STUDIES IN LIGAND UNBINDING TRANSITION STATE PLASTICITY FOR KINETICS-ORIENTED DRUG DESIGN

By

Samuel D Lotz

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Biochemistry & Molecular Biology – Doctor of Philosophy

April 15, 2021

ABSTRACT

STUDIES IN LIGAND UNBINDING TRANSITION STATE PLASTICITY FOR
KINETICS-ORIENTED DRUG DESIGN

By

Samuel D Lotz

The dominant objective in drug design historically has been to improve drug efficacy through increasing affinity of drug binding under equilibrium conditions. However, a complete understanding of drug efficacy in non-equilibrium living organisms requires knowledge of the kinetics of drug (un)binding. While kinetics-oriented drug design has gained popularity it is still hampered by a number of limitations, not limited to the availability of structural models for ligand (un)binding transition states. In general these kinds of simulations are very difficult to achieve due to the long natural timescales of these processes (seconds to minutes) compared to the short timescales at which MD is computed (femtoseconds). In this thesis we address these limitations through computational methods for simulating full, unbiased, unbinding trajectories of inhibitors of drug targets with clinical interest. This is accomplished primarily by applying an enhanced sampling technique, called weighted ensemble (WE), over classical molecular dynamics (MD) simulations. Our approach is drastically more efficient than brute-force simulation methods, requires no biasing forces or other force field modifications, and is shown to work for a variety of systems of interest. Using these methods we are able to model, at all-atom resolution, the structure of unstable transition states for inhibitors of clinical interest of the soluble epoxide hydrolase (sEH) enzyme. This enzyme is implicated in a number of therapeutics including treatment of diabetic neuropathic pain. Critically, we also investigate the role of transition state plasticity in lead optimization. Towards this we developed a model for predicting plasticity from experimental data and a strategy for verifying these predictions which was applied in the context of sEH lead optimization.

To my wife.

ACKNOWLEDGMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

TPPU Inhibitor of sEH in crystal structure 4OD0. ix, xx, xxiv, 25, 26, 142, 143, 144, 156, 162, 166, 169, 178, 180, 182, 184, 187, 189

TS transition state. xi, 6, 102, 105, 129, 132, 133, 142, 143, 154, 155, 156

MD molecular dynamics. xi, 7, 8, 9, 10, 11, 24, 88, 137, 158, 159, 186, 188

BUT 4-hydroxy-2-butanone. xii, 70, 74, 75

DMSO dimethylsulfoxide. xii, 70, 74, 75

DSS methyl sulfinyl-methyl sulfoxide. xii, 70, 74, 75

MFPT mean first passage time. xii, 22, 75, 77, 94

sEH soluble epoxide hydrolase. xv, xx, xxiv, 23, 24, 25, 26, 102, 106, 107, 108, 118, 122, 131, 132, 137, 138, 141, 142, 147, 149, 156, 158, 159, 162, 164, 186, 187

LHS left hand side. xv, 108, 118, 119, 180, 185

RHS right hand side. xv, 108, 118, 119, 120, 121, 180, 182, 185

CP center pinch. xv, 108

SASA solvent accessible surface area. xv, xvi, xxi, 97, 109, 110, 111, 116, 118, 129, 131, 132, 169, 172

4OD0 Protein Data Bank (PDB) entry for sEH bound to TPPU. xxiv, 143

PDB Protein Data Bank. xxiv, 102, 131, 158

AUC Area Under Curve. 1, 3

RT residence time. 4, 22

NMR nuclear magnetic resonance. 5, 7

SPR surface plasmon resonance. 7

SKR structure kinetic relationship. 7, 23, 25, 140, 141, 154, 155, 156

SAR structure activity relationship. 7, 140

GPU graphics processing unit. 9, 28, 70, 71, 87, 103, 141

FKBP FK506 binding protein. 10, 23, 24, 25

TAMD temperature accelerated molecular dynamics. 11

WE Weighted Ensemble. 12, 13, 14, 18, 23, 27, 74, 88, 159, 186, 187

CV collective variable. 13, 14, 15, 16, 17, 18

CSN conformation space network. 24, 102, 104, 105, 109, 114, 119, 124, 131, 161, 168, 169, 174, 178, 180, 182, 184

TSE transition state ensemble. 25, 106, 156, 161, 169, 175, 178, 180, 184, 187, 189

MSM Markov state model. 25, 106, 139, 161, 169, 173, 175, 188

TPT Transition Path Theory. 25, 105, 188

RMSD root mean square deviation. 88, 109

CGENFF CHARMM Generalized Force Field. 103

SI Supplemental Information. 104, 105

GPCR G-protein coupled receptor. 131, 132

PCA principle component analysis. 180, 184

PC principle component vector. 180, 182

CHAPTER 1

INTRODUCTION

1.1 The Role of Ligand (Un)binding Kinetics in Drug Design

The dominant objective in drug design and discovery has historically been to improve drug efficacy through increasing the affinity of drug binding [7]. This has been an effective strategy and has been extensively addressed computationally through the calculation of binding free energies [8, 9, 10, 11]. Thermodynamic drug affinity is the "driving force" of ligand binding in living non-equilibrium systems, but a full description of drug action requires kinetic information as well. The diagram in Figure 1.1 shows how the elimination processes of the organism can effect the overall efficacy of a drug by reducing the integral target occupancy over time spent inside the so-called “therapeutic window” (in general called the Area Under Curve (AUC)).

There are three factors at play here. The first is that there is a minimum effective concentration of receptor-ligand complexes ($[RL]$) which must be formed for the downstream effects of the drugs inhibition to take place. This depends both on the ability of an inhibitor to elicit a response and the overall “vulnerability” of the target. Target vulnerability is a property of the pharmacodynamics of the biological pathways the target is involved in and effectively is not mutable through intervention. High vulnerability targets have lower minimum effective concentrations and are more likely to be “druggable”. Whereas, targets with low vulnerability have higher minimum effective concentrations, requiring larger dosing of drugs.

This leads to the second factor which is the minimum toxic concentration of a drug in an organism. Toxicity primarily occurs when drugs, or metabolites thereof, interact with unintended targets, leading to deleterious side effects. Drugs differ in their levels of toxicity & specificity and must be carefully balanced with the efficacy of the drug. As a drug

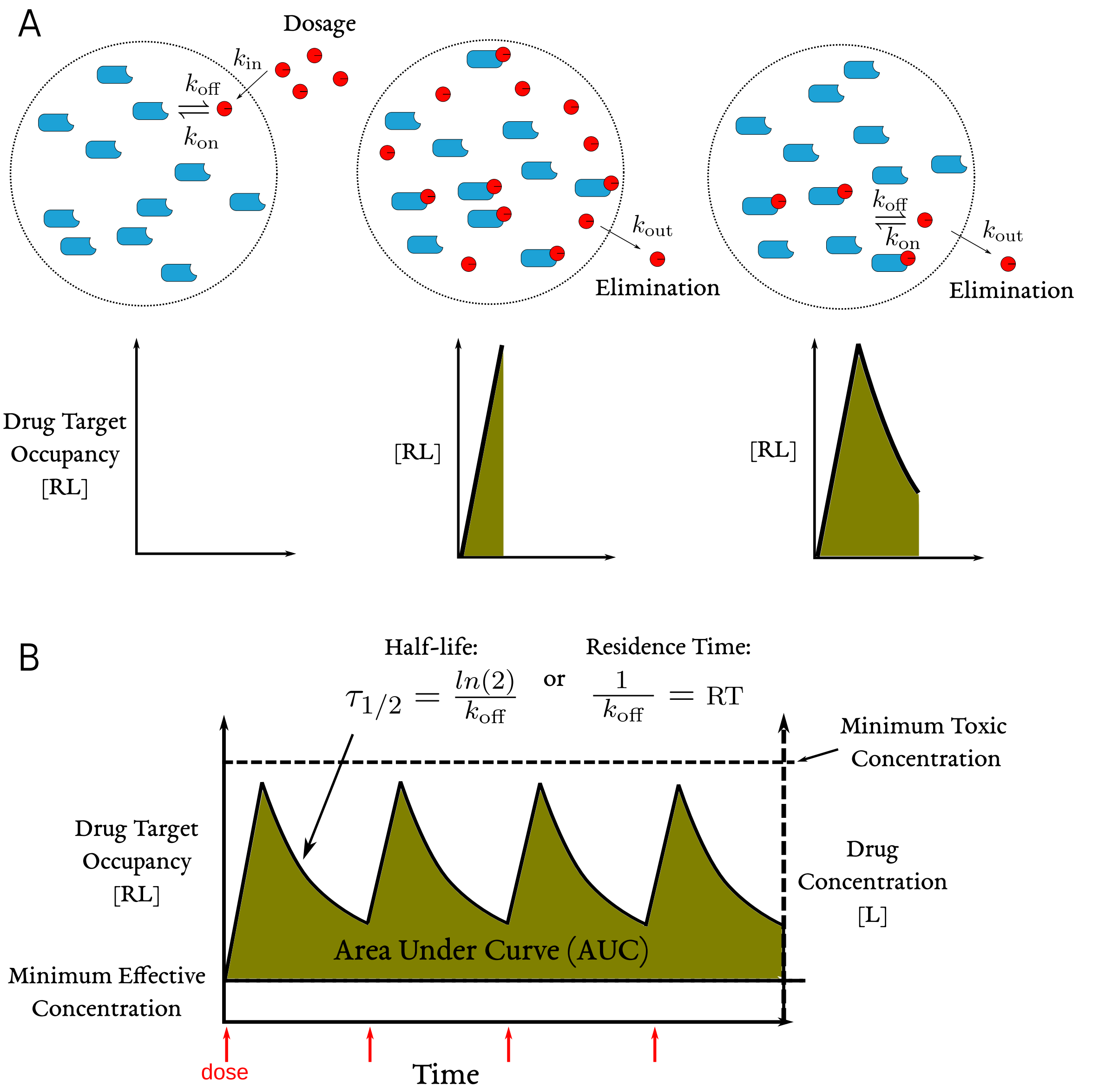


Figure 1.1: Illustrations of (**Panel A**) how the ligand unbinding rate k_{off} affects the overall half-life of a drug in an open (organism-like) system and (**Panel B**) an overview of factors that comprise the “therapeutic window” in a specific dosing regime.

designer may not have much room for change once effective inhibitors are developed some other strategies can be used to limit toxicity. An obvious strategy is to manipulate the dosing regimen so that smaller doses are taken at higher frequency. In this way the AUC is increased while spikes in concentrations of the drug in the body are limited thus limiting toxic exposure.

Optimization AUC in the therapeutic window is also constrained by the third factor which is the rate of elimination of a drug from the organism. We will refer to this simply as the rate of decay (or half-life) of a drug, although the specific model of this decay is subject to specifics pharmacokinetics and is not necessarily a simple exponential decay. A short half-life for a drug means that the concentration of a drug quickly drops after dosing. This exacerbates the issue, again, either larger doses or increased dosing frequencies must be used to compensate.

The half-life of a drug can be controlled in two main ways. First, the molecule can be chosen such that the organism eliminates it less quickly. This essentially affords a long time in which drug ligands have to reach and bind to their targets, as well as allowing significant amounts of rebinding to take place. Secondly, the unbinding rate (k_{off}) from the target can be decreased such that once the ligand is bound to the target it is less accessible to the elimination processes.

On the other hand optimizing for ligand unbinding kinetics has been relatively ignored in the past. This is likely to be for a couple of reasons, first being that the mechanisms by which unbinding rates are determined are either completely opaque or extremely challenging to observe both experimentally or otherwise. Furthermore, knowledge gained about ligand unbinding transition states is not very transferable. Obviously, unique targets don't have similar ligand bound structures and so very unlikely to have similar unbinding transition states.

Developing new drugs is a very challenging field which has manifold constraints on any given problem to be solved. For some drug targets these constraints can coincide such that it is labeled “undruggable”. Undruggable targets are typically valuable targets (by nature of being difficult to target) and so any potential opening up of design space has the potential to lead to a breakthrough in druggability. A focus on optimization of ligand (un)binding rates, which we call “kinetics-oriented drug design”, thus has potential then to add another tool to a drug designers kit in particularly recalcitrant projects.

Somewhat recently in the history of drug design the potential of kinetics oriented drug design has been recognized and an emerging sub-discipline has begun to focus on it [12, 7, 13, 14]. Of particular interest is the residence time (RT) ($1/k_{\text{off}}$) which has been shown to be particularly good at predicting drug efficacy for some targets [15, 16, 17, 18]. Higher RTs, coupled to an appropriate dosing strategy, can be an effective way to protect the drug from being metabolized or eliminated by the body and increase receptor occupancy over longer periods of time [16, 19, 20, 2].

1.2 Theory & Strategies for Kinetics Oriented Drug Design

In the previous section, we have established why there is an interest in optimizing kinetics for drug design. Unfortunately, kinetics oriented design is hampered by a number of fundamental problems that we will attempt to address. In this section we will describe the biophysical and thermodynamic mechanisms by which kinetics are determined for molecular systems as well as a strategy for modeling these mechanisms. The ultimate goal would be to predict kineticss in much the same way that binding free energies are. This thesis attempts to be an early investigation into how this might become possible and outstanding problems to be solved.

1.2.1 Biophysical Mechanisms of Ligand (Un)binding Kinetics

The mechanisms for ligand kinetics – namely the rates of binding (k_{on}) and unbinding (k_{off}) – is in some sense analogous to that of the calculation of binding free energies (i.e. K_d). Using Figure 1.2 as a guide, we remind the reader of how the absolute free energies of the different macrostates determine the binding free energy ΔG and how this relates to the affinity of the ligand to a target. The relative free energy is given by the difference between the absolute free energies of the bound and unbound states:

$$\Delta G = G_U - G_B$$

This quantity then estimates the affinity according to the relation:

$$\Delta G = -RT \ln \frac{K_d}{c_0}$$

where c_0 is the reference concentration of ligand of $1\,\mathrm{mol\,L^{-1}}$. This has an important implication for drug design: to describe the ligand affinity one only needs information on the two stable endpoints.

The features of a ligand that determine G_B are typically much more sensitive and specific to the target structure, electrostatics, etc. Thus, this is where the most drug "design" efforts are expended. While B structures are significantly more difficult to observe than U structures, theoretically this is an energetically stable state and can be observed at macroscopic scales. Structural biology methods like X-ray crystallography, NMR, cryo-electron microscopy, etc. have long been developed and for many systems can readily produce structures at or close to atomic resolution. Structural information is invaluable to drug design as researchers can make and test hypothesis in a much less random manner by matching ligand features to the binding sites of targets, thus stabilizing bound states and generating more favorable affinities.

Taking this idea further many computational methods have been developed to both quantify ΔG from given structures as well as predicting ΔG *de novo* for ligands with no experimental structure. Many of these methods take extensive advantage of the aforementioned fact that no specifics of the unstable intermediate states are strictly necessary to make ΔG calculations. Conversely, the measurement and prediction of kinetics presents a much

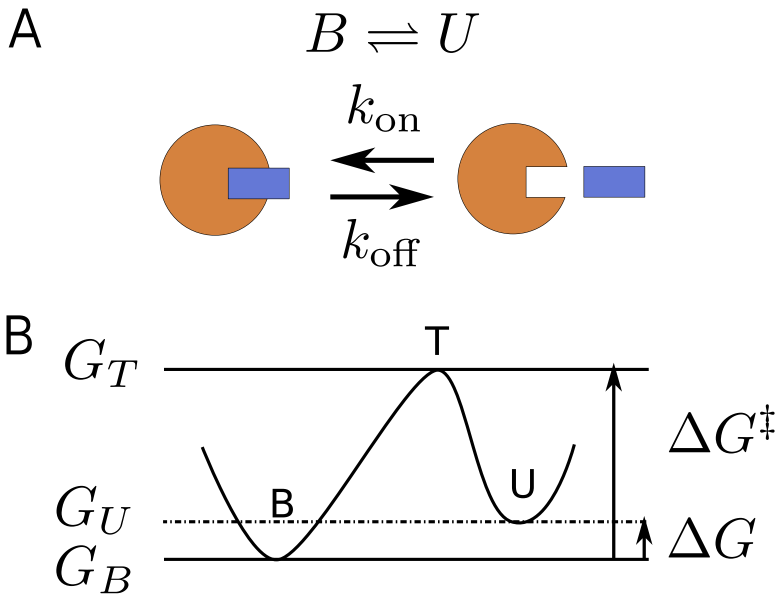


Figure 1.2: Illustration of the free energy landscape for a ligand (un)binding process. **Panel A** describes a two-state system of ligand bound B and ligand unbound U. There are two possible transitions: binding ($U \rightarrow B$) and unbinding ($B \rightarrow U$). The rates associated with each (k_{on} and k_{off} respectively) are the macroscopic quantities we want to describe. **Panel B** shows an illustration of the free energy surface (the curved line) is projected into a single descriptive dimension known as the "reaction coordinate". The two low energy regions of the curve labeled B and U correspond to the stable states in Panel A. The T label corresponds to the "transition state (TS)" between the two stable states. On the left hand side of the graph the absolute free energies of the labeled states are given (G_U, G_U, & G_T). On the right hand side the "relative free energies" which are relevant for our purposes are also labeled. ΔG is the relative free energy of ligand binding and $\Delta G^{\ddagger}$ is the relative free energy known as the "activation energy".

larger challenge as it is dependent not only on the free energies of the stable endpoints, but also that of the unstable transition state, G_T.

Namely the activation energy is given by:

$$\Delta G^{\ddagger} = G_T - G_B$$

And estimates the ligand unbinding rate (k_{off}) according to Transition State Theory [21]:

$$\Delta G^{\ddagger} = -RT \ln \frac{k_{\text{off}}}{c_0}$$

Because rates depend on short lived transition states, this makes them nearly impossible to directly them observe through x-ray crystallography, NMR, or any other method which takes measurements at macroscopic scales.

Fortunately, there are many methods for experimentally determining the macroscopic rates (i.e. k_{off} and k_{on}). Binding kinetics assay methods include surface plasmon resonance (SPR), radioligand binding assays, and fluorescence detection [22, 12, 23, 24]. As will be discussed in this thesis these experimental values are important for evaluating ligand unbinding models, and can even be combined with experimental ligand affinities and solvation energies to make more accurate models.

It is worth noting that structural information that can be extracted through (un)binding rates data in the form of structure kinetic relationship (SKR) models. This involves creating featurizations of a series of ligands (and perhaps targets) and applying machine learning techniques. There is a similar use of structure activity relationship (SAR) models for modeling ligand binding affinities. However these models have the ability to be validated through the use of structural biology techniques mentioned above, unlike SKR models.

For these reasons, in this thesis we take the approach of attempting to model at atomic resolution ligand unbinding transition states through the use of simulation methods. In the next section we outline this approach as will be relevant for the remainder of the chapters.

1.2.2 Resolving Transition States with Molecular Simulation

As we introduced in the last section computational methods are an attractive approach to modeling ligand (un)binding mechanisms. In particular we are interested in atomic resolution of ligand unbinding transition states for kinetics oriented drug design. There are a plethora of different modeling techniques that have been developed specifically for modeling protein-ligand complexes, protein folding, conformational changes, etc. [25, 26, 27, 28, 29]. Some of these methods are directly physics based, such as molecular dynamics (MD), whereas others, such as docking or Rosetta, are more reliant on heuristics that have been found to be useful. Physics based approaches in general have less model error and are relatively under-fit to solving any single problem, however are typically much more expensive to compute [30]. Heuristic based approaches on the other hand are typically cheap to compute, but model error tends to be higher without extensive validation and also requires *a priori* information on the processes of interest to develop models [31]. In reality there is no clear-cut line between the two as even physics based models can have significant model error and heuristic approaches can actually be more accurate depending on the type of observables being predicted. At this time, we are unaware of any heuristic methods that could be applied to modeling ligand unbinding transition states. Thus, we have chosen to use a physics based approach that will provide a more *ab initio* approximation; that method is MD.

We have chosen MD because in theory it allows us to sample unstable microstates (i.e. the ligand unbinding transition state) from a statistical ensemble at atomic resolution and at speeds which are achievable given the current state of computational resources. Additionally, it should be possible to predict macroscopic quantities such as (un)binding rates nonparametrically as the absolute accuracy of force fields are not tuned specifically with rates in mind. That being said biomolecular MD force fields have a long history and are quite accurate for many phenomena [28, 32], and it seems the best choice for this scenario.

The main challenge to using MD simulations, however, is the large discrepancy in the natural timescales of the processes we would like to simulate and the timescales on which MD simulations are computed. Figure 1.3 illustrates the difference in scales between that of MD simulations and some common biomolecular processes of interest. MD simulations time steps are constrained by the fastest fluctuating degree of freedom. In order to maintain stability of these motions time steps are typically constrained to the 1 fs to 2 fs range (denoted t). The real-time cost for a force evaluation is constant for a given algorithm and hardware platform, thus low t values greatly hamstring total throughput (denoted as a "real" time $\bar{T}$).

We should also note that the timescales for macroscopic values (e.g. the rate of unbinding) are ensemble averages (denoted $\hat{T}$). Thus we would expect for a microscopic simulation, like

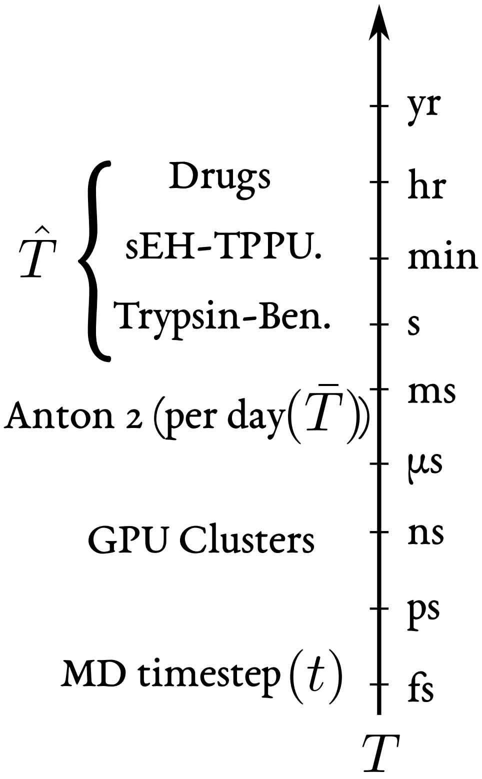


Figure 1.3: Diagram illustrating the order-of-magnitude differences between the timescales of molecular dynamics (MD) propagation time steps, achievable sampling ranges on commodity (e.g. GPU Clusters) & special purpose computer architectures (e.g. Anton 2), and the ensemble average timescales of ligand unbinding processes of interest for drug design.

in MD, the total sampled time (number of time steps × t, denoted T) would correspond to $\hat{T}$.

Fortunately, this problem is partially mitigated by the vast improvements in general purpose computing power [33] as well as development of computer architectures that are particularly suited to MD such as graphics processing units (GPUs) [34] or the Anton series of supercomputers [35, 36]. Even with the state of the art acceleration the sampling range (T) is limited to the microsecond to millisecond range (see Fig. 1.3). As shown in 1.3 this is still very far from the timescales needed to simulate many ligand unbinding processes, which can be as long as minutes to hours [37, 16].

Until recently most simulations of ligand unbinding were limited to model systems, such as small solute-like molecules unbinding from the FK506 binding protein (FKBP) [1, 38, 39, 40, 40] and the protease inhibitor benzamidine binding to trypsin [41, 42, 43, 44, 45, 46, 47]. Small ligands studied for FKBP have low affinities (250 µM to 20 000 µM) and residence times in the 10^{-9} s timescale, and are useful for testing sampling algorithms or off rate calculations. Benzamidine is a ligand with a residence time of 1.7×10^{-3} s on the serine-protease, trypsin [3]. For this reason trypsin-benzamidine is a popular model system for studying ligand unbinding processes and for the development of new sampling algorithms. The use of the MD supercomputer, Anton, has enabled unbiased binding simulations of dasatinib to a kinase receptor (a 200 ms timescale event), but report only 1 to 3 such events using 6 copies of the ligand in 32.5 µs of sampling [48].

So while hardware acceleration is a massive breakthrough in biomolecular simulations [48, 49] it is not a panacea. In the following sections we will discuss some higher level algorithms that can dramatically improve the performance of MD simulations in obtaining data for specific processes of interest, while still taking advantage of these hardware accelerations.

1.2.3 Accelerating Simulations with Enhanced Sampling Algorithms

As we have shown in the last section hardware accelerations are invaluable, yet not enough to perform simulations of the biomolecular processes we are interested in for kinetics oriented drug design. The solution to this issue is a family of algorithms called “enhanced sampling”. In general these are methods which somehow inject *a priori* knowledge about a given phenomena to accelerate sampling of the portions of the search space which are relevant to the researcher. This has the potential to gain information of long timescale processes using only short-timescale simulations that are achievable on available computer hardware. Keeping with the example systems from the last section we outline some of the successes enhanced sampling applications.

Sampling with metadynamics for the dasatinib example was able to observe 12 unbinding events of dasatinib, each taking 150 ns to 750 ns of MD [50]. Other enhanced sampling techniques were used to reveal unbinding trajectories for a family of p38 binding drugs with residence times of approximately 7 s including metadynamics [41, 51] and umbrella sampling [52]. Other methods which do not require progress coordinates (unlike metadynamics), such as temperature accelerated molecular dynamics (TAMD) [53, 54], have been used to simulate ligand unbinding from the Adenosine A_{2A} receptor on a residence timescale of 5.04×10^3 s (84 min) on the Anton supercomputer [49]. Methods such as scaled molecular dynamics [55, 56] and steered molecular dynamics [57], have been shown to be efficient at predicting unbinding rates.

Understandably, the large diversity of enhanced sampling methods can be difficult to grasp. There are roughly two main sub-categories of enhanced sampling methods: 1. those that modify the Hamiltonian of the simulation system (force fields, temperature, pressures, molecular structures, etc.) as the simulation is running, and 2. those that only modify the sampling strategy taken in the simulation (importance sampling, variance reduction etc.). We call the former "bias methods" and the latter "variance methods", as these are essentially two solutions to the well known bias-variance trade off in machine learning [58].

Examples of bias methods that are commonly applied to MD simulations include: replica exchange [59], metadynamics [60, 61, 51, 41], temperature accelerated MD [54, 53], and umbrella sampling [62]. These methods invoke an assumption that the molecular systems are in equilibrium, usually in the form of a canonical probability density function. Additionally, although there are some approaches to approximate rates based on biased simulations [60], this approach makes it very difficult to recover full atomistic detail of transition state ensembles that govern the forward and backward rate constants for a given transition.

These two points have implications for simulations used in kinetics oriented drug design. Firstly, as described in Section 1.1 one of the principle motivations for optimizing residence times is precisely because *in vivo* systems are not at equilibrium. Secondly, as our goal is to recover accurate atomistic structural data for transition states and paths these methods are not suitable. Thus we turn to the variance methods which can simulate long timescale processes using the unbiased Hamiltonian of interest, insofar as the base force fields are unbiased.

Typically variance-based enhanced sampling can be considered "path sampling" techniques. These methods apply a sampling process over a collection of trajectories rather than single conformations [63]. Importantly they do not require modifications to the underlying dynamics model and provide contiguous trajectories of transition paths. This makes them suitable for studying all varieties of path-dependent observables (e.g. (un)binding rates), as well as providing detailed atomistic models of transition states. There are a variety of path sampling methods including transition path sampling (TPS) [64, 65, 66], forward flux sampling (FFS) [67], multilevel splitting [68], and Weighted Ensemble (WE) [69, 63]. Of these, Weighted Ensemble (WE) has advantages in that it does not in general require a Markovian assumption or *a priori* knowledge of full trajectories connecting two states to start [63, 70] and does not require the definition of a progress coordinate, which will be discussed further below. For these reasons and others, in this thesis we make extensive use of the WE path-sampling method and will be described in detail in Section 1.2.4.

1.2.4 Weighted Ensemble in General

The weighted ensemble algorithm (WE) is a general strategy for simulating rare or long-timescale events in stochastic systems [69]. WE is a conceptually portable method that can be applied to any field of study including molecular biophysics [71, 72, 73, 74, 39, 75, 76, 77, 78, 79, 80], systems biology [81, 82, 83, 84, 85, 86, 87, 88], telecommunications [89], and aerospace and engineering [90]. Its fundamental features are as follows.

A set of trajectories (individually called "walkers") are propagated forward in time in a parallel fashion, each one assigned a statistical weight (p). Periodically (on some time interval τ), the trajectories are "resampled" to rebalance the computational effort toward lower-probability regions. This is done by cloning trajectories in sparsely-populated, lower-probability regions, and merging together trajectories in over-populated, higher-probability regions. To maintain proper statistics, the weight of a cloned trajectory is divided amongst its progeny, and the weights of merged trajectories are summed and given to the resulting trajectory. Here we refer to this process as “resampling”.

More formal definitions of WE is discussed in [70, 63, 76], and some useful definitions are made in Section 2.2.1 as well. Importantly, as was shown in Zhang *et al.* [91], WE has been shown to be statistically exact. This feature in essence is inherited from the properties of using resampling as the core mechanism. That is given an ensemble of trajectories sampling a distribution, an infinite amount of sampling will converge to the exact distribution (assuming ergodicity for the domain). If resampling is done at some point in these simulations nothing has changed, since no new samples are being drawn. Resampling does nothing to change the theoretical exactness of the underlying dynamics and only effects questions of variance and rates of convergence. That is resampling can make convergence to a distribution occur faster or slower or may alter the overall variance of the distribution at intermediate steps, but cannot change the identity of that distribution.

Notice also that the specifics of the resampling function are also completely irrelevant to the correctness. New resampling methods do not require new proofs to recover the correctness of the underlying dynamics. This freedom allows for a huge amount of creativity and diversification of resampling algorithms which can be freely mixed and matched to suit the variance characteristics desired. In the next section we will begin to discuss some of strategies for designing effective resampling algorithms.

1.2.5 Algorithmic Strategies for Weighted Ensemble

Given the abstract view of how and why WE works, we must now discuss the various strategies to implementing it efficiently. In this thesis we adopt the use of a class of WE strategies that are well suited to sampling spaces of inherently high dimensions, namely WExplore [92] and REVO [93]. In order to contextualize and motivate these algorithms we first discuss the common problem of determining appropriate collective variables (CVs).

The use of CVs is a common feature of many enhanced sampling methods such as “Adaptive Multilevel Splitting” [68], metadynamics [61], and WE [91]. We then present an alternative to CVs in the form of **distance metrics** and discuss the merits.

CVs are simply a separate, latent, space, K, for which a single point/state in the space of all simulation degrees of freedom, L, can be mapped into by a function F:

$$F(X) \rightarrow Y; \quad d : \mathbb{R}^L \rightarrow \mathbb{R}^K$$

Here the idea is that improving sampling efficiency is easier within K when $dim(K) < dim(L)$ and that the relevant degrees of freedom have not been projected together. This idea is a familiar one and there is an entire field dedicated to dimensionality reduction that can be applied here [94], but in general it is a very difficult problem. Finding $K = 1$ where all relevant degrees of freedom are still captured is identical to finding the reaction coordinate CV for a transition process. Other $K = 1$ spaces can be thought of as “progress coordinates”, if they capture at least some of the important degrees of freedom for the process.

In practice, despite the difficulty of this problem we should be optimistic in searching for low dimension CVs, as long as the probability distributions for the events of interest are non-uniform. However, we shouldn’t rely on being able to obtain arbitrarily small $dim(K)$. Indeed for some sampling problems there isn’t any possible way to reduce $dim(K)$ any further. For example, in simulations of ligand unbinding we may be interested in multiple unbinding pathways. This could be because there is some uncertainty as to which pathway is the correct one (the one with the lowest free energy path integral), or because there are multiple paths with similar probabilities that contribute to the macroscopic process. Extra pathways also give a fuller picture of mechanisms by which ligands unbind from a receptor and should be useful in drug and protein design. In any case, the number of unbinding paths is not known in advance and each would be described by a unique reaction coordinate. Thus each extra path that you want to simulate will incur at least one extra dimension, if we are using reaction coordinates. Likely, and luckily, this won’t be the case and there will be progress coordinates that adequately capture the important degrees of freedom for multiple pathways (such as was the case for a volume based CV [95]). In principle though, more paths means either a larger K-space or an even more complex (or clever) F mapping simulation states to that space.

Ensemble methods like WE utilize a number of parallel simulation processes (called walkers in WE) so that a higher throughput of discovery can be achieved. However – because we can not run arbitrarily many walkers at once – ensemble methods limit the number needed through two mechanisms: 1. more efficient binning of the space and/or 2. establishing priorities among regions and optimizing the load balancing (or scheduling) of walkers that are sampling them. The canonical WE [69] algorithm employs a load balancer that just assigns a constant number of walkers to every region that is defined and populated. We call this "fair sharing" because every region is given the same allocation of sampling every cycle. A fair-share resampler's performance is wholly dependent upon the efficiency of the binning. We can readily see that for inherently high-dimensional search spaces that tesselating with uniform bins would be quite wasteful.

Comparatively, a "proportional-share" resampler that weights sampling time allocations upon some metric of convergence within a region should have better sampling efficiency than a fair share for the same binning strategy. The WExplore and REVO resampling algorithms can be considered proportional-share [93, 92] as the effective binning is determined by back-pressure mechanisms that are scaled to local walker cloning and merging activity (see Sections 1.2.6.1 and 1.2.6.2 for details). The purpose of making this distinction between binning and load balancing is to refocus the admittedly complex design of high-dimensional resamplers, by separating out orthogonal design features. An improved, adaptive load balancing algorithm will always benefit sampling efficiency and could be developed separately. Furthermore, load balancing is essentially transparent with respect to the choice of CVs or distance metrics, and therefore to the end user of the enhanced sampling tool.

For the end user they essentially have two major choices for how they are going to decide to bin (or index) the space. We replace the idea of binning with that of **indexing** instead of binning as it is more intuitive (or general), but formally they are synonyms [96, 97]. This makes sense especially for describing algorithms like REVO, which don't actually define any explicit bins. We identify two basic forms of indexing beyond the trivial one: multidimensional and distance-based. Multidimensional indexing is a strategy for indexing where there are a discrete number of recognizable features that can be used to identify a sample. CVs are simply a multidimensional indexing strategy of samples.

Distance-based indexing differs in that it doesn't assume there any enumerable features of a sample, but does assume that there is a way to compute a distance between two samples. Formally, it is a function, d, that maps two points (also called states), X_i and X_j, in a (metric) space, J, to a single positive real number:

$$d(X_i, X_j) = d_{ij}; \quad d : \mathbb{R}^J, \mathbb{R}^J \to \mathbb{R}$$

additionally where the following also hold:

1. Non-Negativity: $d_{ij} \geq 0$
2. Identity of Indiscernables: if $d_{ij} = 0$ then $X_i = X_j$
3. Symmetric: $d_{ij} = d_{ji}$
4. Triangle Inequality: $d_{ik} \leq d_{ij} + d_{ik}$

Different indexing algorithms may technically not require all of these conditions, but typically performance optimizations can be made if they are satisfied. The fourth requirement is the most difficult to prove but can be optional in some cases, however the majority of optimization techniques leverage it and so is highly desirable for performance reasons [97].

The familiar Euclidean distance metric satisfies all these criterion, and is typically assumed to be a synonym for "distance" in general. Particularly many multidimensional indexing algorithms rely on computing the Euclidean distance for points in the CV space. Thus by association, CV spaces are in addition to being a metric space, also always vector spaces. While vector spaces cover a large swath of the interesting features of samples, they do not cover all of them. The most interesting example are those of string comparison distances. These include important applications like Hamming edit distances for DNA sequences, Tanimoto distances for fingerprints in chemoinformatics, and Jaccard indexes in general. One could easily imagine characterizing a simulation of a protein by a string of booleans indicating the presence or absence of secondary structure or intermolecular interactions. In this situation there is a clear way to compare samples based on these features using distance metrics, but not for CV as sets of these features do not form a vector space. Moreover, there are a number of distance metrics for comparing entire probability distributions (e.g. Root-mean square metrics like RMSD and Mahalanobis distance [98]), continuous topological structures (e.g. Hausdorff distance), and network/graph structures (e.g. Wiener index in chemoinformatics [99]).

Additionally, distance metrics avoid explicitly determining the dimensionality reduction mappings (F) into CV variable spaces. The dimensionality of the latent search space is actually fractional when compared to CVs which increment the number dimensions by whole integers [97]. This is incredibly useful as it provides a way to mitigate the curse of dimensionality when possible. The difference is that the dimensionality of the latent space of a distance metric is implicit (rather than explicit), and actually variable depending upon the samples that it will be comparing. A full proof of intrinsic dimensionality and the difficulties associated with calculating it are given in [100, 96], but briefly, the dimensionality of a distance metric is proportional to the density of distances from a reference sample to other samples for a given cutoff. While the hidden and empirical nature of this dimensionality may be somewhat undesirable, in practice we haven't found this to be a problem. Additionally, the choice of a good distance metric can drastically reduce the dimensionality for distance calculations between the actual samples that have been made while still making complex or subtle distinctions in samples. However, we do suggest developing metrics to monitor the latent dimensionality of distance metrics in order to improve or maintain performance in resampling algorithms.

To summarize, the benefits of using distance-based indexing, over multidimensional indexing are that it [101, 100]:

1. does not a need mapping of a sample to a feature vector in a CV variable space,
2. is a general purpose and theoretically uniform abstraction, and
3. can effectively reduce the dimensionality of the search space in fractional increments unlike CVs.

We finally note that we believe the approachability of an enhanced sampling method by increasingly un-specialized researchers is key to the long term usefulness of it. Ideally, we would like to be able to state the constraints that define an event, and perhaps make some noncommittal guesses about the reaction coordinate, and have this be used to enhance sampling of the event. Distance metrics based resampling algorithms are a very good choice for for this among domain researchers because the choice of distance metrics almost completely encapsulates the domain specific knowledge needed to enhance sampling of simulations. We also note that if some researchers happen to be more familiar and comfortable with defining classical CVs this is still compatible within the framework of distance metrics, since it just involves projecting samples onto the CVs and computing the Euclidean distance between these feature vectors. See Section 2.2.3.2 for a discussion on how the use of distance metrics are incorporated into the design of software for running WE simulations. The primary benefit of this is that a distance metric need only be defined once for all distance metric based resamplers (such as WExplore or REVO).

1.2.6 Design of Some Distance-Based WE Resampling Algorithms

In the last section we have described in theory the benefits of the distance metric based adaptive resamplers. Before explaining the details of the algorithms we outline (and look ahead to) some of the specific successes obtained.

The domain this has been shown most successful is in obtaining preliminary rare event simulations for systems with long waiting times [88, 47, 102, 80, 103] (and as described in this thesis). As we will see in Chapter 5, WExplore simulations produced unbinding trajectories of a drug ligand (TPPU) from its target (soluble epoxide hydrolase) that has an experimentally determined mean first passage time of 11 min, using less than 1 µs of simulation; a speed-up of 10^9-fold [102]. WExplore and REVO have shown to be particularly useful for discovering multiple pathways. Both algorithms were able to discover multiple ligand dissociation pathways for the trypsin-benzamidine system, which requires substantial rearrangement of the loops comprising the trypsin ligand binding pocket [93, 47] (and as described in Chapter 4). In contrast, single and low-dimensional projections like reaction coordinates often constrict the search space to particular paths, which precludes the discovery of alternative paths (and transition states) between macrostates. Finding multiple pathways can be particularly useful for applications like kinetics-based drug-design when we want to understand the structure of the ligand-binding transition state not only for a particular ligand, but also transition states for closely-related ligands. A final benefit is that often adaptive algorithms like WExplore and REVO (as well as the history-augmented Markov state model WE method [104]) require less up-front parametrization such as the definition of bin boundaries. For these reasons we have chosen to use these high dimensional adaptive resamplers for the individual investigations in this thesis.

In the following sections we describe the WExplore and REVO algorithms in more detail.

1.2.6.1 WExplore

The main problem with defining bins in a high-dimensional space is that the number of bins needed to cover the space scales exponentially with the dimensionality of a space. This number of bins is proportional to the overall computational effort of the simulation, as a target number of trajectories are to be run in each bin. The WExplore algorithm was

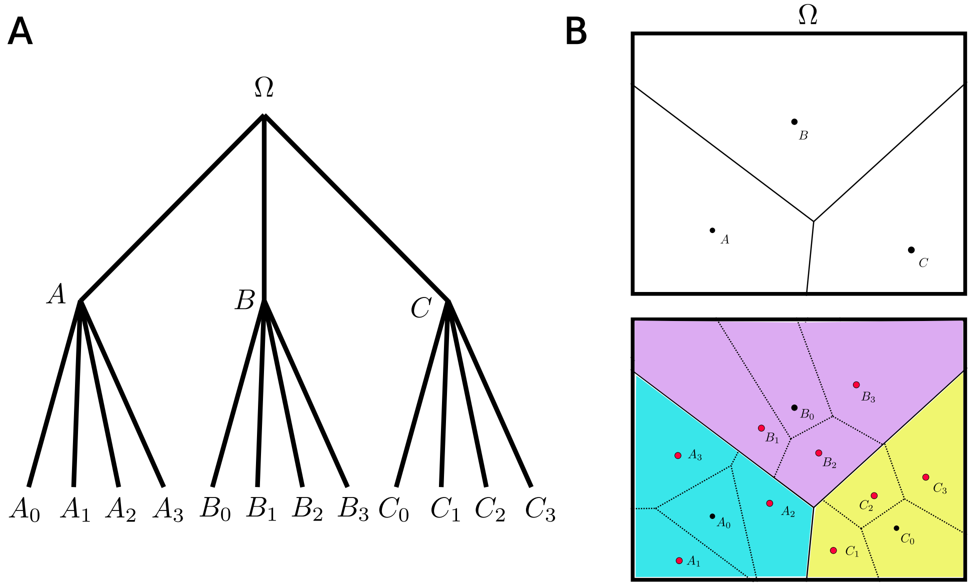

A
B
Ω
A
B
C
A_0 A_1 A_2 A_3 B_0 B_1 B_2 B_3 C_0 C_1 C_2 C_3
Ω
B
A
C
B_3
B_0
B_1
B_2
A_3
A_0
A_2
A_1
C_3
C_2
C_0
C_1

Figure 1.4: Diagram explaining the adaptive binning and walker allocation strategy of the WExplore resampling algorithm. **Panel A** shows the tree structure of the region hierarchy, that corresponds to the Voronoi regions at each level shown in **Panel B**.

developed by Dickson and Brooks to circumvent this difficulty [92]. The key is to construct a set of hierarchical regions: a small set of large regions that tile the entire space, which are each subdivided by a set of smaller regions, which are in turn subdivided by even smaller regions. In the WExplore resampler these regions are Voronoi polyhedra that are defined by points (called "images") in a high-dimensional space. In order to assign a trajectory to a region, we must measure the distance from that trajectory state to each image. The trajectory is then assigned to the region with the closest image. More information on this algorithm can be found in the original paper [92].

In a typical WExplore simulation, we start with only a single image, and define new images as they are visited by the set of trajectories. A new image is defined when a trajectory in the ensemble reaches a new region of space, or more precisely, when the distance to the closest image is greater than a critical value. The list of critical values ($d_{min} = (d_1, d_2, ..., d_n)$) thus control the sizes of the regions at each level of the hierarchy, and are parameters of the WExplore resampler. As the set of images grows over the course of the simulation, the WExplore resampling function (denoted $\mathcal{R}$ below) can change with each resampling step.

1.2.6.2 REVO resampler

Here we briefly outline the REVO algorithm as was introduced in [93].

Although WExplore presents a viable means of indexing tessellations of high-dimensional spaces, there are still some behaviors related to the construction of regions that are non-optimal – most notably, the discontinuous behavior related to reaching new levels of the hierarchy for the first time. For this reason, a new resampling algorithm was recently proposed which avoids the construction of sampling regions altogether. In the REVO algorithm (Resampling of Ensembles by Variation Optimization), an objective quantity called the "trajectory variation" (denoted, V) is used to guide the merging and cloning process [80, 93]. We calculate V before and after each proposed merging and cloning operation and execute only the operations that cause V to increase.

The variation is given by:

$$V = \sum_i \sum_j \left(\frac{d(X_i, X_j)}{d_0} \right)^\alpha \phi(X_i)\phi(X_j) \tag{1.1}$$

where $d(X_i, X_j)$ returns the distance between trajectories i and j, and $\phi(X)$ is a function that describes the importance of individual trajectories. The exponent α allows us to balance the relative strength of the distances and the importance functions, and the d_0 value does not affect resampling, but serves to keep the variation function unitless. On the whole, the variation function increases as the ensemble of trajectories get further apart.

The importance functions can be defined to take the weight of the trajectories into account:

$$\phi(X_i) = \log w_i + C \tag{1.2}$$

where C is a constant. This has the effect of prioritizing not only a broad ensemble of trajectories, but one where the highest weighted trajectories are distributed as far as possible from each other. This strategy can minimize the error in the calculation of observables that depend on the weights of rare trajectories, such as transition rates.

1.2.7 Non-Equilibrium Simulations & Rate Calculations

The studies in this thesis are particularly focused on ligand unbinding and the rates thereof. As simulations are relatively expensive we constrain them to only simulate the one way process of unbinding. In practice this involves definition of some trajectory boundary conditions that qualify a state as one in which the ligand is unbound. Typically this is some distance cutoff from the ligand to the receptor (details of these boundary conditions are given in the methods of each study). When a walker trajectory qualifies the criterion of a boundary condition the simulation is terminated and restarted from one of the initial states; a process we call "warping". Importantly, while the state variables (e.g. atomic positions) of the walker trajectory are altered the weight of the walker is not, which conserves probability in the steady-state ensemble. Also of importance to the computation of rates is that the total amount of weight that passes through the boundary (the probability flux) also be accounted for. Accurate rate calculations can be made when the rate of change of the flux has converged. Such simulations are called "steady-state non-equilibrium" simulations.

To calculate the actual rates and RT (equivalent to the mean first passage time (MFPT) of unbinding) we use the so-called Hill relation which states that the MFPT is equal to the inverse of the probability flux into the sink state [105, 106, 63, 107]. The flux measurement at time t for a single run is calculated by taking the sum of the weighted passage times:

$$\frac{1}{RT(t)} = \text{Flux}_{A \to B}(t) = \frac{\sum_{\gamma \in \Gamma(t)} P(\gamma)}{t}$$

Where $\Gamma(t)$ is the collection of all reactive trajectories that have occurred before time t. A reactive trajectory γ is a walker trajectory in which the sink (B). These trajectories have probability $P(\gamma)$ equal to the weight of the reactive walker when the boundary was crossed.

1.3 Outline of Work

The overarching goal of this thesis is to develop computational methods that aid in kinetics oriented drug design. The specific end goals are to:

- enumerate and characterize the nature of ligand unbinding pathways,
- estimate unbinding rates,
- resolve atomistic structures of ligand unbinding transition states, and
- develop models for optimizing experimental strategies for explaining structure kinetic relationship (SKR) models.

These are relatively understudied topics for real drug target candidates due to the number of technical challenges. As such a large portion of this thesis (Chapters 2, 3, and 4) will be focused on the incremental development and validation of computational techniques that can be leveraged towards these end goals. Ultimately, however we also use the developed techniques to study a realistic drug target, soluble epoxide hydrolase (sEH), in Chapters 5 and 6.

We first describe the a software package `wepy` [108] that was developed during the course of these investigations by the author to aid in the simulation and analysis of the workflows described in the rest of the thesis in Chapter 2. While `wepy` was only used for simulations in Chapter 6, and in the analysis of results for Chapter 5, the overview and perspective of Chapter 2 is relevant to the general understanding of the other chapters. Indeed the investigations in Chapters 3, 4, and 5 were directly used to clarify the specific needs of the software development. The use of `wepy` goes beyond the contents of this thesis and is in use by other researchers to solve similar problems.

Following the description of `wepy`, and associated WE formalism, we discuss the application of the core enhanced sampling algorithms (i.e. WExplore) to two model systems. These are:

- drug-like fragments unbinding from the FKBP protein (Chapter 3), and
- unbinding of benzamadine from trypsin (Chapter 4).

In Chapter 3 FKBP is used as a model system for ligand unbinding in which the timescales for unbinding are very short. From this we begin to gauge the performance and scalability of the WExplore algorithm for sampling ligand unbinding pathways as well as the distribution and structure of ligand associations on receptor binding sites. We are able to evaluate the accuracy of WExplore simulation derived unbinding rates to both experimental values as well as un-enhanced MD; the latter being something that is not feasible with longer timescale processes such as inhibitor unbinding from sEH in Chapters 5 and 6. This is important because in addition to the potentially high-variance of enhanced sampling simulations the accuracy of MD force fields is not well studied or optimized for obtaining accurate rate or transition state predictions. Additionally, we developed workflows for investigating particular intermolecular interactions between ligands and proteins for entire unbinding ensembles.

After our study of FKBP we increase the difficulty of the problem to another model system in Chapter 4: trypsin-benzamadine. This system has a significantly longer unbinding timescale which again tests the ability of our algorithm to scale. Additionally, this is a popular ligand unbinding model system and there are many other studies utilizing both long un-enhanced MD simulations as well as a variety of other enhanced sampling methods. Here we can compare multiple aspects of our approach:

- the accuracy of rate predictions,
- the breadth and diversity of state space sampling, and
- efficiency in terms of total amount of MD sampling needed to achieve these results.

In the trypsin-benzamadine study we also introduce the use of conformation space network (CSN) depictions as a powerful visualization and analytics tool for understanding the complex structure of unbinding pathways. We additionally refine our protocol for investigating the distributions of receptor-ligand intermolecular interactions across entire sample sets.

Following the successes of our protocol in these model systems we then turn to investigating a substantially more difficult system of actual clinical interest in Chapters 5 and 6: soluble epoxide hydrolase (sEH). This is a substantial jump in the magnitude of unbinding timescales from the order 1 ms to 2 ms for trypsin-benzamadine to 500 s to 1500 s for sEH and inhibitors of interest. In Chapter 5 we focus only on a single inhibitor TPPU of particular interest as a drug design scaffold. In this study we again compare rate predictions to experimental values, but we also provide a more in depth structural analysis of the simulations. The structural analysis is important in order to inform medicinal chemists about the relevant interactions that may determine kinetics and affinity. We also introduce the use of a protocol for directly estimating the transition state ensemble (TSE) and dominant pathways using Markov state models (MSMs) and Transition Path Theory (TPT). This information will prove to be invaluable as we can provide a simplified model of the structural features of the rate determining step in ligand unbinding. Additionally, evaluating distinctions between dominant and secondary unbinding pathways is also very important given the much more diffuse definition of unstable transition states.

Following up on this in Chapter 6 we extend our analysis of sEH by simulating several other inhibitors of drug-design interest. This study will push our protocol to the limits and provides a benchmark for the utility of our protocol potentially applied to similar applications. The primary goal of Chapter 6 is somewhat less focused on the structural details of the unbinding paths. Instead we begin to deal with the broader issues of the impacts of:

- transition state modeling methods,
- transition state plasticity (AKA “pathway hopping”) in lead optimization, and
- experimental strategies for efficiently utilizing expensive simulations in structure kinetic relationship (SKR) modeling.

Given the successes of our enhanced sampling simulation protocols for FKBP, trypsin-benzamadine, and sEH-TPPU our attention turns to the post-processing modeling and the outward scaling of these methods to a greater number of phenomenon. Namely, it is one thing to have an isolated success with a single system like sEH-TPPU which may have come at a relatively large expense in terms of man and compute hours, and it is another entirely to apply a method to a large panel of inhibitors. Thus, before choosing an arbitrary set of inhibitors based on the TPPU scaffold we provide a detailed rationale for the choice of these samples such that we can obtain greater insight into the underlying causes of kinetics. This rationale is backed by a theoretical formalism that attempts to predict transition state plasticity for a variety of chemically different ligands. Following the design of an experimental strategy we apply it and attempt to validate the plasticity predictions.

In the final Chapter 7 we revisit the high-level goals and challenges outlined in this section to assess progress, challenges, and the important next-steps for improvement.

CHAPTER 2

WEPY: A FLEXIBLE SOFTWARE FRAMEWORK FOR SIMULATING RARE EVENTS WITH WEIGHTED ENSEMBLE RESAMPLING

2.1 Introduction

In Section 1.2.3 we introduced the need for enhanced sampling algorithms to improve the performance of simulations and why we chose to use the Weighted Ensemble (WE) family of methods for this thesis. An overview of WE in general was given in Section 1.2.4. We then explained the WExplore algorithm in 1.2.6.1 including the benefits to this algorithm. In the current chapter we elaborate on the computational details of WE and describe the implementation of software framework for doing WE simulations.

Despite the advantages of high-dimensional adaptive WE algorithms such as WExplore and REVO, their adoption has been hindered for a number of reasons. Firstly, the implementation of these resampling algorithms are complex and difficult to implement correctly. Secondly, independent implementations of WE algorithms lack interoperability of produced data and so are difficult to compare. Thirdly, the barrier to entry for other researchers to write an implementation of the resampling algorithm as well as progress metrics for their system of interest is prohibitive.

Here we introduce the open source `wepy` software framework for running WE simulations that attempts to address these issues. We first describe a software and data architecture that both reflects a simple mathematical formalism (described in 2.2.1) and also decomposes into multiple modular components. The software architecture allows for reuse of vetted resampling algorithm implementations written by methods researchers with domain specific progress metrics written by users. The data architecture solves interoperability through the introduction of a general purpose decision record design (described in 2.2.3.1).

`wepy` is implemented in the Python 3 programming language and thus allows users to natively leverage a massive ecosystem for scientific computing. Other benefits of a pure-Python implementation are that it 1) increases portability between platforms, 2) has a uniform interface that can be used as a library and embedded into other software easily, and 3) only requires knowledge of a single popular programming language (Python), which if necessary has facilities for writing extremely high performance code (e.g. numba [109], `dask` [110]). Currently `wepy` is tightly integrated with the OpenMM [111] molecular dynamics engine and provides excellent support for running GPU molecular dynamics simulations. The architecture of `wepy`, however, is agnostic to the underlying dynamics engine as well as to any particular parallel computing strategy or framework. The `wepy` project also introduces a high-performance single-file storage format and schema for cloning-merging type simulations implemented in HDF5 [112]. [1] Use of HDF5 also provides "out-of-core" data-structures which allow access to simulation data that does not fit entirely into computer main memory. On top of this an extensive interface (application programmer interface (API)) is provided to make querying, analysis, conversion to other formats of complex path trajectories easy. Part of this interface is to support the intuitive representation of WE trajectories which have been cloned and merged as trees (referred to as "tree-like" data-structures).

While there are two other WE frameworks that have been developed, these have different strengths and design goals. The `AWE-WQ` [76] system provides an implementation of the accelerated weighted ensemble (AWE) along with a "Work Queue" distributed computing framework. However, `AWE-WQ` is less flexible than `wepy` in that it solely implements AWE simulations and is opinionated about the distributed computing framework. The `WESTPA` [113] software suite is a popular implementation of many binning based WE resamplers and provides excellent support for running simulations on large clusters of CPU cores and GPUs. It is implemented in Python and typically run using a combination of UNIX-like shell scripts and YAML configuration files. It shares many of the same benefits of `wepy` discussed above including modularity and the use of HDF5 files for simulation data. `WESTPA` however did not

[1] A thorough description of this format can be found in the documentation for the wepy project `https://github.com/ADicksonLab/wepy`.

satisfy our requirements because the architecture is oriented around fixed-topology explicit binning approaches making implementation of binless algorithms like REVO [93] difficult. Additionally, we also favor the configuration-as-code approach where simulation components are constructed in Python code.

In this chapter we introduce a mathematical formalism that provides an overview of the design and architecture of the system followed by a description of the major software components in `wepy` including how to initialize, run, and analyze simulations. This treatment will focus on the "mechanics" of the algorithm rather than its correctness which is described in Section 1.2.4. Finally, we present an example ligand unbinding scenario, using a Lysozyme model system, along with concrete code examples and explanations.

2.1.1 Calculating transition rates using boundary conditions

The `wepy` software supports the construction of non-equilibrium ensembles to calculate transition rates. In this technique [105, 106, 114] a set of history-dependent ensembles are defined using a set of "basins". For instance, if we are studying ligand binding transition pathways, then the basins will be the "bound" and the "unbound" states. The **unbinding ensemble** is then defined as the set of trajectories that have most recently visited the bound state. When a trajectory from the unbinding ensemble enters the unbound state, it transitions to the **binding ensemble**. In a typical `wepy` simulation, we can initialize trajectories in a given basin and use boundary conditions that terminate trajectories that enter the opposite basin. The weights of these trajectories are used to calculate a transition flux (probability per unit time), which is used to calculate a rate constant (e.g. k_{on} or k_{off}). These trajectories are then "warped" back to the initial state, retaining the same statistical weight.

2.2 Design and Architecture

2.2.1 Formalism

Let us first define ensemble resampling simulations in general. First we define an ensemble to be a finite multiset of walkers, w, of size N:

$$\mathbf{W} = \{w_i\}, \text{ for } i = 0, 1, \dots N$$

where a walker is a set of two elements: a probabilistic weight, p, and a state, X:

$$w_i = (p_i, X_i)$$

An ensemble of walkers defines a probability distribution, $P(X)$, and must be normalized such that:

$$\sum_{0<i<N} p_i = 1$$

There are at least two steps in an ensemble resampling simulation: propagating dynamics and resampling. We can also introduce a third step for so-called boundary conditions that allow for running non-equilibrium simulations and calculation of rate constants. An ensemble resampling simulation process, $\mathcal{A}$ (for "apparatus"), is made up of three components: a runner function $\mathcal{D}$, a boundary condition function $\mathcal{B}$, and a resampling function $\mathcal{R}$.

$$\mathcal{A} = (\mathcal{D}, \mathcal{B}, \mathcal{R})$$

The dynamics can be any stochastic dynamical process such as molecular dynamics, Monte Carlo simulations, etc. Formally, a runner function has the form:

$$\mathcal{D}[\tau](X) \to X'$$

where $\mathcal{D}[\tau]$ is a stochastic function such that multiple evaluations of an input state X might not yield identical X's, and τ is the number of steps of propagation. While $\mathcal{D}$ acts independently on each walker state X it is convenient to write it as

$$\mathcal{D}[\tau](\mathbf{W}) \to \mathbf{W}'$$

The resampling function, $\mathcal{R}$, maps an ensemble, $\mathbf{W}$, to another ensemble $\mathbf{W}'$ along with an updated resampling function $\mathcal{R}'$:

$$\mathcal{R}(\mathbf{W}) \to (\mathcal{R}', \mathbf{W}')$$

The walkers in this new ensemble must the satisfy these constraints:

$$\begin{aligned} \mathbf{W}' = \{(X_i, p_i') : i = 0, 1, \dots N'\} &\text{ where} \\ N' &\text{ is a postive integer} \\ X_i \subseteq \{\mathbf{State}(w_j) &\text{ for } w_j \text{ in } \mathbf{W}\} \\ \sum_{0<i<N'} p_i' &= 1 \end{aligned} \tag{2.1}$$

In words, $\mathbf{W}'$ consists of walkers with states chosen from the states of the $\mathbf{W}$ ensemble. Note that these constraints are necessary but not sufficient to ensure that the sampled ensembles are consistent with the cloning and merging process in the WE algorithm.

The updated resampler $\mathcal{R}'$ allows for algorithms that take into account history dependence. Thus, a resampling function:

$$\mathcal{R}(\mathbf{W}) \to (\mathcal{R}, \mathbf{W}')$$

is said to be stateless and does not take into account any walker history.

While more commonly stateless, boundary conditions and runners also have equivalent history dependent definitions:

$$\mathcal{B}(\mathbf{W}) \to (\mathcal{B}', \mathbf{W}'')$$
$$\mathcal{D}[\tau](\mathbf{W}) \to (\mathcal{D}', \mathbf{W}''')$$

These functions typically utilize an application strategy that follows:

$$\begin{aligned}
\mathcal{R}_0(\mathbf{W}_0) &\rightarrow (\mathcal{R}_1, \mathbf{W}_1) \\
\mathcal{R}_1(\mathbf{W}_1) &\rightarrow (\mathcal{R}_2, \mathbf{W}_2) \\
&\vdots \\
\mathcal{R}_{n-1}(\mathbf{W}_{n-1}) &\rightarrow (\mathcal{R}_n, \mathbf{W}_n)
\end{aligned}$$

To simplify this we write it as:

$$\mathcal{R}^{[n]}(\mathbf{W}_0) \rightarrow \mathbf{W}_n$$

Similarly, a single **cycle** of simulation is the application of the three components of the apparatus $\mathcal{A} = (\mathcal{D}[\tau], \mathcal{B}, \mathcal{R})$ to an initial ensemble, $\mathbf{W}_0$:

$$\begin{aligned}
\mathcal{D}[\tau](\mathbf{W_0}) &\rightarrow \mathbf{W}' \\
\mathcal{B}(\mathbf{W}'') &\rightarrow \mathbf{W}_1 \\
\mathcal{R}(\mathbf{W}') &\rightarrow \mathbf{W}''
\end{aligned}$$

This interleaved application of apparatus components can be simplified to:

$$\mathcal{A}(\mathbf{W_0}) \rightarrow (\mathcal{A}', \mathbf{W_1})$$

where $\mathcal{A}' = (\mathcal{D}', \mathcal{B}', \mathcal{R}')$ which can easily adopt the notation above for n cycles applied to the initial walkers:

$$\mathcal{A}^{[n]}(\mathbf{W}_0) \rightarrow \mathbf{W}_n$$

A final useful construct is the **snapshot** which is the complete state of a simulation:

$$\mathcal{K} = (\mathcal{A}, \mathbf{W})$$

2.2.2 Software Components

In this section we describe concretely the software components that make up the `wepy` framework and how they are integrated. We begin by describing the **simulation manager**, which implements the apparatus described above and handles the reporting of output to the user. We then discuss some unique features of weighted ensemble algorithms that can lead to challenges during data analysis, and tools provided by `wepy` that make this analysis easier.

2.2.2.1 Building and Running Simulations

To run simulations we need two things: an ensemble of initial walkers and a simulation manager. A simulation manager is an object that contains all of the apparatus necessary to move the walkers forward.

The flowchart in Figure 2.1 describes the functioning of a simulation manager acting on a set of initial walkers. First the input walkers have their states propagated by the **runner**, which can be any sort of dynamics engine. The initial release of `wepy` supports OpenMM as a molecular dynamics engine, as well as experimental support for other engines like NAMD. Because the runner propagation is typically extremely compute intensive we also provide another **work mapper** component, which can be customized for different computing environments. `Wepy` includes a serial implementation that can be used for testing on a single core or device, as well as two additional work mappers that use the python built-in multiprocessing module that are suitable for simulations using hardware on a single node. Because the work mapper has been factored out of the implementation of the runner itself these mappers will work for all different MD engines with little modification.

After propagation by the runner, each walker is tested by the boundary conditions function to see if it has met the criteria. If any walker satisfies the boundary condition, the function is free to modify the state of those walkers; this is called a **warping** event. In addition to the new walkers, a set of warping records are produced that describe how the warping event(s) occurred. These records are then made available for generating the final

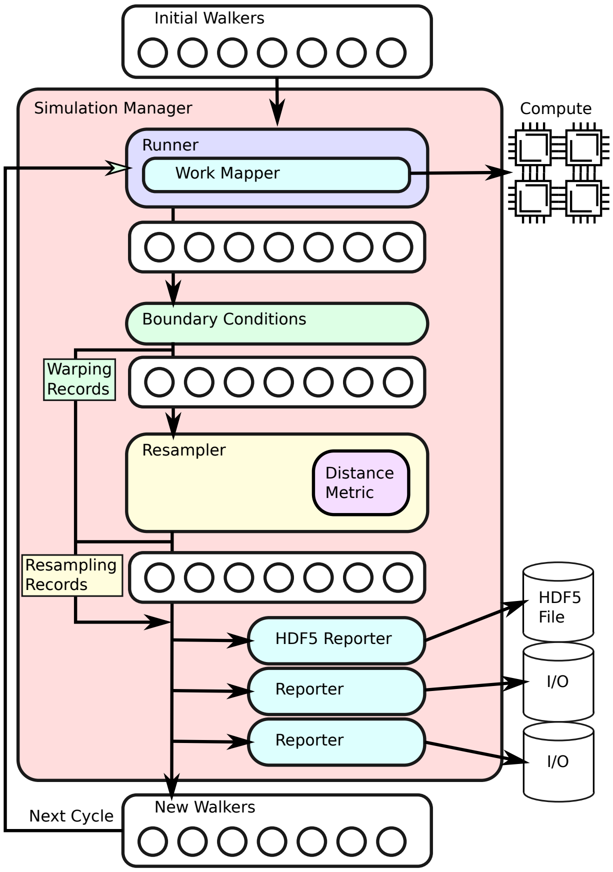


Figure 2.1: Diagram showing the components and flow of data for the main simulation loop of the simulation manager. Dynamics are performed in the Runner component and parallelized onto the compute nodes via the work mapper. The warping and resampling records are shown in the data flow, which are serialized and saved by an HFD5 reporter. Other reporters can be defined to record simulation data, which is indicated by the I/O elements.

data structure and any other report that requests them. In wepy two kinds of warping events are recognized: **continuous** and **discontinuous**. A discontinuous warping event is one where dynamical variables of the walker are modified, e.g. restarting walkers at the initial state after reaching a target state. A continuous warping event is one in which none of those dynamical variables are modified. This could be an auxiliary attribute like a "color" after a walker passes through some boundary [115], or none at all in which case only a record will be produced.

After the boundary conditions are applied the resampler resamples the walkers. Again a collection of records is produced for the resampling, however unlike the warping records a record for each walker is produced every cycle. These resampling records contain critical information about the lineages of walker states that is necessary for reconstructing continuous trajectories. It is useful to use a diagram to depict the histories of each walker as they pass through these stages in a single cycle, an example of which we show in Figure 2.2. Finally, at the end of a cycle the walkers and records are passed off to an arbitrary number of **reporters** which can do whatever I/O is necessary or desired.

While reporters can be customized for a particular simulation, there are a number of useful reporters included with wepy. The WepyHDF5 reporter generates HDF5 output files, which are a major component of wepy and will be discussed on their own later in relation to analysis. The rest of the reporters are designed to give real-time insights to potentially long running simulations as well as to give an accessible summary of the simulation results. The dashboard is an executive summary of the simulation provided in a simple plain-text file that is formatted in emacs org-mode that makes it easy to read a potentially large output file in a hierarchical folding manner. The walker ensemble reporter is for visualizing the current state of the walkers simply outputs a reference topology file and a DCD trajectory file that can be used for visualizing in any of the main 3D molecular visualization programs. While these snapshots of simulations show only transient information they are useful to indicate rough sketches of progress or for debugging purposes. Finally, the resampling tree output

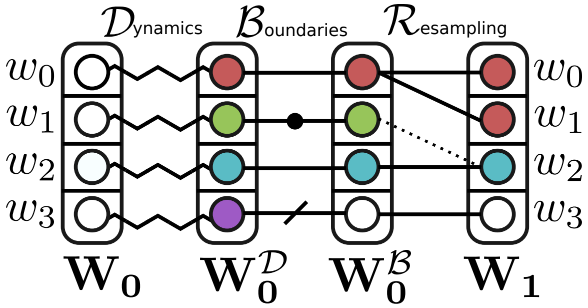


Figure 2.2: A walker history schematic. The vertically stacked boxes are the ensembles of walkers with labels below. The index of the "box" is the walker index in that ensemble. The color of the circle inside the box represents the state of the walker, where we treat white as the initial state of the simulation. The jagged lines indicate propagation of the walker state through the dynamics of the runner. The lines for the boundary condition step indicate whether a warping event occurred, where the dot indicates a continuous warp occurred and the slash indicates a discontinuous warp, in this case returning the walker state to the initial conditions. The lines in the resampling phase show cloning and merging, where a solid line indicates that a child inherits the state of its parent and a dashed line indicates a transfer of weight to another walker. In this example w_0 has cloned itself to produce two child walkers while w_1 has been "squashed" and its state forgotten. It is "merged" and its weight was added to the resultant w_2.

will generate a GEXF formatted XML file that shows the “family tree” of all walkers in the simulation. This is useful for helping understand both when and where resampling and warping events are taking place. These reporters are provided only for convenience as all of this information can be generated from `WepyHDF5` files.

2.2.2.2 Data Format and Analysis

The weighted ensemble algorithm is built on branching trajectories: simulations that have a single starting point may have multiple ending points. We call these kinds of tree-like branching trajectories **non-linear** and a diagram of the difference between them and more traditional linear trajectories can be seen in Figure 2.3. In `wepy` we call a locally-linear selection of frames from the entire non-linear dataset a **trace**. These are shown in Figure 2.3 as the solid colored lines drawn next to the frames they encompass. Non-linear trajectories introduce additional complexity and typically require modifications to common time-dependent analyses.

Firstly, visualizing a single linear trajectory requires a selection step and the input of information. For instance in Figure 2.3 the single red and blue traces for the linear trajectories is trivial, whereas there is some redundancy in the non-linear trajectories from the tree. In a non-linear tree, we can choose a frame for which we are interested in its history and recover its **lineage** as a trajectory. These linear trajectories can then be exported to a common format (e.g. CHARMM DCD) and visualized.

Secondly, state network models like Markov State Models (MSMs) are a common method for representing and understanding large amounts of simulation data. This is a perfect match for weighted ensemble simulations, as the trajectory segments are typically generated using unbiased dynamics. However, current tools for constructing MSMs do not support construction of the transition matrices from non-linear trajectory trees. Fundamentally, the problem involves avoiding the double-counting frames when counting the transitions for lag times greater than our cycle length (τ). A comparison for the solution for the non-linear case is compared to the linear case in Figure 2.3, which is implemented in `wepy`.

Thirdly, when calculating free energies of macrostates the weights of trajectory observations must be taken into account. This simply amounts to doing weighted sums rather than counts for macrostate bins, and requires careful association of trajectory frames to the instantaneous values of trajectory weights.

Finally, observations in different trajectories cannot be assumed to be statistically independent. Take for example a set of observed warping events taking place where a ligand molecule reaches the threshold of unbinding for a receptor. Each single instance contributes to the overall rate of unbinding proportional to its weight via the Hill relation [63]. However, the calculation of macroscopic observable uncertainties (like the probability of an unbinding event) is complicated by the duplication of single observations via cloning. The degree to which two observations can be seen as "independent" depends on the time point of their last common ancestor.

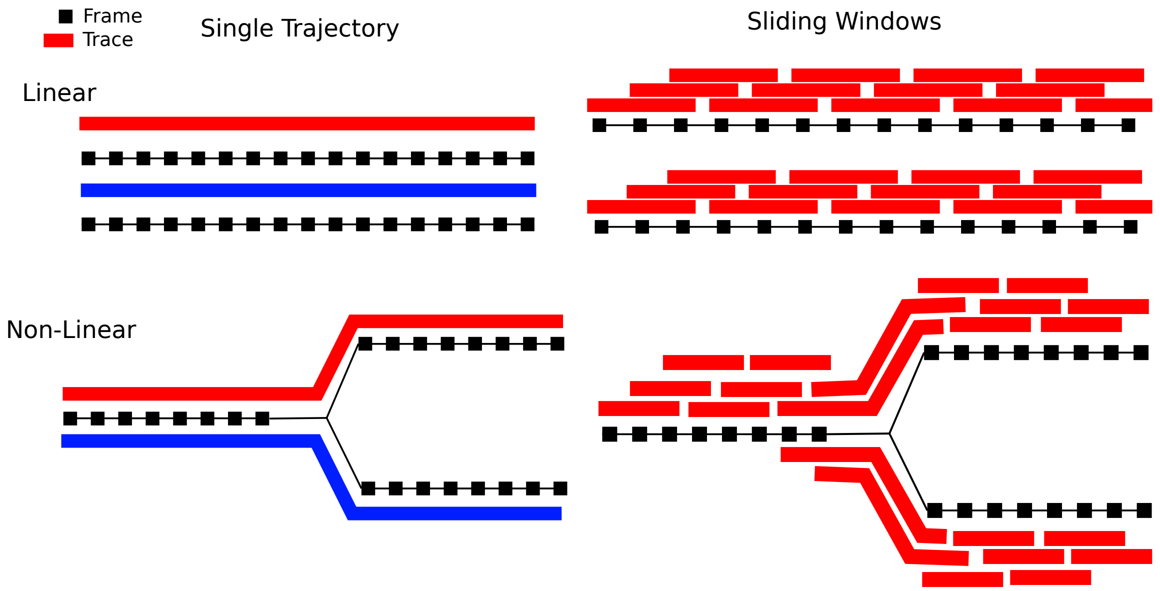


Figure 2.3: Non-linear trajectories comparison. The black squares indicate single frames which are connected in time by the thin black lines. The larger lines indicate a locally linear trajectory, here called a "trace".

To deal with these fundamental differences in the abstract structure of data generated by a WE simulation `wepy` implements a new mechanism for storing trajectory data. To accomplish this we have designed and implemented a storage layer using the common HDF5 format, which is suited to storing large amounts of heterogeneous data. In this implementation all data results from a simulation are contained in a single monolithic binary file. While the data can be accessed with any tool that supports HDF5, in `wepy` we provide an extensive API for creating, accessing, querying, processing, and transforming the data at a useful semantic level.

In the `WepyHDF5` format, we bundle together the three critical and interdependent data pieces into a single file: walker data (including weights and states), resampling data, and boundary condition warping data. This combination is what allows us to generate views of linear trajectories from non-linear data. The resampling data informs the parent-child relationships between walkers and the warping data alerts to the presence of discontinuities in dynamics of these walker lineages. These primitives solve most of the major problems listed above for dealing with non-linear trajectories with tools provided in `wepy`.

2.2.3 Creating and Developing Resamplers

As mentioned in Section 2.2.1 resamplers can do whatever they want as long as they satisfy the constraints given in Equation 2.1. There is thus great flexibility in `wepy` for advanced users who wish to design new resampling algorithms. However, in practice not much is developed in a vacuum and much of the functionality between resamplers can be shared. Here we describe two core abstractions that are provided by `wepy` to aid in design and construction of resamplers: decision classes and distance metrics. These two components are the foundations for the two high-dimensional resamplers provided in `wepy`, WExplore and REVO, in addition to other in-progress research resamplers.

2.2.3.1 Decision Classes

The first useful abstraction we identify is the **decision class** (denoted D). We would like to report on the resampling process such that there is no loss of information. While this is not completely necessary and resamplers have the privilege of not divulging how it resampled a given ensemble of walkers, it is rather useful to know *post hoc*. Indeed if we don't have information on how an ensemble of walkers was derived from a former one, then we have no way of connecting them together for visualization or analysis.

One way to represent the resampling process that satisfies these requirements is to model a resampling process as a set of discrete actions applied to each walker. We then require that the resampler generate an **action record** for every walker in a single applicative step:

$$\mathcal{R}(\mathbf{W}) \rightarrow (\mathcal{R}', \mathbf{W}', A)$$

where $\mathcal{R}$ is the resampler, $\mathbf{W}$ is the set of walkers, and A is a list of the actions a_i that were applied to each walker i, where is n is the number of walkers in $\mathbf{W}$:

$$A = (a_0, a_1, ...a_{n-1})$$

In order to support more general situations we allow for this “net” action record to be described as a set of micro-actions, that when applied successively, k times, yield the net action record: $\bar{A} = (A_0, A_1, ...A_{k-1})$. This structure was chosen because it supports a multi-pass iterative approach without losing any information. For single pass approaches we can use a set of micro-actions with size one, i.e. $\bar{A} = (A_0)$. However, the net action record, $\bar{A}$, is all that is needed when analyzing ancestries and for this discussion we can ignore the micro-actions.

This representation is quite general as it leaves the structure and content of the individual action records a_i unspecified. The decision class adds structure to these records by regarding each action record as a decision that was made regarding the fate of a walker. Minimally, this simply means defining a set of decision types (or symbols) and choosing one for each walker.

As an example, for the canonical WE cloning and merging use case we choose the decision symbols: "CLONE", "MERGE", "SQUASH", and "NOTHING". Where "CLONE" indicates to make a copy of the walker, "MERGE" indicates that the state of this walker will be kept and weight from a selection of the "SQUASH" walkers will be donated to it, and "NOTHING" indicates that no action will be taken for this walker. The definition of these symbols is implemented as a python `Enum` that assigns an integer value to each symbol (useful for efficient storage). To encode the meanings of these symbols in a resampler, a decision class uses two methods: `action` and `parents`.

The `action` method can be seen by expanding the previous application of the resampling function into a two-step process:

$$\mathcal{R}(\mathbf{W}) \to (\mathcal{R}', A)$$

$$\texttt{action}_D(\mathbf{W}, A) \to \mathbf{W}'$$

The `clone-merge` decision class in `wepy` covers most use cases and allows for multiple clones from a single walker. In addition to the decision symbol it also requires an additional list of indices indicating the other walkers they target. For “CLONE” the target indices are the locations (and thus implicitly reveal the number of clones) that the newly minted walkers will occupy. For “SQUASH” the target indicates which “MERGE” walker it will donate its weight to. Target indices for “MERGE” and “NOTHING” designate the index of the surviving walker. It is up to the resampler to ensure the integrity and consistency of these records.

As an example, the action record A from the resampling step of Figure 2.2 would take the form:

$$\begin{aligned} A = ((\text{CLONE}, &\quad (0,1)), \\ (\text{SQUASH}, &\quad 2), \\ (\text{MERGE}, &\quad 2), \\ (\text{NOTHING}, &\quad 3)) \end{aligned}$$

This decision class is the only one supported in `wepy` currently, although others are currently under investigation.

The `parents` function is a function which allows for the recovery of the lineage of walkers:

$$\texttt{parents}_D(A) \rightarrow (p_0, p_1, ..., p_{n'-1})$$

where p_i is the index of the walker in $\mathbf{W}$ that is the parent of the walker i in $\mathbf{W}'$, and n' is the number of walkers in $\mathbf{W}'$. For the example above the parents would be $(0, 0, 2, 3)$.

Resamplers can then be equipped with decision classes both in order to aid in generating the records of resampling and identify the decision protocol that it adheres to. This both reduces the amount of code a resampler must implement but reifies an interface such that multiple components can identify a resampler as having certain properties. Reification of the decision class drastically improves the serializability of weighted ensemble simulation data, and thus the interoperability. This can be seen in the straightforward representations of these records into the HDF5 storage format as a simple table of values.

2.2.3.2 Distance Metrics

See Section 1.2.5 for a detailed discussion on the theory and strategies for designing distance metrics. In both WExplore and REVO, a central object is the distance metric, which is used to compare walkers to each other (in the case of REVO) or to compare walkers to a set of "images" that are constructed over the course of the simulation (in the case of WExplore).

Distance metrics are defined as independent objects in `wepy` that can be used to build resamplers, boundary conditions, or for analysis. These can be defined very generally and need not be Euclidean. One example is the characterization of molecular conformations by a "string" of booleans indicating the presence or absence of secondary structure features or intermolecular interactions. While this does not form a vector space it is still able to be compared using metrics based on Jaccard distances like the Tanimoto distance which is frequently used in chemoinformatics for comparing molecular structure features. In fact there is a great diversity of non-vector space distance metrics that could be used to express the goal of a simulation:

- Root-mean square metrics like molecular RMSD [116],
- Mahalanobis distance for characterizing protein surfaces [98]
- Hausdorff distances for characterizing shapes from continuous topologies,
- network/graph structures Wiener index for characterizing network/graph structures again used in chemoinformatics [99].

Distance metrics have a simple and uniform abstraction that is mathematically well studied . Generally, a distance metric can be defined as a function, d, that maps two states, X_i and X_j, in a metric space, J, to a single positive real number [101, 100]:

$$d(X_i, X_j) = d_{ij}; \quad d : \mathbb{R}^J, \mathbb{R}^J \to \mathbb{R}$$

An example of a simple `wepy` distance metric is given below. The necessary details of this are that the class inherits from the Distance super-class and that it implements a method called `image_distance`. With this implementation we are free to run simulations with either WExplore and REVO because they both support this interface.

In practice distance metrics should be defined to reflect the goals of a particular simulation. This meshes well with the `wepy` implementation as one can use any external library available in the vast Python ecosystem. We highlight, for instance, that the `scipy.spatial` library provides over 20 different general purpose distance metrics at the time of writing [117]. Furthermore, more molecular focused analysis tools like `MDTraj` [118], `MDAnalysis` [119], ProDy [120], and our own `geomm` can be leveraged for constructing more complex distance metrics.

```
import wepy.resampling.distances as wepy_dists

from wepy_dists import Distance as Dist

import numpy as np

class XYEuclideanDistance(Dist):

  def image_distance(self,
           image_a,
           image_b):

  xx = (image_a[0] - image_b[0])**2
  yy = (image_a[1] - image_b[1])**2

  return np.sqrt(xx + yy)
```

2.3 Results

In this section we first discuss code examples on how to set up, run (Section 2.3.1), and analyze (Section 2.3.2) a simulation for the test system used in this paper[2]; Lysozyme ligand unbinding in implicit solvent. We then describe all the parameters and details about the simulations that were run, the results of which are briefly discussed. This small experiment is shown to give an example of the kinds of results and analysis that are typical for wepy simulations.

2.3.1 Code for running simulations

Here we give a brief sketch of how these components were constructed and put together into a simulation manager in a python script. The complete code is given in the Supplemental Information. The system and parameter choices will be discussed in Section 2.3.3.

We first need to set up our wepy runner for OpenMM, which requires OpenMM system and integrator objects. Using OpenMM-Systems helper library (installed with wepy) we can easily create a ready-to-go MD system.

```
from openmm_systems.testsystems import LysozymeImplicit
test_sys = LysozymeImplicit()
```

The integrator can be constructed using the OpenMM constructor:

```
from simtk.openmm import LangevinIntegrator
from simtk.unit import kelvin, picosecond

integrator = LangevinIntegrator(
```

[2] Due to typesetting limitations code examples do not necessarily represent good style.

```
                300.0*kelvin,
                1/picosecond,
                0.002*picosecond)
```

where the three arguments are the temperature, friction coefficient and dynamics timestep, respectively. We can then create a runner object that contains everything needed to propagate the system, where the `Reference` platform specifies a cross-platform reference implementation on a CPU. Note that our production simulations for this work were run with the `CUDA` platform.

```
from wepy.runners.openmm import OpenMMRunner


runner = OpenMMRunner(
            test_sys.system,
            test_sys.topology,
            integrator,
            platform='Reference')
```

Second, we need to create the initial ensemble of `Walker` objects from which to start our simulations. We use a `wepy` helper function `gen_walker_state` to generate state objects directly from OpenMM systems.

```
from wepy.runners.openmm import gen_walker_state
init_state = gen_walker_state(
               test_sys.positions,
               test_sys.system,
               integrator)
```

`init_state` is a `WalkerState` object which can be put inside a `Walker`. Because we are using a Langevin integrator which has a stochastic component (required by all weighted ensemble simulations) we can copy the same structure for all starting replicas. Each worker manually has a weight assigned to them; in this case it is a uniform distribution.

```
from wepy.walker import Walker
from copy import copy


n_walkers = 48
init_weight = 1.0 / n_walkers


init_walkers = []
for i in range(n_walkers):
    walker = Walker(copy(init_state),
                    init_weight)
    init_walkers.append(walker)
```

Third, we construct a resampler to use. Both the WExplore and REVO resamplers require a metric, which is a way of measuring the distance between two `Walkers`. Details regarding distance metrics are discussed in more detail in Appendix 2.2.3.2. For receptor-ligand based systems there are distance metrics already included in `wepy`:

```
import wepy.resampling.distances as wepy_dists
from wepy_dists.receptor import UnbindingDistance


distance = UnbindingDistance(
                   ligand_idxs=lig_idxs,
                   binding_site_idxs=bs_idxs,
                   ref_state=init_state)
```

where lig_idxs and bs_idxs are the atomic indices of the ligand and binding site in the system. A user can also easily make their own distance metrics, a recipe for which is shown below:

```
import wepy.resampling.distances as wepy_dists

from wepy_dists import Distance as Dist

from geomm.rmsd import calc_rmsd
from geomm.superimpose import superimpose

class MyUnbindingDistance(Dist):

    def __init__(self,
                 ligand_idxs,
                 binding_site_idxs):

        self.lig_idxs = ligand_idxs
        self.bs_idxs =  binding_site_idxs

    def image_distance(self,
                       state_a,
                       state_b):

        sup_b = superimpose(
                     state_a[’positions’],
                     state_b[’positions’],
                     idxs=self.bs_idxs)
        lig_rmsd = calc_rmsd(
                       state_a['positions'],
                       sup_b,
                       idxs=self.lig_idxs)


        return lig_rmsd
```

The dependencies here are the `Distance` base class and some helper functions for performing geometric operations on arrays from our custom library `geomm`. The distance object is created in a similar way, except we don't need the reference state (which is needed in `UnbindingDistance` for performance reasons):

```
my_distance = MyUnbindingDistance(
                      lig_idxs,
                      bs_idxs)
```

Fourth, we construct the resampler we are going to use. Here we create an instance of the `WExploreResampler` using the `distance` and `init_state` objects we created earlier as well as some additional algorithm parameters (discussed in Section 1.2.6.1):

```
import wepy.resampling.resamplers as wepy_resamplers
from wepy_resamplers.wexplore import WExploreResampler


NUM_REG = (10, 10, 10, 10)
REG_SIZE = (1, 0.5, 0.35, 0.25)
resampler = WExploreResampler(
                    init_state=init_state,
                    distance=distance,
                    max_n_regions=NUM_REG,
                    max_region_sizes=REG_SIZE,
                    pmin=1e-12,
                    pmax=0.5)
```

Fifth, to set up a non-equilibrium ligand unbinding simulation, we will construct boundary conditions that capture walkers as they cross into the unbound state. `Wepy` also comes with some built-in modules for receptor-based boundary conditions, which we can import and parametrize as follows:

```
import wepy.boundary_conditions as wepy_bc
from wepy_bc.receptor import UnbindingBC

CUTOFF = 1.0 # nanometers

bc = UnbindingBC(
           cutoff_distance=CUTOFF
           initial_state=init_state,
           topology=json_top,
           ligand_idxs=lig_idxs,
           receptor_idxs=prot_idxs)
```

where `json_top` is an internal JSON-based topology format in `wepy`, more information on this is given in the Supplemental Example script. Again this is easy to customize for your application; we show a simplified example of a custom Boundary Condition below. The `warp_walkers` function computes the minimum distance between the ligand and the receptor atoms for each walker and if it exceeds a threshold we "warp" it by replacing that walker state with the initial bound state, while keeping its weight constant.

```
import wepy.boundary_conditions as wepy_bc
from wepy_bc.boundary import BoundaryConditions as BC
```

```
class MyUnbindingBC(BC):

  def __init__(self,
               initial_state,
               ligand_idxs,
               receptor_idxs,
               cutoff_distance=1.0):

    self.initial_state = initial_state
    # etc
    # ...

  def warp_walkers(self,
                   walkers,
                   cycle):

    new_walkers = []
    for walker in walkers:

      dists = compute_distances(
                             walker,
                             self.ligand_idxs,
                             self.receptor_idxs)

      if (min(dists) >= self.cutoff_distance):
```

```
      warped_walker = Walker(
                         self.initial_state,
                         walker.weight)


      new_walkers.append(warped_walker)
      else:
        new_walker.append(walker)


    return new_walkers, [], [], []
```

Finally, we assemble these components into a simulation manager:

```
from wepy.sim_manager import Manager


sim_manager = Manager(
                  init_walkers,
                  runner=runner,
                  resampler=resampler,
                  boundary_conditions=bc)
```

Once the simulation manager is constructed all we need to do now is to tell it to run. A simulation for 10 cycles, each having 10 000 steps per walker per cycle is run as follows:

```
final_walkers, final_apparatus = sim_manager.run_simulation(
                                                      10,
                                                      10000)
```

This returns the walkers at the end of the simulation along with the final state of the runner, resampler, and boundary conditions which are contained in `final_apparatus`.

Another important component is the use of reporters. In addition to the walkers and apparatus returned by the simulation manager functions, wepy supports a plugin system to output data as the simulation is progressing. In the following example we will show you how to use two different reporters: one for creating the HDF5 files and another that produces a plain-text dashboard file every cycle to show the progress of long running simulations in a high-level overview. If you have the resampler and boundary conditions already constructed it is very simple to make the HDF5 reporter. This will work out of the box, but should likely be customized as the default is to save all data, including the velocities, at each step.

```
from wepy.reporter.hdf5 import WepyHDF5Reporter


h5_reporter = WepyHDF5Reporter(
                      file_path='myresult.wepy.h5',
                      topology=json_top,
                      resampler=resampler,
                      boundary_conditions=bc)
```

The dashboard is composed of different sections for the different components. There is a generic one that displays the number of cycles run, walker weights, and total simulation time, and component specific sections for information on resampling, boundary conditions, etc.

```
from wepy.reporter.dashboard import DashboardReporter, BCDashboardSection


from wepy.reporter.openmm import OpenMMRunnerDashboardSection


import wepy.reporter.wexplore as wexplore_reporter


from wexplore_reporter import WExploreDashboardSection
```

```
import wepy.reporter.receptor as receptor_reporter

from receptor_reporter.dashboard import UnbindingBCDashboardSection

runner_dash = OpenMMRunnerDashboardSection(runner)

wexplore_dash = WExploreDashboardSection(resampler)

bc_dash = UnbindingBCDashboardSection(bc)

dash_reporter = DashboardReporter(
                     file_path=’myresult.dash.org’,
                     step_time=0.002*picosecond,
                     runner_dash=runner_dash,
                     resampler_dash=wexplore_dash,
                     bc_dash=bc_dash)
```

These reporters are attached to the simulation manager at creation time:

```
sim_manager = Manager(
                 init_walkers,
                 runner=runner,
                 resampler=resampler,
                 boundary_conditions=bc,
                 reporters=[h5_reporter,
                            dash_reporter])
```

As mentioned in Section 2.2.2.2, the HDF5 reporter is of utmost importance and is a purpose-built fully-featured storage format implemented in HDF5. Saving your data in the HDF5 format will let you use an extensive API designed specifically for manipulating weighted ensemble data (the `WepyHDF5` class in the `wepy.hdf5` module) as well as allowing lower-level manipulations via libraries like `h5py`. Furthermore, a fairly comprehensive analysis toolkit is made available in the `wepy.analysis` module. This toolkit makes it easy to transform and structure data to interoperate with other analysis toolkits like `scipy` [117], `dask` [110], `mdtraj` [118], and `gephi` [6]. In the next section we show some examples of how to leverage these tools to perform common analyses and visualizations.

2.3.2 Code for Analysis & Visualizations

2.3.2.1 Probability Distributions

One of the most common ways to visualize simulation data is to project structural data to a small number of dimensions and then plot this as a probability distribution. In practice, free energy profiles are computed by binning the projection domain and then computing the weighted histogram over those bins by summing the weights of the samples in those bins. The bin values are then normalized to get a probability distribution, which is then transformed by $-ln(p)$ to get a free-energy like value. Typically the term free-energy refers to probability distributions over equilibrium ensembles. However, for convenience here we refer to probability distributions transformed as described as “free-energy” or more specifically “non-equilibrium free-energy” even if the underlying ensemble is not necessarily an equilibrium one. In a single linear MD trajectory the frames are equally weighted, but in an ensemble of walkers in WE the weights of each frame can be different and vary over time. As such, the trajectory coordinate data is associated with the weights in the `wepy` HDF5 file.

Here we show how to create 1-D free energy profiles for each experimental group projected onto the ligand RMSD relative to the bound pose (see Fig. 2.4). The process is to first compute the ligand RMSD for all of the simulation frames as an “observable” and then to compute the free energy profiles which can be plotted with a tool like matplotlib [121].

To compute the ligand RMSD observable we open the file with the `WepyHDF5` API in Python:

```
from wepy.hdf5 import WepyHDF5


wepy_h5 = WepyHDF5('my_data_file.wepy.h5',
                   mode='r+')


wepy_h5.open()
```

Then we define a function for computing RMSD that will be mapped over all of the data:

```
import numpy as np
from geomm.rmsd import calc_rmsd


def calc_rmsd_observable(fields):


  rmsds = []
  for frame in fields['positions']:


      rmsd = calc_rmsd(
                 reference_positions,
                 frame,
                 idxs=ligand_idxs)


  return np.array(rmsds)
```

We then apply the function to the data:

```
observable = wepy_h5.compute_observable(
                             calc_rmsd_observable,
                             [’positions’],
                             (),
                             save_to_hdf5=’lig-rmsd’)
```

where the second argument asks to retrieve the relevant data from each frame, which here is just the positions. This also saves the observable into the database as the field ‘lig-rmsd’. By default the `calc_rmsd_observable` function is applied in serial. In `wepy` we also provide a distributed version which can connect to a `dask` cluster server for distributed parallelism on large computing clusters (see documentation).

Now we can create the free energy profile for this observable. One intermediate step though is to use another representation called the “contig tree”, which makes combining multiple contiguous simulations (such as when a simulation is restarted) much easier to analyze. Construction of a contig tree requires: 1) a dataset, 2) a “decision” (used internally by the resampler see Appendix 2.2.3.1 for more details), and optionally 3) the boundary condition class that was used for the simulation if any.

```
from wepy.analysis.contig_tree import ContigTree

decision = WExploreResampler.DECISION

bcc = UnbindingBC

contigtree = ContigTree(
                     wepy_h5,
                     decision_class=decision,
                     boundary_condition_class=bcc)
```

With the contig tree constructed, we can then feed it to the profiler, bin the domain, and calculate the free energies for each.

```
from wepy.analysis.profiles import ContigTreeProfiler


profiler = ContigTreeProfiler(contigtree)


bin_edges = profiler.bin_edges(
                           'auto',
                           'lig-rmsd')


fe_profile = profiler.fe_profile(
                               0,
                               'lig-rmsd',
                               bins=bin_edges)
```

A line plot of `fe_profile` against `bin_edges` is shown for each simulation type in Figure 2.4 and discussed in Section 2.3.4.1.

2.3.2.2 Visualizing Ligand Unbinding Events

To make rate estimates for a process or to analyze the transition path ensemble, we need an efficient way to examine the pathways from one basin to another. There are tools in `wepy` to help analyze this kind of boundary condition data, which we call the "warping" records. The data for all the warping events can be found in the HDF5 and is accessed through either the `WepyHDF5` object or the `ContigTree`. Here we can make a `pandas` dataframe table for run 0, which can easily be exported to any number of formats.

```
warp_df = wepy_h5.warping_records_dataframe([0])
```

Each row of this table contains the index, weight, and point in time that the event took place.

To get trajectories from these warping events to the starting structure we use the following functions on a `Contig` object which is a single simulation from a `ContigTree`:

```
contig = contigtree.span_contig(0)


with contig:
  warp_points = contig.warp_contig_trace()


  lineage_trace = list(contig.lineages(warp_points))[0]


  contig_h5 = contig.wepy_h5


  trace_fields = contig_h5.get_trace_fields(
                                   lineage_trace,
                                   ['positions',
                                    'box_vectors'])


  traj = contig_h5.traj_fields_to_mdtraj(trace_fields)


traj.save(filepath)
```

On the first line we first generate a linearized `Contig` from the `ContigTree` which is necessary for traversal. Then we open a context for the contig to open the HDF5 file where we can then access the simulation data. We then query for a “trace” of all frames in which a warping event occurred as the `warp_trace`. From each of these events we then produce the individual histories (“lineages”) of that walker to the initial state. For this example we only choose to look at one of these, which is set as `lineage_trace`. We now want to actually retrieve the simulation data to be able to visualize and perform analysis on it. To do this we use the lineage trace and choose which attributes we want. For this example we get only the positions and box vectors since our purpose is to generate visualizations. Now that we have the `trace_fields` for this trajectory we can easily convert to an `mdtraj` object which can then be used to save the trajectory to a file format visualization software can read.

2.3.3 Simulating Lysozyme Ligand Unbinding

To showcase the use of `wepy` for performing simulations of rare events in real systems we simulate the small molecule p-xylene unbinding from the T4 lysozyme L99A mutant protein; which we refer to henceforth as "lysozyme". Lysozyme is a common model system for ligand unbinding studies both experimentally and computationally [122]. Lysozyme unbinding pathways have been simulated through a variety of enhanced and brute-force methods [123, 124, 125, 126]. Here we simulate lysozyme interacting with the p-xylene ligand in implicit solvent where the event of interest, ligand unbinding, is not trivial to observe but is still tractable with straightforward MD simulations.

The system was prepared in an OBC GBSA implicit solvent, using Amber ff96 for the protein force-field and a GAFF and AM1-BCC parametrization of the ligand. A Langevin integrator was used with a 2 fs time step, temperature of 300 K and a friction coefficient of 1 ps^{-1}. Here we determine a target p-xylene residence time in implicit solvent using a total of 3.385 µs of straightforward MD simulation, where p-xylene is “warped” back to the binding site upon unbinding a total number of 11 times. This resulted in an unbinding rate of 3.25 $\mathrm{\mu s}^{-1}$, which we use as a target rate for comparison with both WE simulations and trajectory ensembles using `wepy` without resampling. The test groups that were simulated are as follows:

- ensemble of 48 walkers with no resampling i.e. straightforward (SF group)

- ensemble of 48 walkers with the REVO resampler (REVO group)
- ensemble of 48 walkers with the WExplore resampler (WExplore group)

For each group, 4 independent simulations were run, each for a total sampling time (summation across all replicas of ensemble) of 1 µs. All simulations used the same boundary conditions criterion, which is that the minimum of all ligand-protein interatomic distances is greater than 1.0 nm. Boundary conditions are checked at the end of every cycle, which was $\tau = 20$ ps for every group. The weighted ensemble simulations (REVO and WExplore) employed the following parameters: a minimum walker weight of 10^{-12}, and a maximum walker weight of 0.5. Both REVO and WExplore used the `UnbindingDistance`, where the distance between two walker structures is computed as the distance between their ligands after alignment of their binding sites to the initial starting structure. This is included in `wepy` and has been used with success in other ligand release simulations [39, 47, 102, 80, 103].

For the WExplore simulation the same parameters from previous publications [39, 47, 102] were used; four region hierarchy levels with cutoffs d = 10 Å, 5 Å, 3.5 Å and 2.5 Å with a maximum of 10 sub-regions per parent region. For REVO [93] the following parameters were used: a characteristic distance of 1 Å, merge distance 2.5 nm [3], distance exponent 4, and we use the weight-based importance function as described in Eq. 1.2.

2.3.4 Analysis of Simulation Results

2.3.4.1 Probability Distributions

Figure 2.4 shows a series of plots of free energies of each simulation group projected onto the RMSD of the ligand to the starting pose as a function of aggregate simulation time. Again, the "free-energies" here are more accurately the negative logarithm of the non-equilibrium

[3] The large value chosen here allowed overly dissimilar walkers in the ensemble to be merged. We expect that a smaller value for this quantity could improve REVO performance, e.g. 2.5 Å.

probability distributions because we are warping unbound walkers back to the starting position. Accordingly, we see a local free energy minimum at RMSD close to 0 Å but no corresponding minimum in the unbound state.

REVO and WExplore reach much larger ligand RMSD values than the straightforward (SF) simulations. Both WExplore and REVO observed close to 5 nm RMSD values at 0.052 µs total sampling time (Figure 2.4B), whereas SF has only reached ligand RMSD values of around 1.7 nm. The probability of large RMSD states tends to be underestimated early on, as can be seen by comparison between panel A and B and the final estimates in panel D.

By the end of the simulations (Figure 2.4D) all the profiles are very similar between the different groups until around 4 nm. The curves for WExplore and REVO beyond that are noisy because of the difficulty in reaching very large distances (e.g. 7 nm) without reaching the unbinding boundary condition. In general, note that early time predictions of weighted ensemble free energy profiles can differ significantly from equilibrium free energy profiles. These should instead be viewed as direct estimates of a conditional time-dependent probability distribution: $P(x, t|x_0, t_0)$, or the probability of being at a point x and time t, given that we began at point x_0 at time t_0. While we expect these to converge to equilibrium probabilities only for $t \rightarrow \infty$, we can often learn valuable information from the tails of these distributions.

We note that a peculiar artifact in this system is that the p-xylene ligand sticks strongly to the surface of lysozyme, and breaking out of the binding pocket was much easier than leaving the surface of the protein. This likely explains high-variance ligand RMSD values above 4 nm and performance could be improved by incorporating the latter process (leaving the surface) more directly into the distance metric that governs the resampling process.

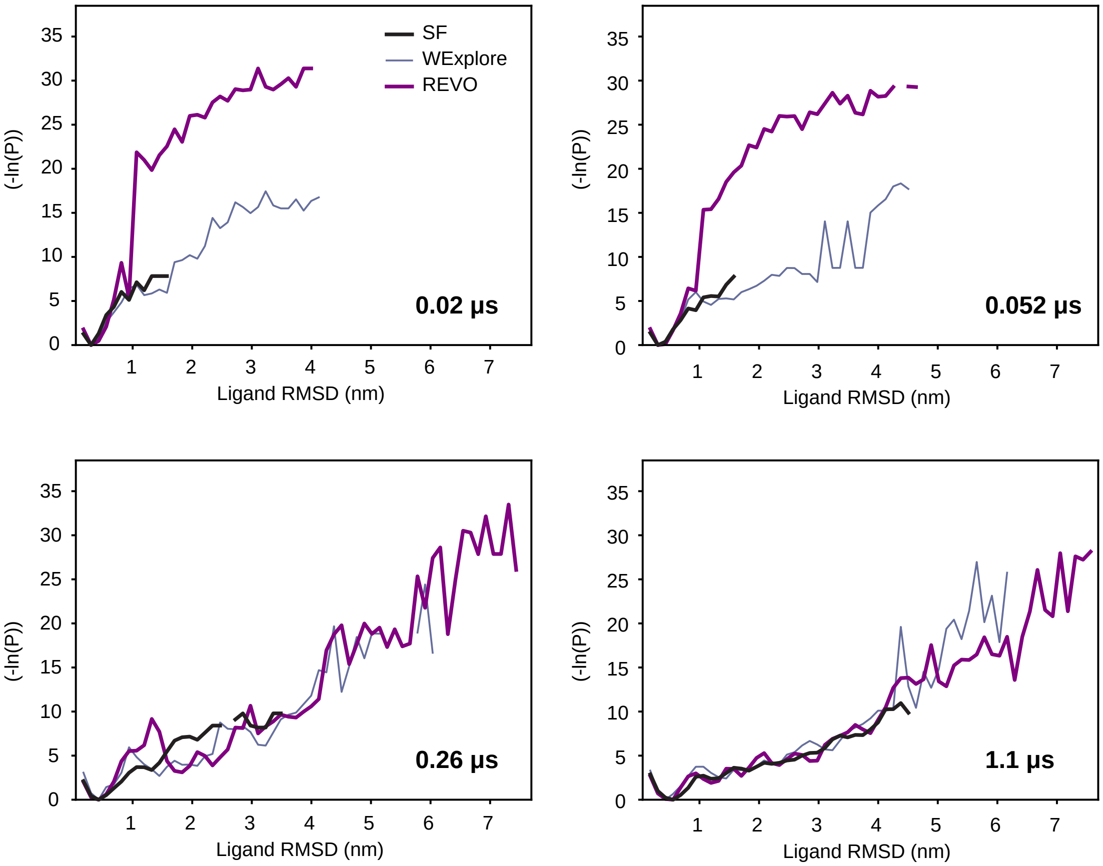


Figure 2.4: Series of $-\log(p)$ distributions for ensemble simulation groups at different time points. The time shown in the bottom right of each panel is the total amount of sampling across all replicas in the ensemble.

2.3.4.2 Unbinding Events & Trajectories

A primary interest in the study of rare events in non-equilibrium systems is to understand the kinetics of transition paths. While we can use a variety of simulation methods to estimate the rate constants associated with events, it can be much more difficult to obtain accurate data on the actual mechanistic determinants of these rates. Primarily this is the structural details of transition states, but all structures along a path can potentially be useful for design purposes. Transition paths obtained with path-sampling methods, like WE, are especially meaningful as the Hamiltonian is un-perturbed throughout the sampling process. Here we show how `wepy` can easily obtain and analyze continuous, unbiased trajectories of ligand

Table 2.1: Unbinding events and final rate estimates of simulations

Group	Num Warps
SF (numerical target)	11
SF (ensemble)	39
WExplore	486
REVO	4337

unbinding.

Table 2.1 shows that both REVO and WExplore outperform the SF methods in terms of the absolute number of unbinding events that are observed. More importantly though is that the initial times of the first observations for REVO and WExplore are much faster than the SF ensemble simulations. All replicates for REVO and WExplore simulations have exit points within the first 100 ns, while the first among SF simulations takes about 3 times as long (at 300 ns) and is highly variable among replicates. The slowest SF simulation does not observe an unbinding event until around 700 ns. This highlights the utility of high-dimensional resamplers like WExplore and REVO for generating observations of rare events with a more modest investment of computing power. In Figure 2.5 we show structures along the first unbinding paths generated by the SF and REVO simulations. From this we can see that the REVO simulation (panel B) takes a much shorter, more directed path to unbinding compared to the SF one (panel A). The ligand in the SF simulation can be seen to move around to multiple other locations on the surface of the protein before ultimately unbinding.

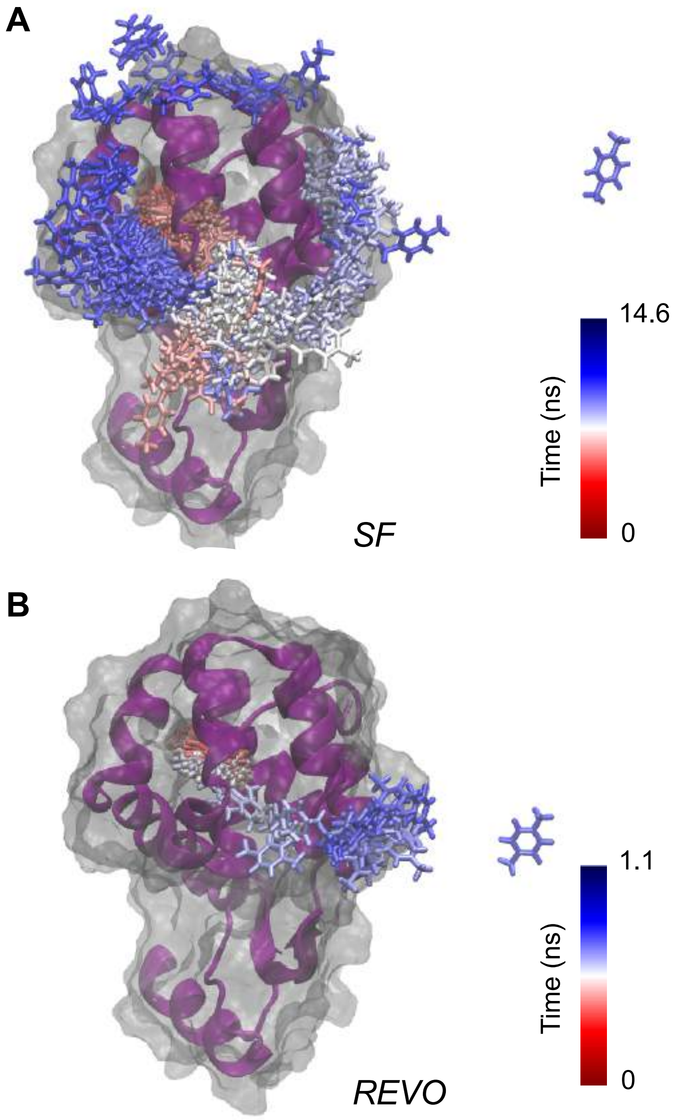


Figure 2.5: Ligand unbinding trajectories. These 3D renderings of lysozyme protein (from the trajectory seed structure) showing the positions of the p-xylene ligand from the bound positions (red) to unbound positions (blue). A cartoon representation of lysozyme is in purple surrounded by a grey surface representation. Positions of the ligand are shown for the first observed unbinding trajectory from each of the two simulation groups shown at the end of each Weighted Ensemble cycle (τ, i.e. every 20 ps). **A**) An unbinding trajectory from the SF group (no resampling). **B**) A trajectory from the REVO group.

2.4 Discussion

2.4.1 Successes

Wepy is a useful and flexible implementation of advanced weighted ensemble simulations with a growing number of applications [80, 93, 103, 127]. In our experience wepy has been particularly useful in three major ways. First, it has made the prototyping of new methods very easy even for researchers with little experience in programming or the Python language. The initial goal of wepy was to simplify and modularize the original implementation of the WExplore algorithm [92]. Following this the REVO algorithm was fully prototyped, tested, and eventually used as a resampler for a variety of problems [93, 80, 127]. This process was made much simpler and faster by the sharing of the same infrastructure that was already developed. In addition, the binless nature of the REVO algorithm was actually conceived of partly as a response to the abstractions formalized by wepy.

This highlights the second point: not only does wepy provide useful software to perform simulations but also a common language with which researchers could communicate with each other about their simulations. We have found the use of diagrams like Figure 2.1 and 2.2 to be valuable not only for reasoning about our programs but also in explaining WE and non-equilibrium simulations in general.

And lastly, wepy has made the process of analyzing WE simulations dramatically simpler. The biggest contribution here is the support for out-of-core data-structures (via the HDF5 format) and expressive in-memory representations that reflect the tree-like trajectory structures (i.e. WepyHDF5 and ContigTree). Using this data-structure, wepy also provides facilities for easy analysis and visualization of "resampling trees".

As mentioned in the Introduction (Section 2.1) a big asset to the weighted ensemble method is that the microscopic trajectories are generated using the unbiased Hamiltonian. In wepy we ensure that this data can be fully leveraged however the user sees fit. One major use case for this data is the construction of Markov State Models (MSMs) with long lag-times, which in turn can be used to predict steady-state probabilities or identify transition states. Although not mentioned here, wepy provides facilities for constructing macrostate models (like Conformation State Networks (CSNs) and MSMs).

Others are encouraged to use, share components for, or even contribute to wepy which is open source with an MIT license. The source code is currently hosted on github (https://github.com/ADicksonLab/wepy) and documentation is currently available at https://adicksonlab.github.io/wepy. At the time of publication the 1.0 release of wepy has been made and has been archived and given a persistent identifier from Zenodo.org (DOI: 10.5281/zenodo.3973431).

2.4.2 Opportunities for Future Development

While wepy provides many essential and novel features there are areas for potential improvements that could make it even more widely useful. Currently, it has not yet been integrated with major MD engines such as GROMACS [128], CHARMM [129], Amber [130] and Desmond [131], and there is currently only experimental support for NAMD [132] and ASE [133]. We note that the architecture of some of these engines makes it difficult to interface without going through a UNIX-like system environment (a difficulty for all software using these tools). We note that OpenMM allows for the use of force fields native to each engine (e.g. CHARMM, AMBER) to be used within it, so the issue is more about choice of implementation rather than the content of the simulations.

Secondly, the priority for wepy developers has been on implementing and prototyping new adaptive and high-dimensional resampling algorithms rather than implementing standard static binning methods or accelerated WE [115, 76]. Fortunately, many of these methods are not complex to implement as resampler objects in wepy and we hope that these will either be included in future releases or as standalone libraries. We note that a goal between WESTPA [113] and wepy developers is to enable a modular program design, where resampler objects could be used interchangeably between the two WE implementations.

Thirdly, for protein and other macromolecular simulations `wepy` resorted to using an *ad hoc* serialization format in JSON for molecular topologies (the schema of which was borrowed from an internal representation in the `mdtraj` HDF5 implementation). In principle `wepy` is agnostic to topology formats but in practice this is an extremely important component in building simulations and performing analyses. Despite there being a large number of software packages implementing in-memory representations of molecular topologies, there are no formats suitable for serialization and communication between software. We encourage users to consider the merits of the JSON topology used in `wepy` but by no means recommend it as a general purpose standard, nor do we wish it to become a *de facto* standard.

Lastly, while the use of an HDF5 based file format has proven to be a good choice for many reasons, it is currently not supported natively by any molecular visualization software. Thus, trajectories must be converted and duplicated into separate files with a supported format (a simple task using the integration with the `mdtraj` library). In our work flows we treat these files as temporary intermediates, however this can bloat the necessary disk space needed as well cause some time delays when (re)generating them. We note though that visualization tools could benefit immensely by adopting a random-access format like HDF5 which would allow for visualizing single trajectories which do not fit in memory; a problem we frequently encounter when attempting to view very long trajectories.

CHAPTER 3

LIGAND RELEASE PATHWAYS IN FKBP OBTAINED WITH WEXPLORE: RESIDENCE TIMES AND MECHANISMS

3.1 Introduction

Computational studies of ligand binding are dominated by methods that consider only the endpoints of the process. Docking is widely used to determine bound ligand poses and to estimate binding free energies using empirical scoring functions, and can be conducted on a large scale to facilitate receptor-based virtual screening [134, 135, 136]. More robust methods to calculate binding free energies also typically consider only the bound and unbound states of the ligand [137, 9, 138, 139]. Accordingly, while much is known about the relationship between ligand structure and binding affinity, much less is known about the general properties of the pathways of small molecule binding.

Simulations employing "steering" forces to encourage ligands to unbind from their pockets [140, 141, 142, 143] are useful to determine general properties of pathways, especially of deeply buried ligands, but they can introduce significant perturbations to unbinding mechanisms. Other techniques such as metadynamics [144, 145, 41] and umbrella sampling [62, 52] can recapitulate binding pathways as well as binding kinetics [60, 46], although a good choice of collective variables is crucial, and often not obvious. Enabled by advances in computer hardware and sampling algorithms, full unbiased ligand binding trajectories have emerged for a number of systems. This includes long, continuous trajectories of ligand binding [48], as well as large numbers of shorter trajectories synthesized by Markov state models [43, 1, 44]. It was shown recently that a dramatic sampling improvement could be obtained by iteratively constructing Markov state models and using these to inform sampling on-the-fly, with trajectory management steps happening on the 10 ns timescale [45]. Previously a similar method was developed called WExplore [92], which also uses unbiased trajectories that are actively managed on a much shorter timescale (20 ps), but does not rely on a Markovian assumption. WExplore defines a set of regions in a high-dimensional order parameter space, and uses trajectory cloning and merging steps – as in the weighted ensemble algorithm [69] – to improve the sampling of low-probability regions, such as transition states that lie between two high-probability basins. In this work we apply WExplore to protein-ligand systems for the first time, allowing for a more thorough exploration of ligand unbinding pathways.

Previous ligand binding simulations have provided a mixed picture of general ligand binding mechanisms, particularly in regard to the extent of surface diffusion prior to ligand release. Trajectories of dasatinib binding to Src kinase from Shan et al [48] have shown extensive surface diffusion prior to binding, involving almost the entire surface of the protein (this behavior is referred to herein as the "surface diffusion model"). Moderate surface diffusion was found in trajectories of benzamidine binding to trypsin [43]. In contrast, some enhanced sampling restraint potentials, such as funnel metadynamics [41], do not allow for substantial surface diffusion in their predictions of the ligand binding free energy maps. As these reconstructed binding pathways establish first contact with the protein directly at the binding site, we refer to this as the "dartboard model".

We use the WExplore method to examine ligand release distributions for a well-studied ligand binding model system with three ligands: FK506 binding protein (referred to here as FKBP) with three low affinity ligands. FKBP belongs to the peptidylprolyl cis/trans isomerase (PPIases) family of proteins [146, 147]. It has been the subject of much interest since its discovery, with structures and free energies of binding determined to a large number of drug-like ligands, both experimentally [147, 148, 149, 150, 151, 152] and computationally [8]. The interaction of FKBP with a set of low affinity small-molecule ligands has also been characterized, with bound structures and biophysical properties [153]. Further analysis was conducted recently in silico by Huang and Caflisch [1], which determined the landscape of binding and unbinding pathways in great detail, and determined the unbinding kinetics for a series of ligands. This extensive characterization, as well as the fast unbinding time that is tractable to conventional molecular simulation, makes FKBP an ideal test case to demonstrate WExplore sampling applied to ligand binding systems for the first time.

3.2 Methods

3.2.1 Molecular dynamics sampling

The trajectory segments are run using CHARMM [129], through the OpenMM-CHARMM interface in order to utilize GPU hardware, with the CHARMM36 force field for the protein. Ligand parametrization is done using the CHARMM Generalized Force Field (CGenFF) [154, 155]. Solvation, minimization and equilibration are done separately for the three ligands: 4-hydroxy-2-butanone (BUT), dimethylsulfoxide (DMSO) and methyl sulfinyl-methyl sulfoxide (DSS), using structures 1D7J, 1D7H and 1D7I, respectively [153] (Figure 3.1). Each structure is solvated in a cubic box allowing 12 Å of space between the protein and the cell boundaries, resulting in 13623, 13512 and 12350 water molecules for BUT, DMSO and DSS, respectively. One chloride ion is used to neutralize each system. SHAKE is used to constrain the hydrogen atoms along their covalent bonds, with a tolerance of 10^{-8}. Particle-mesh Ewald summation is used for non-bonded interactions with a van der Waals switching function that scales the non-bonded interactions to zero at 10 Å, starting from 8.5 Å.

After solvation, the solvent and ions are minimized using 500 steps of steepest descent

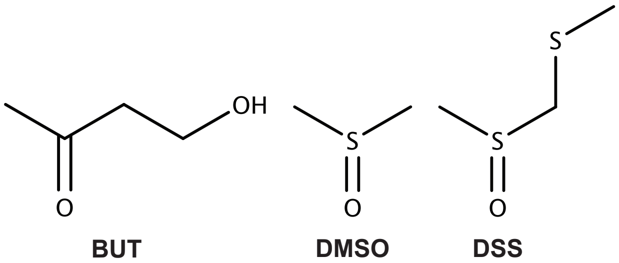


Figure 3.1: **Three low-affinity ligands used in this study.** Left: 4-hydroxy-2-butanone (BUT). Middle: dimethylsulfoxide (DMSO). Right: methyl sulfinyl-methyl sulfoxide (DSS).

and 500 steps of the adopted basis Newton Raphson method. The entire system is then minimized in the same way. The system is then heated from 50 K to 300 K over 100 ps, using 2 fs timesteps, followed by equilibration for 0.5 ns at 300 K. The resulting structure is used as an initial condition for all 48 replicas in the WExplore method.

Langevin dynamics is performed in a constant pressure ensemble with a barostat coupled to a reference pressure of 1 atm, and volume moves attempted every 50 steps. We use a group of eight GPUs, each running an instance of CHARMM with OpenMM 6.3 [111] to run trajectories for the replicas. These trajectories are managed by a Perl script that implements the WExplore method outlined below. This implementation of WExplore will be referred to as 'WExplore-perl' and is distinguished primarily from the implementation of WExplore in `wepy`.

3.2.2 WExplore Algorithm & Sampling

The WExplore method [92] is an extension of the weighted ensemble (WE) algorithm [69] that facilitates usage in high-dimensional spaces, and is particularly useful in simulating low-entropy to high-entropy transitions, such as protein unfolding or ligand release. In the original WE algorithm, sampling of low-probability states is enhanced by defining a set of regions that span the space, and enforcing that all regions are sampled evenly. Multiple copies of the system ("walkers") are run in parallel, and each is given a weight with which it contributes to statistical averages; the sum of weights over all walkers is equal to 1. Periodically throughout the simulation, walkers are merged in over-represented regions and cloned in under-represented regions in order to keep the number of walkers in each region as even as possible. When a walker is cloned, its weight is split up evenly amongst the clones. When two walkers A and B, with weights w_A and w_B, are merged, the resulting walker has weight w_A+w_B, and takes on the configuration of walker A with probability $w_A/(w_A+w_B)$, and otherwise takes on the configuration of walker B.

Recently, we introduced an extension to WE that allows for efficient application to high-dimensional spaces, and allows for the number of walkers to be far less than the number of regions. The method has been described in detail in previous work [92, 75]. Here we discuss the main details of the algorithm, focusing on aspects that are unique to this study. WExplore uses Voronoi polyhedra to define sampling regions in a high-dimensional space. These Voronoi polyhedra are defined by a set of "images", which are conformations of the system, and are defined to all be a critical distance apart (d_c) from each other, using an RMSD-based (root mean squared distance) distance metric. Importantly, these regions are defined using a hierarchy: the top level regions tile the entire space, and these are tiled by smaller regions, which are themselves tiled by even smaller regions, and so on. Each level of the hierarchy has its own value of d_c, with the top level regions having the highest, and the bottom having the lowest. This hierarchical structure allows for replica balancing across multiple length scales: a balancing procedure is first performed at the top level of the hierarchy, and then at each lower level, prioritizing balancing computational effort amongst regions that are the farthest apart. The structure also helps with region assignment: at each level of the region "tree", the distance to each image needs to be calculated for only a small number of branches.

Here we apply WExplore to study small molecule binding and release for the first time. As noted above, the key quantity to run a WExplore simulation is a way of measuring the distance between two conformations of the system. Our distance metric for small molecule binding is defined as follows. First the two conformations are aligned using a selection of residues that are local to the binding site. These are determined as the set of protein residues that are within a cutoff of 8 Å in the crystal structure binding pose. The distance is then determined using the RMSD of the absolute positions of the ligand, that is without any further translation or rotation. This metric captures both internal ligand motions, as well as ligand translations and rotations.

In WExplore, regions are defined dynamically over the course of the simulation, and do not require any *a priori* information about the dynamics of interest. We use a 4-level sampling hierarchy with region sizes of 2.5, 3.5, 5.0 and 10 Å, and a maximum number of images of 10 at each sampling resolution, for a theoretical maximum of 10000 regions. Each sampling cycle consists of all 48 walkers running for 20 ps, followed by region assignment and cloning and merging steps. The aggregate simulation time of a WExplore simulation after 250 cycles is then 240 ns. We run two WExplore simulations for each of the three ligands, as well as a conventional simulation of the same length (48 trajectories, 240 ns total). In total 2.16 μs of simulation is performed.

Two aspects of our implementation are significantly different in previous work. Firstly, to minimize the sampling expenditure on conformations where the ligand is far away from the protein, we apply the following rule: if a walker has a minimum interatomic distance to the protein that is between 5 and 10 Å, it is not allowed to define new regions, and it is excluded from walker cloning and merging operations. Secondly, we institute a minimum and a maximum weight for the walkers, and we do not clone walkers that are below the minimum weight (p_{min}), and do not merge walkers that are above the maximum weight (p_{max}). p_{max} is set to be 0.1, in order to ensure that there is always at least 5 to 10 walkers in the original binding site. In order to efficiently determine unbinding rates, p_{min} should ideally be set close to the probability of the rarest conformations that are still of importance to the unbinding ensemble. In other words, p_{min} should be close to the probability of conformations in the transition state, but since this is not known a priori, we use the following procedure. We initialize p_{min} at $p^i_{\text{min}} = 10^{-40}$, and allow this to grow as follows:

$$p_{\text{min}} = \begin{cases} p^i_{\text{min}} & \text{if } p_{\text{off}}/10 < p^i_{\text{min}} \\ p_{\text{off}}/10 & \text{if } p^i_{\text{min}} < p_{\text{off}}/10 < 1/2N \\ \frac{1}{2N} & \text{otherwise} \end{cases} \tag{3.1}$$

where p_{off} is the sum of the weights of all exited trajectories collected so far. The net effect of this procedure is that even after we have an estimate of the transition state probability, we are still giving the system the resources it needs to make transitions with higher transition state probabilities.

3.2.3 Exit rates

Following previous work using Nonequilibrium Umbrella Sampling [105, 156] and the string method [106], we can determine the rate of ligand release in a WExplore simulation by examining the probability flux into a given basin of attraction, averaged over a period of time. For our purposes, consider basin A as the set of ligand configurations that are within a certain RMSD cutoff of the native binding pose (e.g. 3 Å). Consider basin B as the set of ligand configurations where the closest protein-ligand interatomic distance is greater than another cutoff, d_U, set here to 10 Å. Here, we simulate the non-equilibrium ensemble where trajectories are initialized in basin A and are terminated in basin B, and the exit rates are formally defined as the transition rate from state A to B.

See Section 1.2.7 for a more rigorous explanation of rate calculations from WE simulations.

3.2.4 Clustering

Clustering is performed using MSMBuilder version 3.3.1 [157], using the "RegularSpatial" algorithm [158]. For each ligand we use a set of protein-ligand distances connecting three points in the ligand to the 50 closest heavy atoms in the protein: 150 distances total. We use the Canberra distance metric between the two sets of distances:

$$d(A, B) = \sum_i \frac{|a_i - b_i|}{|a_i| + |b_i|} \tag{3.2}$$

where $\mathbf{a}$ is the set of atomic distances in structure A, and $\mathbf{b}$ is the set of atomic distances in structure B. The Canberra metric is appropriate for our purposes as it emphasizes differences in distances that are small (i.e., close to the binding pocket). A cluster size of 10 is used in each case, resulting in 485, 497 and 565 cluster centers for BUT, DMSO and DSS, respectively.

Table 3.1: Mean first passage times of unbinding for three ligands from FKBP, obtained by WExplore simulation (two replicates), straight-forward simulation, and previous simulations by Huang et. al. [1]. In the first three cases, standard errors are obtained by block averaging the last 50% of the data into 10 evenly spaced blocks, in each case. The experimentally-determined binding affinity is also shown for each molecule.

	Unbinding time [ns]				Binding affinity [μM]
	WExplore1	WExplore2	SF	Huang 2011	Experiment [1]
BUT	4 ± 1	7 ± 2	4 ± 1	8 ± 2	500
DMSO	5 ± 1	13 ± 3	6 ± 1	4 ± 1	20000
DSS	19 ± 5	27 ± 6	22 ± 4	18 ± 3	250

3.3 Results & Discussion

3.3.1 Residence times

As this is the first application of WExplore to a protein-ligand system, we used conventional simulations as a benchmark, both those conducted here and simulations run previously [1]. The final estimates of the MFPT of unbinding are shown in Table 3.1, and show solid quantitative and qualitative agreement between WExplore and conventional simulations conducted here. Averaging over the three simulations: BUT shows the fastest dissociation at 5 ns, DMSO is the next fastest at 8 ns, and DSS is the slowest with a MFPT of 23 ns. Previous simulations documented similar results [1], and the discrepancies are comparable to estimates of standard error. Interestingly, experimental binding affinities only partially correlate with these results [153]. DSS has the highest binding affinity, but only by a factor of 2. The binding affinity of DMSO was found to be 400 times weaker than BUT (and 800 times weaker than DSS), however the residence times of DMSO and BUT are comparable.

Figure 3.2 shows the convergence of the MFPT estimates as a function of time for both WExplore simulations and the conventional simulations. Although all three ligands dissociate quickly enough to be captured by conventional simulation, the slowest dissociating ligand (DSS) provides some insight into the future performance of WExplore simulations for high-affinity ligands. For conventional simulation the first exit point is not observed until after over 104 ns of total simulation, while exit paths are observed in WExplore simulations using 13.4 and 17.3 ns of total simulation, respectively.

3.3.2 Exit point distributions

Even for low affinity ligands, WExplore simulation generates distributions of exit points much more efficiently than conventional MD. Figure 3.3 shows the number of exit points generated as a function of simulation time for both WExplore (the average of two simulations) and conventional MD. In each case, the most points are determined for BUT, the second most for DMSO, and the least points were determined for DSS. This ordering is what would be expected from the exit rates for the three ligands. For DSS this is particularly stark: only 9 exit points are obtained in 250 ns of simulation with conventional sampling, compared to 101 and 95 from the two WExplore simulations.

3.3.3 Surface binding maps

To help investigate the origins of the difference between BUT and the other two ligands we examined ligand binding to various regions of the protein along the exit pathways observed in the WExplore simulations for each ligand. Figure 3.4 compares binding probabilities, as measured using an atom-atom cutoff distance of 6 Å, for the three different ligands mapped onto the protein surface. The views on the left hand side show the front of the molecule, with the binding pocket in the middle. In accord with the results above, BUT shows a much higher density of bound ligand in the lower left hand quadrant than either DSS or DMSO. This can be seen more clearly in the side views for each ligand, shown in the right hand column. The surface density for DSS is much patchier than for the other two ligands, indicating a smaller number of more tightly bound poses.

To investigate the structural details of these poses we perform a clustering procedure and pull out representative conformers (see Methods). The same cluster radius is used for each ligand allowing for comparison across the ligand set. We define a pseudo Shannon entropy as $S = -kT \sum_i p_i \ln p_i$, where the sum is over the set of clusters, and the probability of a

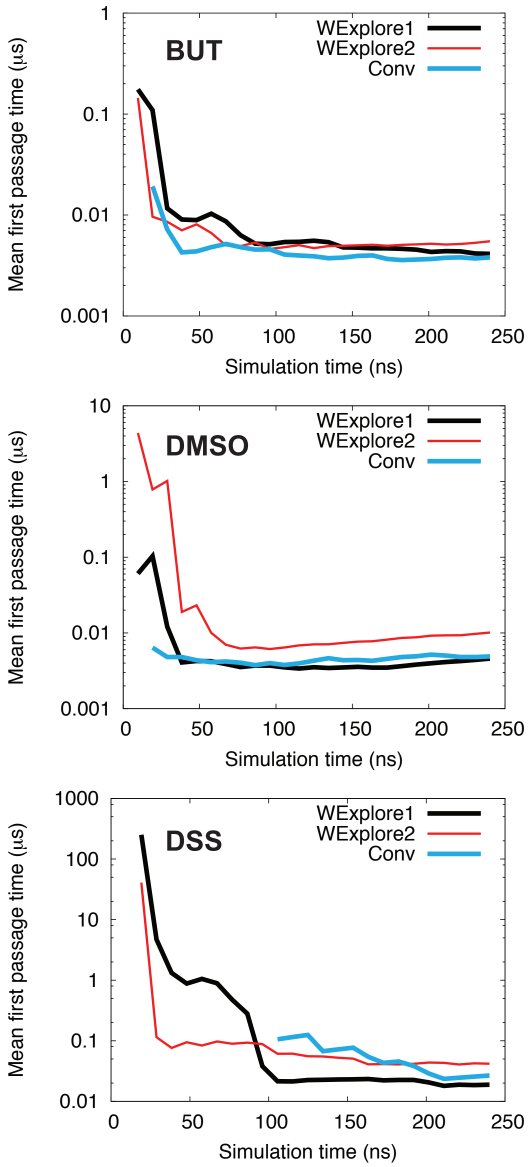


Figure 3.2: **Estimated mean first passage time (MFPT) for unbinding as a function of simulation time**. Every 10 simulation cycles the best estimate of the unbinding mean first passage time (MFPT) is computed, and is shown here for both WExplore simulations and a conventional sampling simulation for each ligand. The exit rates are computed as described in the Methods section, and the MFPT is the inverse of this rate.

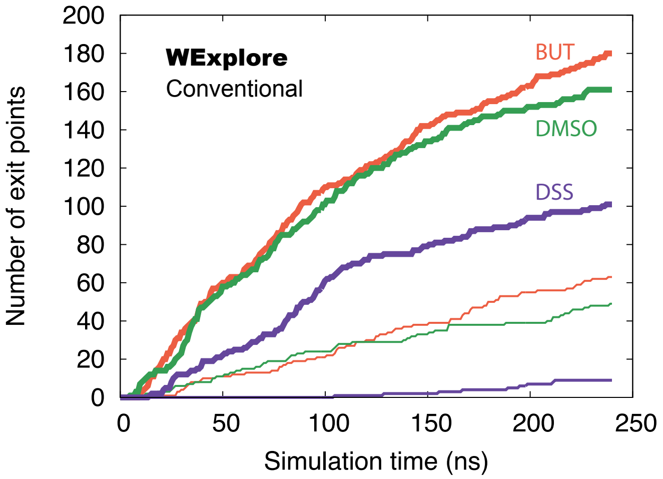


Figure 3.3: **WExplore more efficiently generates exit points**. A running total of the number of exit points observed over the course of the simulation is plotted for both WExplore (thick lines) and conventional sampling (thin lines). The WExplore curves are the averages of the two WExplore simulations performed for each ligand.

Table 3.2: Pseudo-Shannon entropy of the bound ensemble, and the number of clusters found for each ligand, using a constant cluster radius.

	N_c	S (kT)
DMSO	497	4.31
BUT	485	4.49
DSS	565	3.59

cluster (p_i) is the sum of the probabilities of each conformation in the cluster. Note that this is not equal to the true Shannon entropy, as we are only using statistics from the non-equilibrium unbinding ensemble. Table 3.2 shows the S values for the three ligands, and these are consistent with the ligand surface densities: DSS has a significantly lower conformational entropy in the bound ensemble. Interestingly, this is not reflected in the number of clusters that are found for DSS, which underscores the ability of WExplore to broadly sample low probability conformational ensembles.

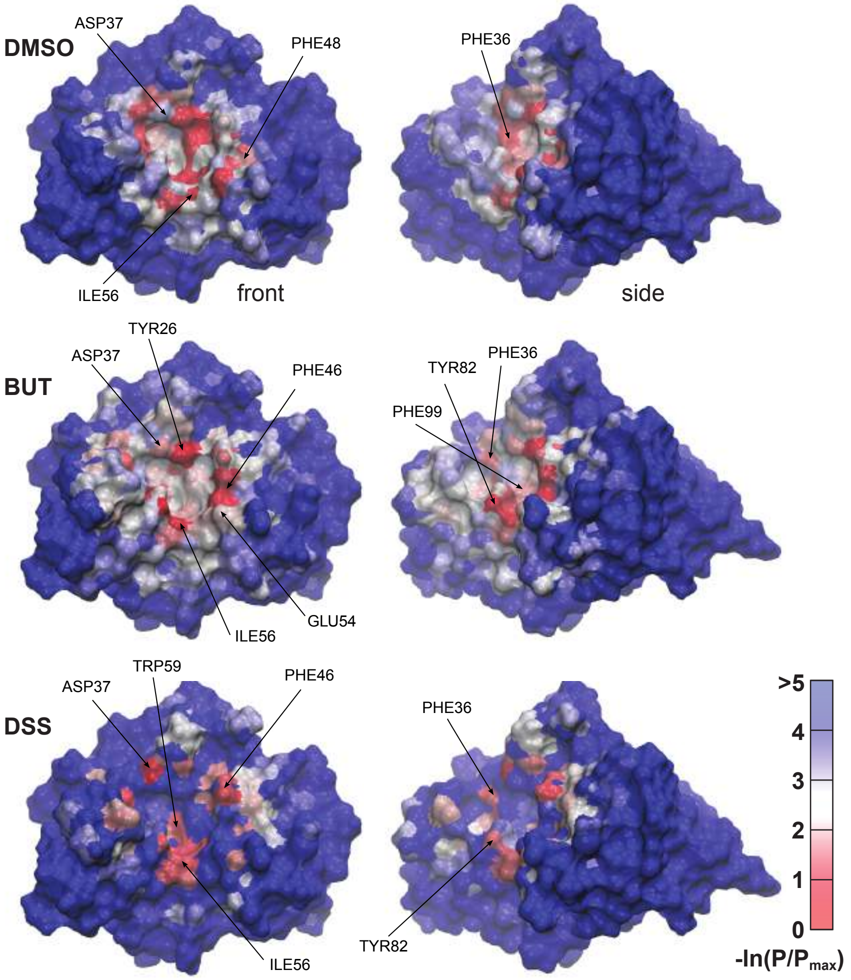


Figure 3.4: **Ligand surface densities**. Probabilities of ligand association are calculated in the unbinding ensemble on a per-atom basis. These probabilities are shown here mapped onto the structure of FKBP. Note that only one FKBP conformation is shown for easy comparison, but these association probabilities are averaged over entire WExplore simulations, involving substantial changes in FKBP conformation. The color map shows a pseudo free-energy, using the negative natural logarithm of the probability. Residues involved in ligand-protein interactions (Figure 3.6) are labeled, along with those that show high association probabilities.

3.3.4 Binding pose analysis

The most probable bound structures are also determined by choosing a representative structure from the highest probability regions for each ligand. These structures are overlaid with the crystallographic poses in Figure 3.5. After aligning the structures based on a set of FKBP residues in the binding pocket, the root mean squared distance between the highest probability poses and the crystallographic poses is determined. All three ligands show general agreement with respect to the position of the ligand. For DSS and DMSO, these poses show high agreement with the crystal structure poses, with RMSD values of 1.03 Å and 0.77 Å respectively. The highest probability BUT pose found here differs in orientation from the pose in the crystal structure with PDB ID 1D7J [153] (RMSD = 3.08 Å), with the carbonyl group pointing the opposite direction (in our case, towards the NH group of residue I56). Interestingly, this pose was also identified in a set of high probability poses of BUT identified by Huang and Caflisch [1], although it was not found to be the most probable.

Non-covalent interactions for hydrogen bonding and pi-cation interactions are then profiled with the PLIP (Protein-Ligand Interaction Profiler) program [159]. For hydrogen bonds to be detected the acceptor and donor must be less than 4.1 Å apart and at an angle between 180° and 100° [160]. Pi-cation interactions are detected when the charged atom and aromatic ring center are less than 6.0 Å apart [161]. We identify interactions in the representative conformers of high-weight cluster centers ($w > 0.005$), and show all such unique interactions for each ligand in Figure 3.6.

These protein-ligand interactions largely explain the surface densities and pseudo-Shannon entropies for each ligand. BUT forms hydrogen bonding interactions in 128 representative structures (compared to 88 and 69 for DSS and DMSO, respectively), contributing to its higher entropy of clustering and its broad surface density. DSS has more restricted surface densities (Figure 3.4), and lower entropy (Table 3.2) than both BUT and DMSO, which may be explained by a high-probability pi-cation interaction of a partially positive sulfoxide (0.310 D) and the aromatic rings of TRP59, specifically the -0.61 D indole N (N1) (panel

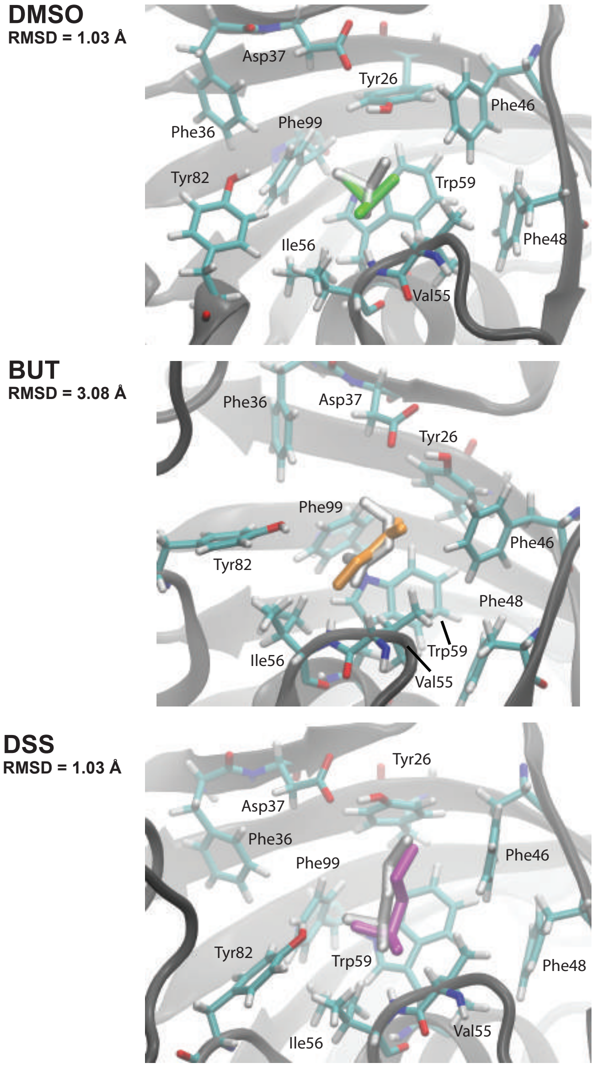


Figure 3.5: **Comparison of highest probability poses with crystallographic poses**. Representative structures from the highest probability clusters are shown for each ligand. Residues that compose the binding pocket are shown in stick representation, and are labeled. The ligand pose from simulation is shown in color: green for DMSO, orange for BUT and purple for DSS. The corresponding crystallographic ligand pose, after alignment to the set of local protein residues shown here, is shown in white. The position of the ligand agrees very well between the two sets of structures, although the orientation of BUT is flipped.

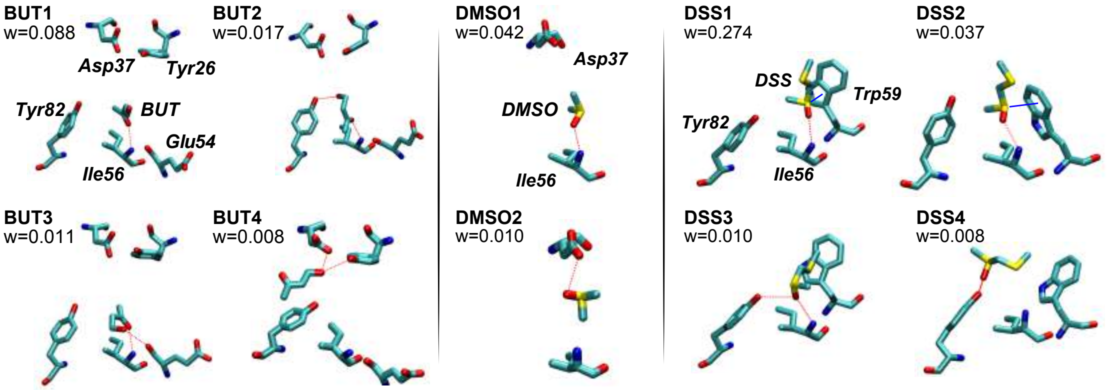


Figure 3.6: **High probability ligand-protein interactions**. Representative structures for cluster centers with unique ligand-protein interactions and weight greater than 0.005 are shown for each ligand numbered in descending order of cluster weight. Red dashed lines indicate hydrogen bonds and solid blue lines indicate pi-cation interactions.

DSS1 of Figure 3.6) and the −0.115 D carbons of the six member indole ring (C4-7) (panel DSS2 of Figure 3.6). The interaction with N1 has a much higher probability ($w = 0.274$) than the most probable interactions for BUT ($w = 0.088$) and DMSO ($w = 0.042$).

It is interesting that even though DMSO is a structural subset of DSS (also with a 0.31 D sulfoxide), we did not observe probable poses involving pi-cation interactions nor were any poses stabilized to the extent of DSS1. This is likely due to the small size of DMSO, which lacks the steric confinement that can stabilize electrostatic interactions in larger ligands like DSS. This suggests that fragment-based screening approaches using very small ligands may miss important electrostatic interactions such as pi-cation interactions, salt-bridges, and halogen bonds that can help drive receptor specificity.

3.4 Conclusion

Ligand release in general is an extremely demanding computational problem, and even for these low affinity ligands, the enhanced sampling method WExplore enabled our collection of a statistically relevant set of exit pathways. This work is the first demonstration of WExplore on a ligand binding system. The distance metric used to define the sampling regions captures both internal ligand motions as well as global ligand translations and rotations with respect to the binding pocket. We expect that this distance metric will also work well for flexible ligands with binding pathways that feature motion along internal degrees of freedom. One limitation is that this distance metric does not enhance sampling of protein motions that are distant from the binding site. Thus if ligand release is triggered by allosteric motions far from the binding site (e.g. as revealed in the work of Plattner and Noé [44]), a different distance metric should be used that would also incorporate those motions in order to efficiently sample ligand release. We note however, that the method described does nothing to inhibit protein motions, only that they are not enhanced, and if this is the slowest degree of freedom will likely be the bottleneck to observing unbinding events.

WExplore is well suited to ligand binding applications as it does not require a specific single order parameter, and thus does not bias our collection of exit paths in any particular direction. Numerous other enhanced sampling methods such as temperature accelerated molecular dynamics [54, 53], orthogonal space random walk [162], self-guided Langevin dynamics [163], random acceleration molecular dynamics [164] and random expulsion molecular dynamics [165] are also suitable for this purpose, although WExplore is unique in that its trajectories are run using the unbiased Hamiltonian, allowing for direct incorporation into Markov state modeling frameworks [166] and network visualization tools [167, 168]. Looking forward, a robust comparison of the performance of enhanced sampling methods for the generation of ligand binding and unbinding trajectories is needed, from the standpoint of efficiency as well as accuracy.

The coming years should see a growth in simulations of full binding and unbinding trajectories, due to growing computational power as well as improvements in ligand force field parametrization. General, quantitative metrics that measure properties of ligand binding pathways, together with visualization tools will go a long way to help define principles and models for the basic physics of ligand binding. A deep physical understanding of the binding process will improve our understanding of binding kinetics, helping both directly to design drugs that have the desired kinetic properties – such as long residence times [7, 16, 169] – and indirectly to inform scoring functions used in ligand- and receptor-based virtual screening.

CHAPTER 4

TRYPSIN: MULTIPLE UNBINDING PATHWAYS AND LIGAND-INDUCED DESTABILIZATION REVEALED BY WEXPLORE AND CONFORMATION SPACE NETWORKS

4.1 Introduction

The pathways traveled by ligands as they bind to their molecular receptors are important to drug design. Although the binding thermodynamics is purely determined by the endpoints of these pathways, analysis of the entire paths can reveal binding transition states that govern the kinetics of the binding process. Under-appreciated until recently, long residence times have been shown in a handful of systems to be more predictive of in vivo efficacy than the thermodynamics alone [169, 14]. Conversely, fast binding and release could also be preferable in some applications, including enzyme inhibition [170], and for systems where fast clearance of the drug is essential. Robust methods that can predict structure-kinetics relationships would thus be of tremendous value to drug design efforts. Unfortunately, structural details of ligand-binding transition states are difficult to capture experimentally, and ligand binding and release typically occur on timescales that are inaccessible to conventional molecular simulation.

Recently, a handful of cutting-edge applications of molecular dynamics, using either specialized hardware [35, 48], large parallel sampling efforts synthesized with Markov state models [43, 171, 44], or customized enhanced sampling algorithms [41, 52, 46, 42], have been applied to study full ligand binding or unbinding pathways. These have revealed an intricate interplay between the conformations of the ligand and receptor, and are beginning to reveal how biological molecules are controlled by exogenous factors, which is important both for our understanding of biology, and for our ability to design drugs that elicit a desired biomolecular response. Despite some progress, the principles that govern the general relationship between ligand binding and protein stability or protein activity remain elusive. General biophysical properties of protein-ligand interactions are needed to elucidate and predict phenomena such as allosteric signaling networks [172], and ligand-induced stability changes [173]. This necessitates a general knowledge of how ligand binding is coupled with conformational change in the binding site.

The binding of the ligand benzamidine to trypsin has in recent years served as the system of choice to demonstrate emerging enhanced sampling approaches to study ligand binding [43, 41, 45, 11, 46, 44, 42]. Long simulations of ligand binding synthesized with Markov state models obtained binding rates that showed good agreement with experiment [43, 45, 44], but the unbinding rates were consistently over-predicted, owing to the steep free energy barrier of ligand unbinding. Particularly, Plattner and Noé used hundreds of microseconds of simulation to show a dynamic picture of trypsin with two main binding channels and multiple long-lived trypsin conformations [44]. Approaches using metadynamics with path-based order parameters have also obtained unbinding rates [46], but these were significantly under-predicted, although again the binding rates showed excellent agreement. Teo et al [42] used the Adaptive Multilevel Splitting method to obtain excellent agreement with the experimental rate with modest computational cost, but did not observe some of the long time-scale conformational transitions seen by previous investigations.

Here we use our own technique, WExplore [75], to investigate a broad set of ligand release pathways in the trypsin-benzamidine system.

This and related methods have been used to study protein unfolding, hydration changes near a fluorophore [174], long time-scale conformational transitions in a RNA helix-helix junction [92] and to generate the ensemble of unbinding pathways of small ligands from the protein FKBP [39]. Like MSM approaches, it uses trajectories that are run with the unbiased Hamiltonian and are suitable for a network-based conformation analysis [167, 1, 168], but it is based on a weighted ensemble of trajectories, and obtains unbinding rates by a different mechanism that does not rely on a Markovian assumption of transitions between regions. A set of trajectories are run in parallel, each with a statistical weight, and these are actively managed every 20 ps using cloning and merging steps that maximize the heterogeneity of the trajectory set. As in the original weighted ensemble algorithm [69], during cloning the weights are split, and during merging the weights are added. Observables are then computed using weighted averages.

One such observable is the flux of trajectories that cross into the unbound state (defined here as all states where the minimum protein-ligand distance is greater than 10 Å). In the non-equilibrium ensemble where trajectories are initiated in the binding site and are terminated in the unbound state, the flux of trajectories into the unbound state per unit time is equal to the unbinding rate.

4.2 Methods

4.2.1 Molecular Dynamics Simulations

Dynamics are run in CHARMM [129] on GPUs using the program OpenMM version 6.3. The system is constructed using the coordinates from PDBID 3PTB, preserving the crystallographic calcium ion and the 62 crystallographic water molecules. The system is then solvated with a 12 Å cutoff surrounding the protein and the ligand, resulting in 12592 waters. Nine chlorine ions are added to neutralize the system, resulting in 41006 atoms total. Cubic periodic boundary conditions with a box size of 74.3 Å.

For dynamics we use a 2 fs timestep. Dynamics are performed in the constant pressure, constant temperature ensemble, coupled to a Langevin heat-bath with temperature 300 K and friction coefficient of 1 ps^{-1}, and a Monte Carlo barostat with a reference temperature of 1 atm, and volume moves attempted every 50 timesteps. The ligand is parameterized using the CHARMM Generalized Force Field (CGenFF) [154]. We compute non-bonded interactions using Particle Mesh Ewald, with a switching function that scales the non-bonded interactions to zero at 10 Å, starting at 8.5.

The solvent is first minimized using 500 steps of steepest decent followed by 500 steps of the Adopted Basis Newton-Raphson method, and the entire system is then minimized in the same way. After minimization, we gradually heat the system from 50 K to 300 K in ten steps of 10 ps each, followed by equilibration at 300 K for 500 ps. the resulting structure is then used as the initial conformation for all 48 walkers in the WExplore sampling method.

4.2.2 WExplore Sampling

The WExplore methodology has been described in detail in previous work [92, 75], as well as its application to ligand unbinding simulations [39] (see also Sections 1.2.4 and 1.2.6.1 of this thesis). In this study we used the following parameters for WE and WExplore. Simulations were run with a static number of walkers numbering 48, where each walker is run for 2 ps of MD sampling time between resampling cycles. The distance metric employed calculates the root mean square deviation (RMSD) between the ligands of two sample after aligning the binding sites to the initial structure. For the WExplore hierarchy four levels were used with the following region creation thresholds respectively d = 10, 5, 3 and 1.7 Å. We use a minimum walker weight of 10^{-12} and a maximum weight of 0.1, these are enforced by preventing cloning and merging steps that would violate these rules.

4.2.3 Clustering

To visualize the results in conformation space networks, we first cluster the joint data set of all 5 WExplore simulations. This is done in MSMBuilder [157], using a set of ligand-protein distances. The set of distances is constructed using the 50 closest heavy atoms in the protein to the ligand in its crystallographic conformation (set A), and the nine heavy atoms in the ligand (set B). We use every possible connection between sets A and B for clustering: a set of 450 distances. These are clustered using the KCenters algorithm and the Canberra distance metric, which highlights differences between quantities which are small. This is ideal for our purposes as it helps avoid over-clustering poses in the unbound state, which have large distances between the ligand and receptor.

4.2.4 Hydrogen Bond Profiling

We first introduce some terminology which is used througout. The term “Interaction Classes” describes a specific *potential* interaction between two specific features in a molecular system. This is accompanied by a series of constraints, typically geometric. The term “Interaction Instance” is a specific molecular structure where the constraints for an Interaction Class have been met. I.e. it is an actual example of an Interaction Class rather than only the potential.

Instances of hydrogen bonds were detected with the following constraint parameters between interaction class features:

1. acceptor donor distance < 4.1

2. $100^\circ < angle < 180^\circ$.

Descriptive statistics of community interactions is given in Table 4.1. The summary of the number of interactions per node, by community in Figure 4.2.4 shows that B and B* have a high average number of interactions per node, but also the largest ranges. P3 stands out from P1, P2, and P3* in having a fairly high average number of interactions per node indicating that perhaps a higher level of coordination between multiple stabilizing interactions is necessary for unbinding via P3.

The software to identify hydrogen bonds used in this study is available at `https://github.com/ADicksonLab/mastic`.

Table 4.1: **Descriptive statistics of interactions by community** In the table the left hand column "Community Name" is the abbreviation for the network community that is described on that row. The "# of Interaction Classes" is the total number of interaction classes for which there was at least one instance. The "# of Interaction Instances" is number of total instances for any interaction class in the community. The "Average Frequency of Pairs" is number of nodes any specific Interaction Class was found in. The "% Total Frequency" is the percentage of total number of observed Interaction Instances for that community. The "Average Frequency of Nodes" is the average number of Interaction Instances observed in each node, and the "Std. Dev. Per Node" is the standard deviation of that average.

Community Name	# of Interaction Classes	# of Interaction Instances	Average Frequency of Pairs	%-Total Frequency	Average Frequency of Nodes	Std. Dev. Per Node
B	76	2793	36.8	32.4	4.44	1.96
B*	59	1092	18.5	12.7	4.94	1.99
P1	68	842	12.4	9.77	2.23	1.63
P2	82	710	8.66	8.24	1.31	1.48
P3	101	1314	13.0	15.2	2.89	1.60
P3*	43	181	4.21	2.10	1.72	1.40
U	214	1646	7.69	19.1	0.998	1.30

4.2.5 Specific Structure Analysis

We obtained community representatives by first selecting all nodes from each community that contain an instance of the interaction pair of highest frequency in that community, and then from those choosing the one with highest weight. These structures are shown in Figure 4.2.5, where the highest frequency hydrogen bond is indicated and the principle ASP189 is shown as a point of reference. The structure for B is the highest weighted node in the network and is similar to the crystal structure. As expected the P1 (highest weighted unbinding pathway) structure features the ligand simply backing out of the pocket and the highest frequency hydrogen bond occurs with the adjacent SER190 side-chain. The U community is not well represented by a single high frequency interaction, but the representative structure seems to be, unsurprisingly, related in position to the P1 unbinding pathway. Benzamidine hydrogen bonding in B* also involves ASP189, but there is a conformational change of the blue loop that opens the P2 exit pathway. The B* structure appears to be a precursor to P2 as ASP189 is flipped out of the pocket allowing hydrogen bond formation with a backbone oxygen on TRP215 (and likely pi-pi stacking against the indole ring) guiding the ligand away from the binding pocket. P3 and P3* are both related in their localization in the network pathways as well as in the conformational changes in the blue and orange loops. In both there is a closing of the binding site by the blue loop and the opening of gaps in the orange loop, likely as exit sites. It also appears that P3* is a precursor to P3 as the ligand is much closer to the original B position and orientation in P3*. However, among unbinding pathways P3 appears to be much more multi-modal and this relationship is likely to be more complex. The identification of B* and P3* indicate that the use of graph theoretic methods will likely continue to be useful in identifying and refining unique states along complex unbinding pathways and ultimately identifying the salient intermolecular interactions useful for developing drug targets.

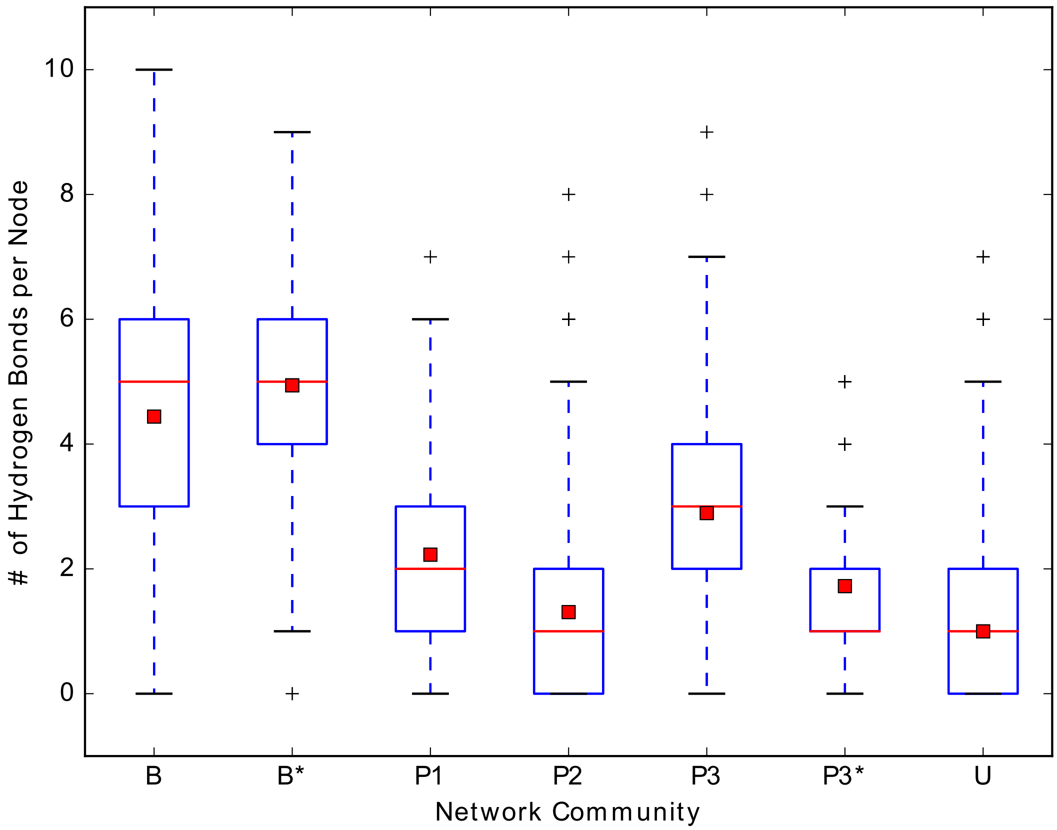


Figure 4.1: **Box-plots of the number of hydrogen bonds per node by community**. The vertical axis gives the number of hydrogen bonds per node. The red square indicates the average for the community and the red line is the 50th percentile. The box bounds the 25th and 75th percentiles and the dashed line indicates the range with outliers are drawn as crosses.

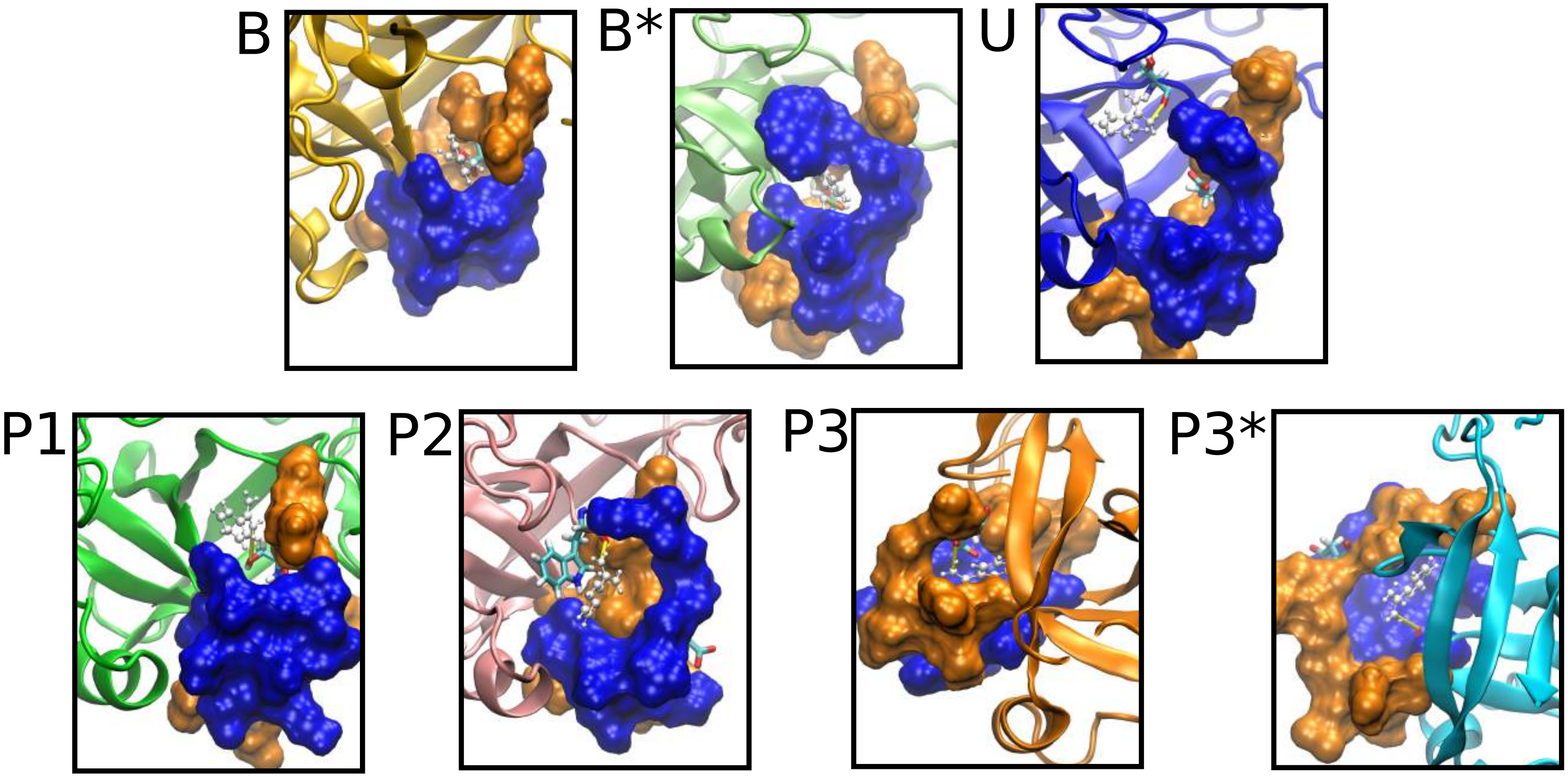


Figure 4.2: **Binding poses of community representative structures**. The panels correspond to the representative structures that were chosen for each community based on the highest weighted with it's highest frequency hydrogen bond (shown as a yellow cylinder). The ligand is shown in white CPK representation and it's acceptor partner is shown in licorice representation. The blue and orange loops in surface representation correspond to residue indices 209-218 and 179-190, and are the same as in Figs. 2 and S5. The native binding site residue ASP189 is shown in licorice representation as a point of reference. Structures for B, B*, U, P1, and P2 are shown from the binding site side of the protein (blue loop), while P3 and P3* show the alternative exit pathway through the orange loop.

4.3 Results

Figure 4.3 shows the predicted mean first passage time (MFPT) (MFPT = $1/k_{\text{off}}$) as a function of simulation time for five independent WExplore runs. Each run uses 48 trajectories total that are cloned and merged repeatedly throughout the simulation, and the total sampling time for each run averages 820 ns. A total of 4.1 μs of simulation time is used to generate the average curve (thick orange line) that obtains a final prediction of 180 μs, using the last 10% of the data. Significantly, the individual results differ over eight orders of magnitude, owing to large differences in the weight of the trajectories that break out of the binding pocket. Large, downward jumps in residence time occur (e.g. Run 3, Run 5) when a new exit point is recorded that has a significantly higher weight than the others recorded so far. As such, we expect that extensions of runs 2 and 4 forward would eventually converge toward the mean, although we have found that multiple shorter runs are more efficient than single long ones, as the weight distributions within a run are much more highly correlated than those between the runs. Despite this variability, the averaged trajectory flux gives a MFPT is within an order of magnitude of the experimental value of 1700 μs (Table 4.2).

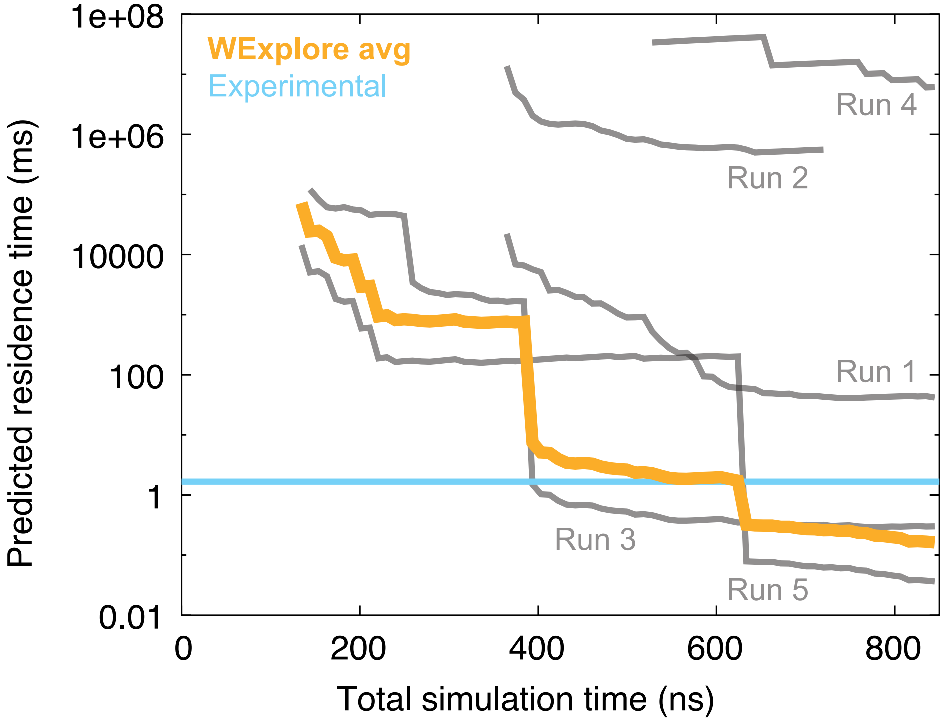


Figure 4.3: **Mean-first passage time of ligand unbinding**. Predicted residence times are shown for all five WExplore runs (grey). The residence time computed using the average probability flux across all WExplore simulations is shown as a thick orange line, and shows reasonable agreement with the experimentally determined residence time [3], shown as a horizontal blue line.

Table 4.2: Mean first passage times of unbinding as measured here (*) and in previous simulations. Also tabulated is the method used and the total simulation time.

	Experiment [3]	*	Plattner and Noé [44]	Tiwary et al [46]	Doerr et al [45]	Buch et al [43]	Teo et al [42]
MFPT (μs)	1700	180	76	110000	100	11	3800
Methodology	–	WExplore	MSM	Metadynamics	Adaptive MSM	MSM	AMS [1]
Total sim. time (μs)	–	4.1	150	5.0	10	50	2.3

As these simulations are conducted using the unbiased Hamiltonian, we can use conformation space networks to synthesize our findings [167, 1, 168]. Figure 4.4A shows the complete network of states visited by all five simulations, created by clustering using a set of 50 ligand-protein distances (see Section 4.2.3). Node sizes show the state probabilities; the biggest nodes in the top right are the bound states closest to the crystal structure used to initialize the simulations (PDBID 3PTB). Nodes are colored here by solvent accessible surface area (SASA), which reveals a large number of states that are kinetically far from the crystal state, but are still completely buried inside the protein. We find three transition paths that connect the bound and unbound basins (Figure 4.4B). Path 1 is the direct exit pathway that has been found by all previous investigations. In Path 2, the loop shown in blue (residues 209-218) undergoes a conformational change and creates an alternative pathway for benzamidine release. This path was previously observed by Plattner and Noé [44], and significant loop motions in this region were also observed using metadynamics [46]. Path 3 involves a similar conformational change in another loop shown in orange (residues 179-190) creating a large set of bound, buried states that have not been previously observed.

We break up our network into communities using a fast stochastic modularity-based community detection algorithm [175] (Figure 4.5A). We obtain seven communities: two of each representing the bound (B,B*), and path 3 (P3,P3*) states, and one of each representing unbound (U), path 1 (P1) and path 2 (P2). To study these communities we first profile the entire set of ligand-protein hydrogen bonds (H-bond) in the network. For each H-bond that we observe in our simulations, Figure 4.5B shows the frequency with which it is observed in each of the seven communities. 276 unique acceptor-donor pairs are found with 8621 H-bonding instances total (see 4.2.4). B and B* distributions are dominated by a few high frequency pairs, while U has many low to moderate frequency pairs. The remaining unbinding pathway communities (P1, P2, P3, and P3*) have somewhat heterogeneous distributions but feature some high frequency interaction pairs that are mostly non-overlapping between pathways. This suggests that each pathway may be characterized uniquely by only a few

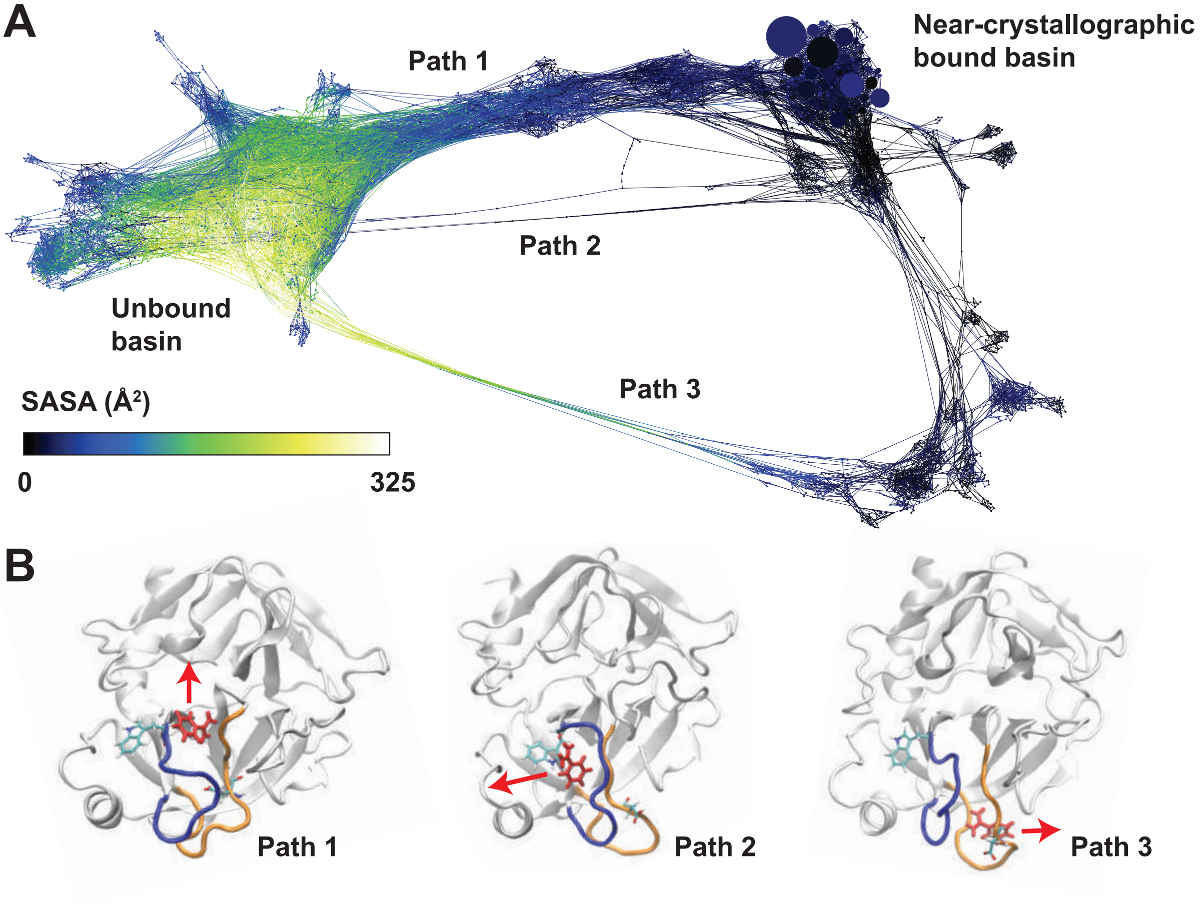


Figure 4.4: **Trypsin-benzamidine unbinding network shows three exit pathways.** (A) The conformation space network of the trypsin-benzamidine system is shown. The size of the nodes corresponds to the weight of the states, and the node color shows the SASA of a representative structure from that region. The bound and unbound basins are connected by three discrete transition paths, which are labeled. (B) Representative structures are shown that characterize the mechanism of the three transition paths. Benzamidine is shown in red, and the general direction of exit is shown with a red arrow for each pathway. Residues TRP208 and ASP186 are shown in licorice representation, and the loop regions 179-190 and 209-218 are shown in orange and blue, respectively.

specific interactions. The highest-weighted structures for the highest frequency pairs in each community are shown in Section 4.2.5.

Each of the three pathways is not observed by every WExplore simulation (Figure 4.6). Path 1 is observed in runs 2, 3 and 5, Path 2 is observed only in run 1 and Path 3 is observed only in run 4. Figure 4.7 shows the free energy of each state, which shows Path

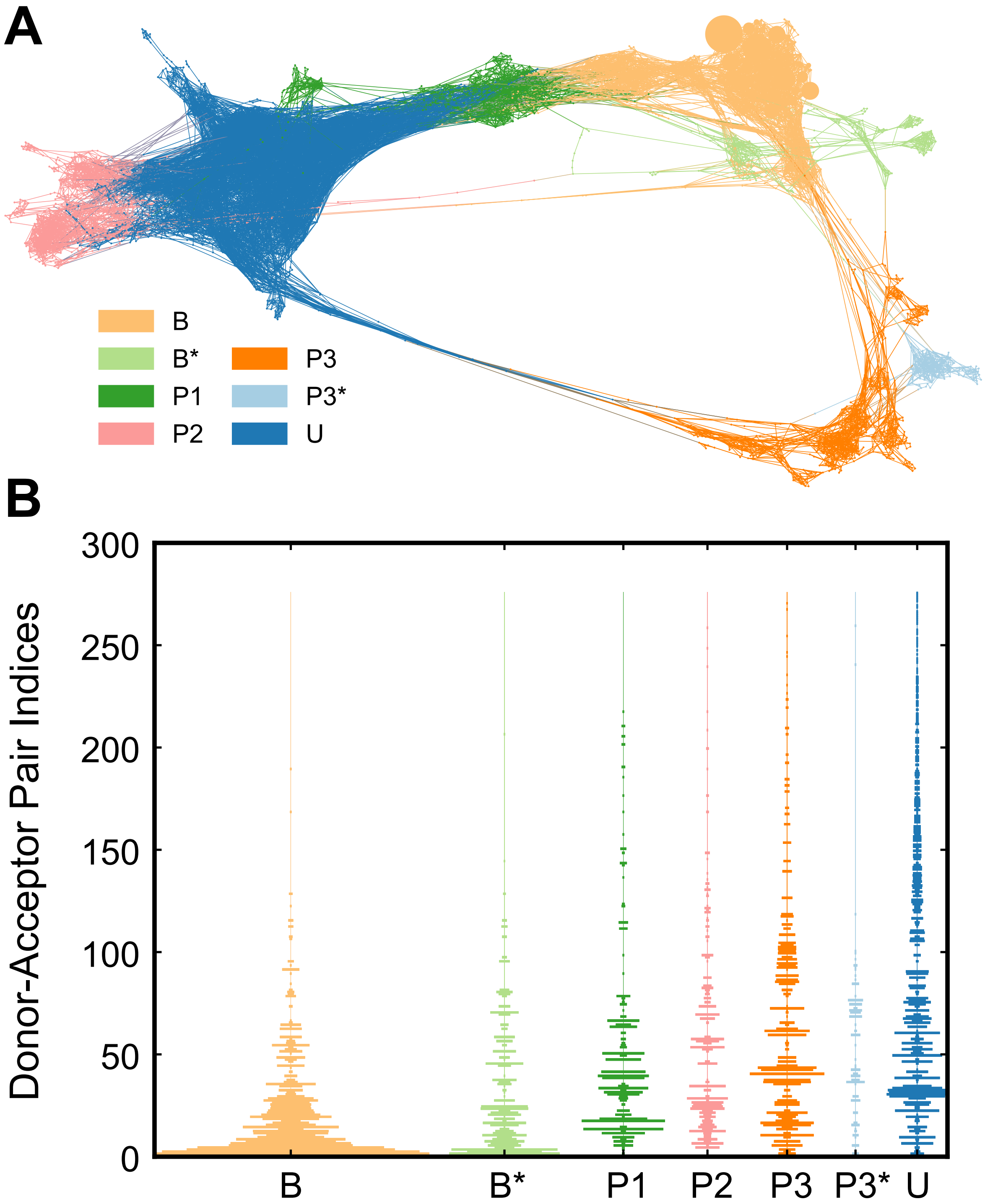


Figure 4.5: **Community Detection and Hydrogen Bonding Frequencies**. (A) Network plot showing communities of the network. The labels B and B* correspond to the two bound state communities and U corresponds to the unbound states. P3* is classified as a distinct component of the P3 pathway. (B) Violin bar plots of hydrogen bond frequencies. The vertical axis shows the donor-acceptor pairs sorted by their frequency in the whole network. Each violin shows the frequencies observed within each community.

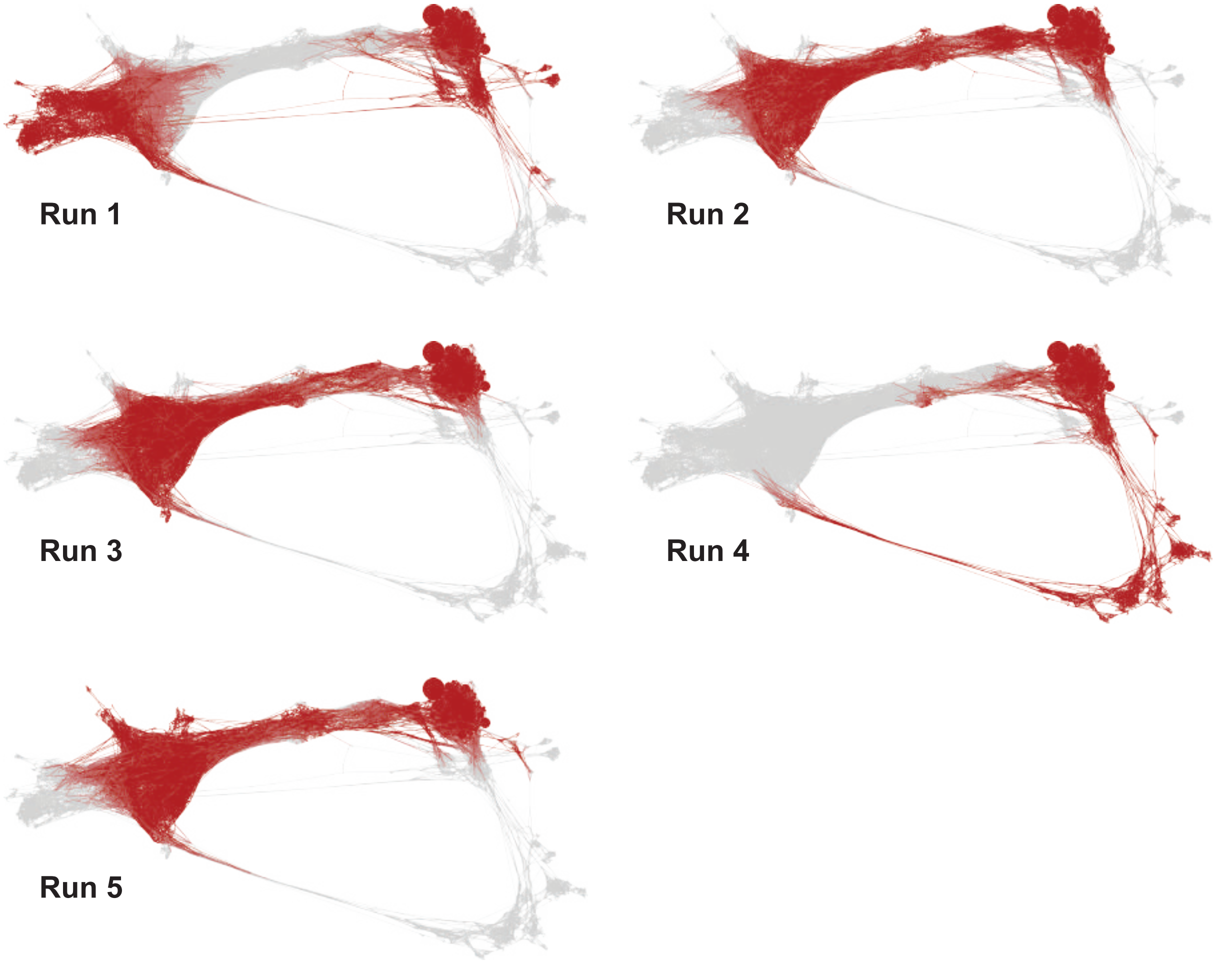


Figure 4.6: **Conformation space networks colored by the contribution from the five WExplore runs**. In each figure, a node is colored in red if it is sampled in that run, and in light grey if it is not.

1 to be by far the most probable, Path 2 to be the next-most probable, and Path 3 to be the least probable, consistent with Figure 4.3. However, this result underscores the ability of WExplore to discover alternative bound conformations, even those that are separated by large free energy barriers, requiring significant rearrangement of local protein structure.

The solid agreement with experimental rates, the broad sampling of pathways and poses, and the relative efficiency of our technique bode well for future applications of WExplore. Drug-like ligands can have residence times approaching minutes or hours, which will be

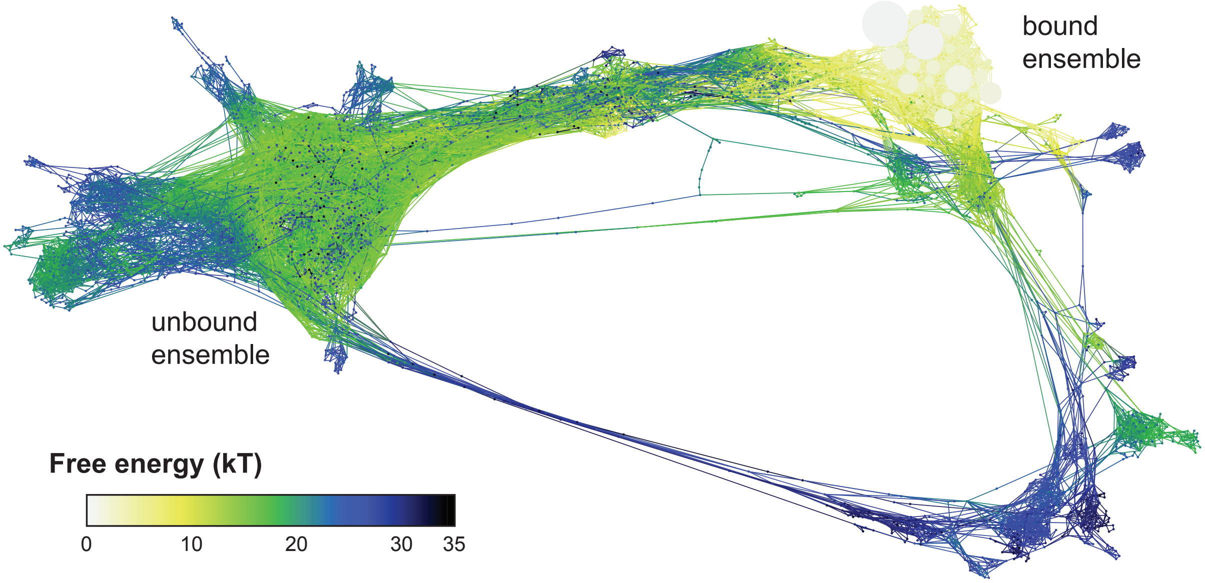


Figure 4.7: **Conformation space network colored by the free energy**. The free energy is shown in units of kT.

prohibitive to straight-forward molecular dynamics for the foreseeable future, but is well within the residence time that we predict for benzamidine dissociating via Path 3, which involves substantial rearrangements of the protein that occur on extremely long timescales. Further testing is needed on ligand dissociation events that occur on longer time scales, which could reveal important information about the optimization of kinetic properties for drugs under development. (Un)binding pathways can also reveal important molecular motions in the receptor that can be used to design new ligands that stabilize alternative receptor conformations. As an example, many states are identified here where the ligand is still deeply buried (SASA ≈ 0) that are kinetically far from the crystallographic starting structure. It is easy to imagine this approach being used to identify such states, which can serve as templates for the design of new ligands that bind via an induced-fit mechanism.

CHAPTER 5

UNBIASED MOLECULAR DYNAMICS OF 11 MIN TIMESCALE DRUG UNBINDING REVEALS TRANSITION STATE STABILIZING INTERACTIONS

5.1 Introduction

In our previous work we applied our simulation and analysis techniques to a series of more challenging model systems (Chapters 3 and 4 respectively). In this study we present unbiased pathways of a drug development intermediate (1-trifluoromethoxyphenyl-3-(1-propionylpiperidin-4-yl)-urea, or TPPU) unbinding from its receptor, sEH, using WExplore. TPPU forms several stable interactions with residues that are buried within the sEH protein. The unbinding of TPPU has an experimentally determined residence time of 11 min, which is significantly longer than the previously studied systems.

With WExplore we generate a series of unbinding trajectories using unbiased dynamics, and construct a portrait of the ligand-protein free-energy landscape through a conformation space network (CSN). We extensively profile ligand-protein hydrogen bonds throughout the network, and use these to develop an understanding of distinct steps in the unbinding mechanism. Furthermore, we identified the likely ensemble of states surrounding the unbinding transition state (TS) and begin to investigate potential TS stabilizing interactions.

5.2 Methods

5.2.1 Molecular Dynamics

Initial atomic coordinates were taken from the PDB entry 4OD0 [2], which has two domains with few inter-domain contacts and are connected by a flexible linker. We removed the domain which does not bind to the ligand of interest, including the associated crystallographic waters, magnesium ion, and phosphate ion, retaining residues Ser231 to Arg546 (PDB serial numbers) and 10 crystallographic waters. To solvate the system, a box of TIP3P water molecules were placed around the protein and ligand with a 12 Å cutoff. The completed system included 316 amino acids (5052 atoms) for the protein, 45 ligand atoms, 16 831 water molecules, and 7 neutralizing sodium ions for a total of 21 935 atoms.

The protein is parametrized by the CHARMM36 force-field [176] and the ligand force-field was derived using the homology based CHARMM Generalized Force Field (CGENFF) algorithm [177]. Before dynamics the solvent molecules are minimized using 500 steps of steepest descent followed by 500 steps of the adopted Newton-Raphson method. The whole system is then minimized in the same way.

Molecular dynamics are run using the CHARMM program [129] with an OpenMM (v6.3) [111] interface to allow use of GPUs. Timesteps of 2 fs are calculated in the constant pressure and temperature ensemble by coupling to a Langevin heat-bath, temperature of 300 K and a friction coefficient of 1 ps^{-1}, and a Monte Carlo barostat with reference pressure at 1 atm where volume moves attempted every 50 timesteps. Non-bonded interactions are calculated using the particle mesh Ewald method with a switching function that scales interactions to zero from 8.5 Å to 10 Å. We start the heat-bath at 50 K and run dynamics for 10 ps before increasing the temperature in ten even steps to 300 K. The resulting structure is then equilibrated at 300 K for 500 ps before being used in WExplore simulations.

5.2.2 WExplore Simulations in the Non-equilibrium Unbinding Ensemble

Weighted ensemble is described in Section 1.2.4 and the WExplore algorithm is detailed 1.2.6.1, here we only present the parameters used.

In our simulations we use four levels of a hierarchy with cutoffs $d = 10$ Å, 5 Å, 3.5 Å and 2.5 Å. Cloning and merging steps occur every $\tau = 20$ ps between the set of 48 walkers. As in previous work [47], we enforce a maximum weight of 0.1 and a minimum weight of 10^{-12} by disallowing cloning and merging steps that would violate these conditions. Each separate run of WExplore was run with 48 replicas and was run for 1 µs of simulation time. The parameters used here, including d, τ, the number of walkers, minimum and maximum weights, and the definition of the unbound state, are identical to those used previously for the trypsin-benzamidine system [47], which emphasizes the generality of the WExplore approach.

Simulations are run such that trajectories start in the bound state (source; A) and end in the unbound state (sink; B). The source state is the initial bound pose and the sink is the set of conformations in which the shortest protein-ligand distance is greater than 10 Å. Walkers that have crossed the transition barrier normally decay very quickly to this absorbing boundary at the free-energy minima of the unbound state and thus contribute little probability to this portion of the state space, and consequently have very high free-energies. Thus the free energies calculated from the weights in our simulations are not the more familiar equilibrium free-energies. The equilibrium free-energy can be recovered if the non-equilibrium binding ensemble is also calculated and combined with the unbinding ensemble, which has not been done here. For our purposes the unbinding ensemble is sufficient for determining the structure of the transition state, and subsequently calculating the unbinding rate.

5.2.3 Clustering and Network Visualization

To reduce the dimensionality of the sampled trajectories and to summarize the contained dynamical information we create a conformation space network (CSN). We first cluster our data according to a feature vector consisting all possible distances between TPPU atoms 2, 5, 10, 13, 15, 22, and 24 (see Fig. 5.1 for numbering) and all non-hydrogen protein atoms within 8.0 Å of any of those atoms in the bound crystal structure. A total of 2478 atom pairs were included in the MSMBuilder [157] featurizer (see Supplemental Information (SI) for details). The KCenters clustering algorithm in MSMBuilder [157] was used to create 2000 clusters with a Canberra distance metric to avoid a proliferation of unbound clusters. Edges in the CSN were defined between states that are connected in the sampling trajectories, and given a strength equal to the number of transitions observed between the nodes in either direction.

Network visualization was performed in Gephi version 0.9.1 [6] using the ForceAtlas2 layout algorithm [5], followed by a short minimization to prevent the overlap of nodes. The node sizes are proportional to the weight of each cluster, with a minimum and maximum node size chosen for clear visualization. In CSN plots the nodes and edges can be colored in order to highlight different properties calculated for the clusters. The network layout (node positions and edge lengths) is not quantitatively reproducible and layouts from separate minimizations may differ slightly due to the stochasticity of the ForceAtlas2 algorithm. However, the network layout is a useful tool for data exploration, allowing us to visualize the entire free energy landscape explored by our simulations in a single plot, and giving insight into the heterogeneity of transition paths connecting the bound and unbound states.

5.2.4 Interaction Profiling

We profile the frequency of hydrogen bonds across the set of clusters using a single structure from each cluster that was randomly chosen. Possible donor and acceptor atoms in both the protein and ligand are detected by RDKit [178]. A hydrogen bond is counted between a given donor-acceptor pair if the distance between the atoms is less than 4.1 Å and the donor-hydrogen-acceptor angle is within 100° to 180° [160]. To achieve this we use the software Mastic [179], which we have developed, and is parallelized with the SCOOP library [180], which uses the ZeroMQ message passing protocol. See SI for further details and usage.

5.2.5 Committor Probabilities and Highest Flux Pathways

Using Transition Path Theory (TPT) we estimate the highest flux pathways through the CSN and the clusters near the TS, using the TPT module in MSMBuilder [157]. For this purpose we use a trimmed network containing 1987 nodes, excluding nodes that are not connected to main connected component by both in-flow and out-flow. We first select a set of source and sink nodes, which are shown below in Fig. 5.15. The set of source nodes have cluster centers that are closer than 0.7 from the center of the cluster containing the initial MD structure according to a Canberra distance metric applied to the clustering feature vectors previously described in Section 5.2.3. The sink nodes are defined as the set of clusters whose centers' minimum distance between the ligand and protein is greater than 4 Å. Using these we construct a MSM, and calculate the unbinding committor probability ($p_{B \to U}$) using the `committors` function, then calculate the net forward flux (f^{+}) using the `net_fluxes` function in MSMBuilder [181, 182, 157]. From the unbinding committor probability we designated the set of clusters with $0.4 \leq p_{B \to U} \leq 0.6$ as the transition state ensemble (TSE). We also identified the highest flux reactive pathway in the network using Djikstra's algorithm (the `top_path` function in MSMBuilder [157]).

To find the contributions of different pathways to the reactive flux, we identify the edge with the lowest flux (the "bottleneck") along the highest flux reactive pathway. This is then removed, and the highest flux reactive pathway is again computed, identifying the next bottleneck. The reactive flux contribution of a bottleneck is equal to the flux through the bottleneck in the modified network. This procedure is continued until all but $1 - 10^{-10}$ of the reactive flux is accounted for. This procedure is implemented in the `paths` function in MSMBuilder, except for the identification of the actual edge that is the bottleneck.

5.2.6 Soluble Epoxide Hydrolase (sEH)

The protein soluble epoxide hydrolase (sEH) is found in most mammalian tissues and catalyzes the conversion of epoxyeicosatrienoic acid (EETs) to dihydroxyeicosatrienoic acids (DHET) [183]. It plays a physiological role in blood pressure, anti-inflammation, neuroprotection, and cardioprotection [183], and as well as being a target for treating chronic obstructive pulmonary diseases (COPD), atrial fibrillation, and diabetic neuropathic pain, it has had many drugs in clinical trials [184]. The binding site of sEH (see Fig. 5.1(C)) is large and deeply buried [183]. In crystal structures it is partially occluded by a center pinch (CP) formed by two loops coming together. Typical inhibitors of sEH are derived from urea and amide derivatives but other scaffolds include chalcone oxides, carbamates, and acyl hydrazones [183]. In this study we simulate unbinding of the TPPU inhibitor from the piperidyl-urea scaffold family (see Fig. 5.1(A)). The bound state of TPPU is coordinated by three hydrogen bonds from Asp105 and Tyr236 in the back of the sEH binding pocket and Tyr153 near the front of the pocket (see Fig. 5.1(B)) [185, 2].

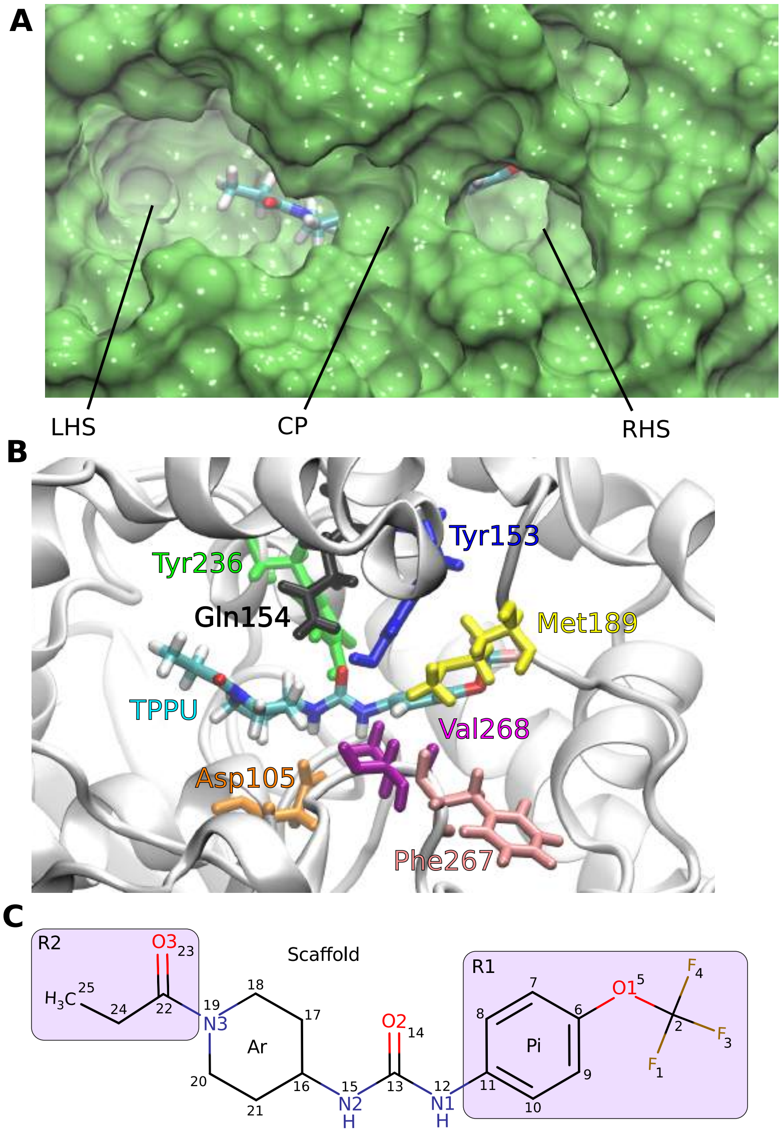


Figure 5.1: Binding site anatomy of sEH and structure of TPPU. (A) Surface representation (1.4 Å probe radius) of the binding pocket of sEH and the TPPU ligand in licorice representation colored by element, from PDB: 4OD0. The left hand side (LHS), the right hand side (RHS), and the center pinch (CP) of the binding site are labeled. (B) Identification of some important amino acids along the unbinding path of TPPU. The three residues with hydrogen bonds in PDB: 4OD0 include two in the back of the binding site: Tyr236 (green) and Asp268 (orange); and Tyr153 (blue) in the upper front. Other residues that form hydrogen bonds during dynamics are Gln154 (black), Met189 (yellow), Phe267 (pink), and Val268 (purple). (C) TPPU with the oxygens, nitrogens, piperidyl ring (Pi), and aromatic ring (Ar) labeled. The two R groups (R1 and R2) from the piperidyl-urea scaffold are highlighted in the boxes. All heavy atoms are labeled according to the serial numbers in PDB: 4OD0.

5.3 Results

5.3.1 Convergence of Sampling

Six WExplore runs were conducted starting in the equilibrated crystallographic binding pose. All of these runs generated sampling poses that differed significantly from the crystal structure, and five runs sampled complete ligand unbinding pathways. Before we begin to analyze our results for the existence of specific molecular determinants of kinetics we first examine the convergence of our sampling with respect to ensemble-averaged observables. For this purpose we created non-equilibrium free energy profiles along two general progress coordinates: i) ligand SASA and ii) the RMSD of the ligand with respect to the initial structure after alignment of the binding sites. These are calculated for all frames in the simulation and cumulative histograms are used to investigate convergence (Fig. 5.2). Each curve shows a histogram of $-ln(\sum_i^n p_i)$ with 30 bins along the progress coordinate, where n is the number of frames in a bin and p_i is the weight of the ith frame, and each curve is shifted vertically so its minimum is equal to zero. Due to differences in the lengths of runs there are 86 112, 172 512, 258 912, 332 829 and 353 180 points, respectively for the curves. This shows that as sampling progresses the average probability of the unbound state stabilizes. The anomalous feature at very low SASAs with very high free-energy (above the cutoff of 35) is the result of a few extremely low weight states where the ligand burrows further into the protein.

5.3.2 Conformation Space Network (CSN) Features

The entire ensemble of conformations generated by all WExplore runs is depicted by a CSN shown in Fig. 5.4, and the individual contributions from each run are shown in Fig. 5.3. The sizes of the nodes correspond to their statistical weight determined from sampling and nodes that are close together tend to interconvert more quickly than nodes that are far apart (see section 5.2.3). This results in a depiction that allows us to visualize the complete free

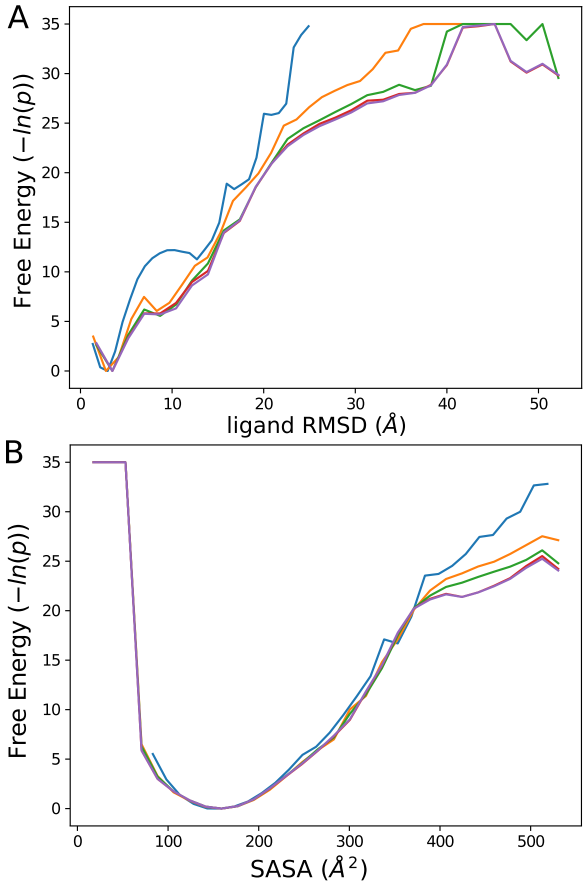


Figure 5.2: Cumulative non-equilibrium free energy profiles of A) ligand RMSD to the initial structure after superimposing the protein binding sites and B) ligand SASA [4]. Each line consists of values taken from the frames of cycles indicated in the legend, *i.e.* in increments of 300 cycles. Raw weights (p) were binned and summed across the progress coordinate values (30 bins) within each set of cycles before computing the free energy, $\mathrm{FE_{bin}} = -ln \sum_i^M p_i$, where M is the number of observations in a bin.

energy landscape sampled by our simulations, as well as identify major collections of states and the paths between them [167, 1, 47, 75, 186].

Using a network modularity algorithm [175] we break up the network into communities, where the connections within a community are stronger than the connections between communities. We will refer to these communities using the labels in Fig. 5.4 (*e.g.* $\mathrm{B_1}$, $\mathrm{B^*}$, $\mathrm{P_1}$, $\mathrm{P_2}$, and U). The spatial density of each community is compared in Fig. 5.7 and shown individually in Fig. 5.6.

The majority of the sampling weight is concentrated in the bound ensemble states ($\mathrm{B_1}$),

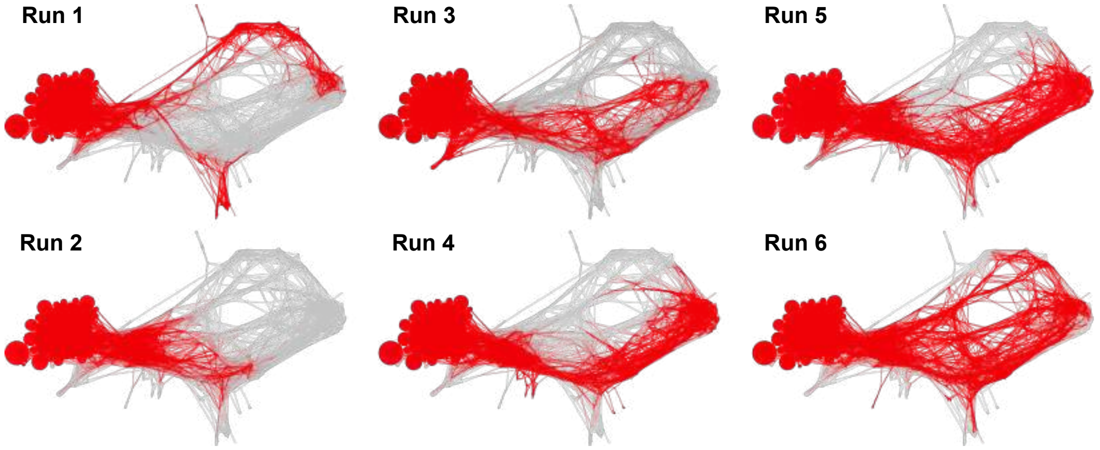


Figure 5.3: The contributions of each individual run to the Conformation State Network (CSN). Nodes (and the connected edges) are colored red for a run in which at least one frame from the run was included in the cluster.

which includes the crystal structure. The B_1 community forms a tight, compact cluster of states both in the network layout (Fig. 5.4), indicating that they interconvert quickly, as well as the spatial density plot (Fig. 5.7 and Fig. 5.6), which additionally shows that the states are localized deep in the binding pocket. The B_1 ensemble can be characterized by its low free energy and low SASA (Fig. 5.8). The slowest interconverting states (identified as having the longest edges in the network) are those in R_1 and R_2 in which the ligand is bound, but has a reversed orientation relative to the crystal structure (Fig. 5.5). The location of these states in the network suggests that they are not directly accessible from B_1, and that the ligand needs to first visit U to access R_1 and R_2. These states demonstrate the breadth of sampling that was obtained with WExplore.

5.3.3 Release from the Deeply Bound State

The deviation in ligand position is very restricted in B_1 as shown in Fig. 5.7, but in the near-bound intermediate community B^*, the ligand position density moves outwardly. To understand the driving forces for this transition we performed an interaction profiling across the network, where every protein-ligand interaction is identified from a representative frame

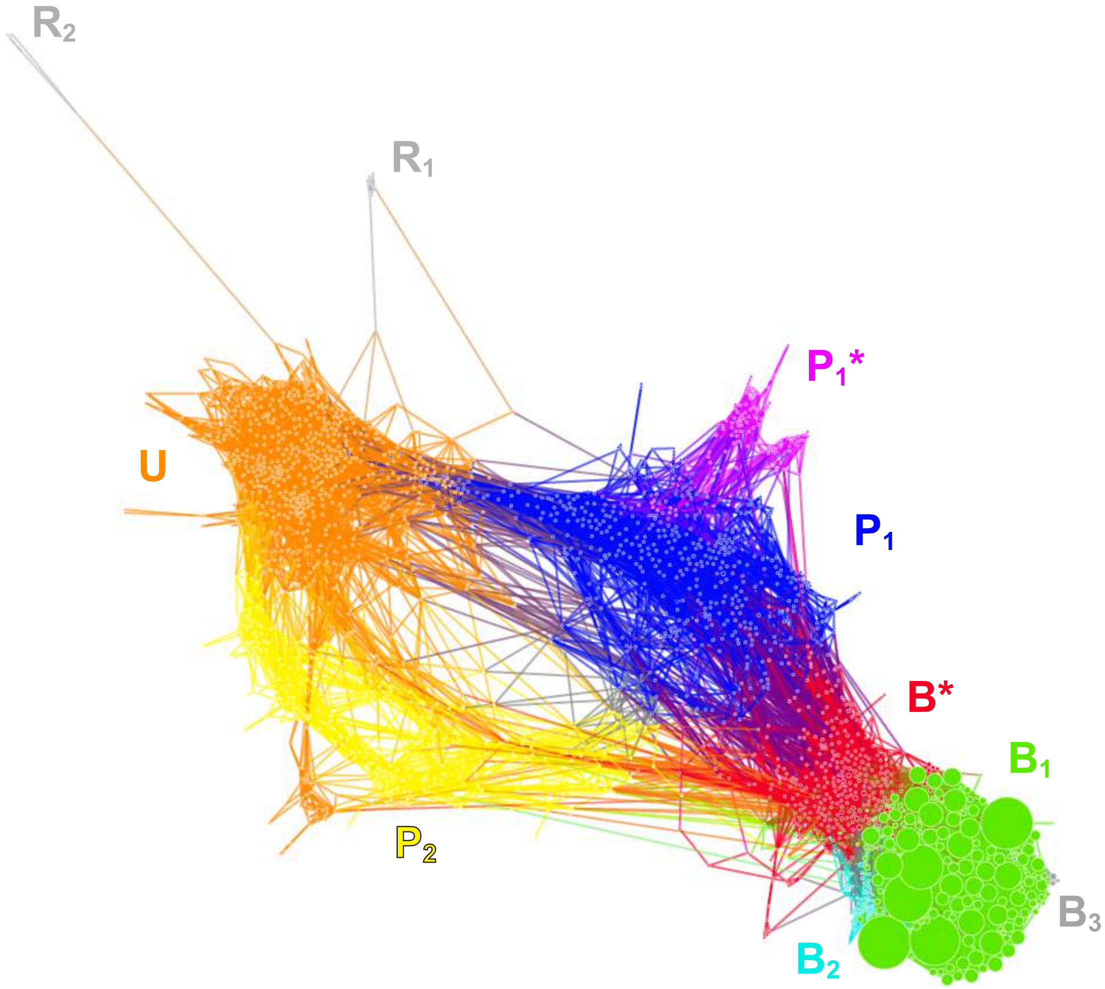


Figure 5.4: A conformation space network shows all poses of TPPU associated to sEH during sampling, including bound, transitioning, and unbound conformations. Each cluster is represented by a node, whose size is proportional to the weight of the cluster. The layout of the nodes and edges was determined using the ForceAtlas2 algorithm [5] in Gephi [6]. The nodes are colored according to their community: B_1 (Primary bound ensemble, light green), B_2 (secondary bound ensemble, cyan), B^* (bound branchpoint ensemble, red), P_1 (primary pathway 1 ensemble, blue), P_1^* (secondary pathway 1 ensemble, purple), P_2 (pathway 2 ensemble, yellow), and U (near-unbound ensemble, orange). Minor communities detected include B_3, R_1, and R_2 (grey). R_1 and R_2 feature ligand poses of reversed orientation relative to the native orientation (see Fig. 5.5). R_2 is removed from other CSN depictions for convenience of display.

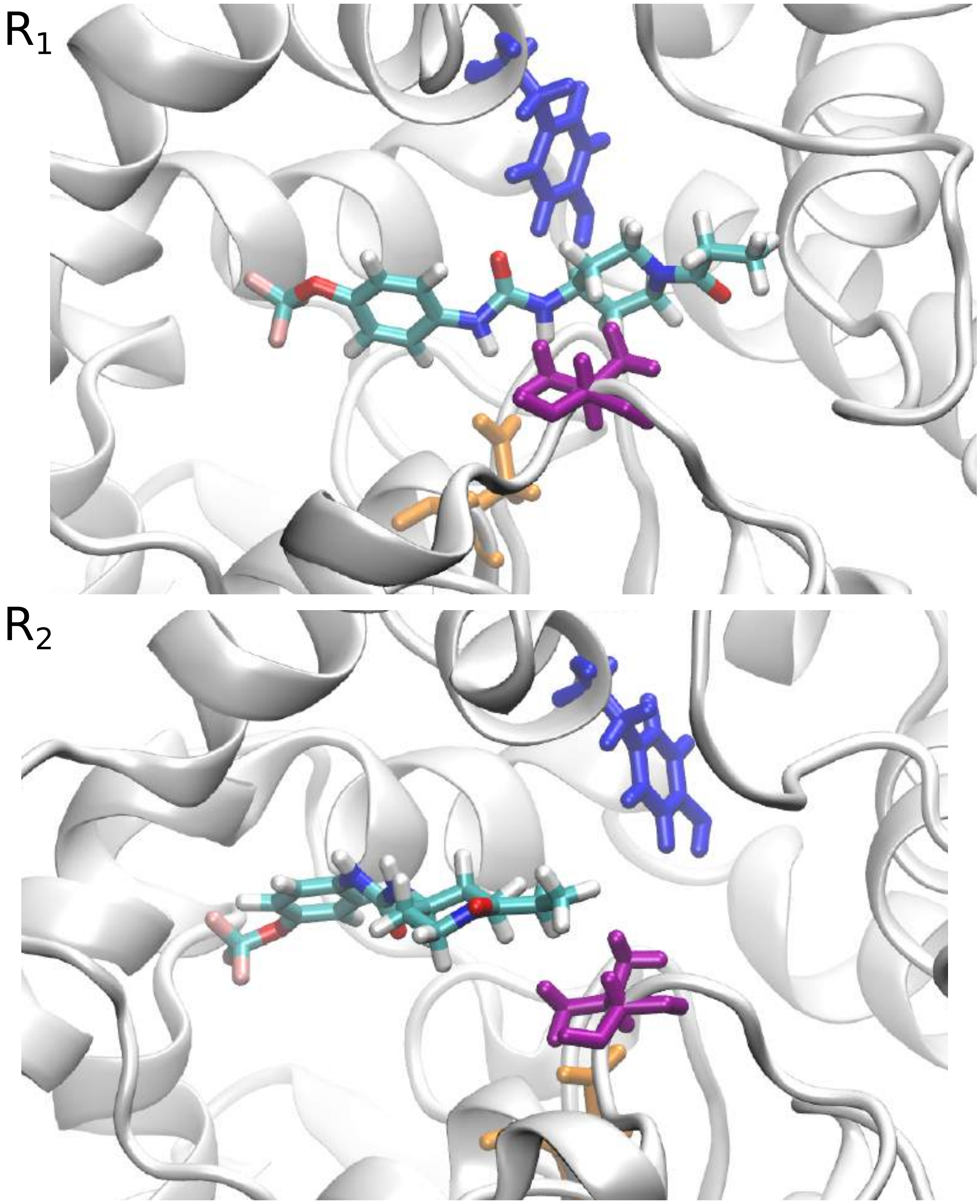


Figure 5.5: Example ligand poses for the two reversed orientation communities R_1 (top) and R_2 (bottom). The protein is shown in right from the front. Tyr153 (blue), Val268 (purple), and Asp105 (orange) are shown for reference.

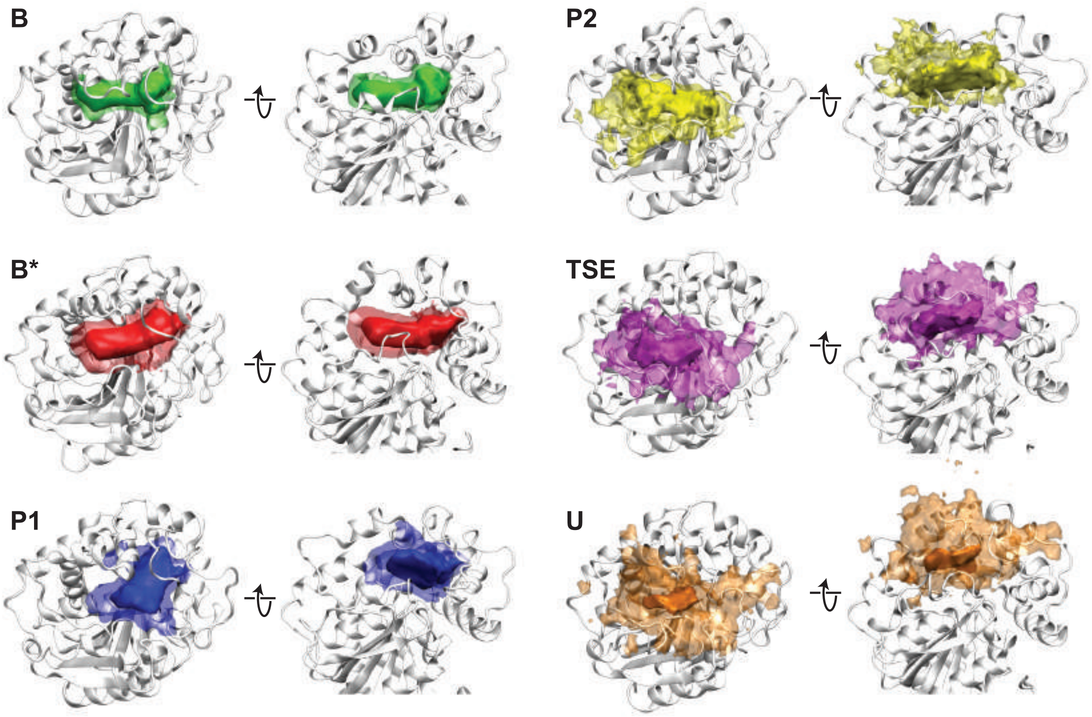


Figure 5.6: Density isosurfaces for selected major communities of the sEH-TPPU network, as well as the transition state ensemble (TSE). Each surface is computed using a single cluster representative for each node in an ensemble. Two density isosurfaces are shown for each: an opaque high density isosurface ($\rho = 0.2$) and a translucent low density isosurface ($\rho = 0.03$). The viewpoints are the same in all panels and show both the front and bottom views as before. One protein conformation from the ensemble is shown for each in white. The high density regions are the same as those shown in Fig. 3. The colors are roughly the same as those in other figures for ensembles as in the main text.

from each node in the CSN. Fig. 5.9 shows ligand position density isosurfaces for some of the CSN's highest frequency ligand-protein hydrogen bonds (Tyr153-OH:Lig-O2, Asp105-OD2:Lig-N2, Val268-O:Lig-N1, and Met189-N:Lig-O2; see Fig. 5.9 caption for explanation of interaction nomenclature), as well as the distribution in the CSN. The frequencies of these interactions are shown separately for each network community in Fig. 5.10. Interactions which occur deep in the binding site (*e.g.* Asp105-OD2:Lig-N2) are shown to be much more compact than interactions with Tyr153. The occurrence of Tyr153-OH:Lig-O2 across the bound ensembles (B_1), near-bound ensemble (B^*), and parts of the path 1 (P_1) and path 2 (P_2) ensembles correspond with the positioning of Tyr153 at the front of the binding site.

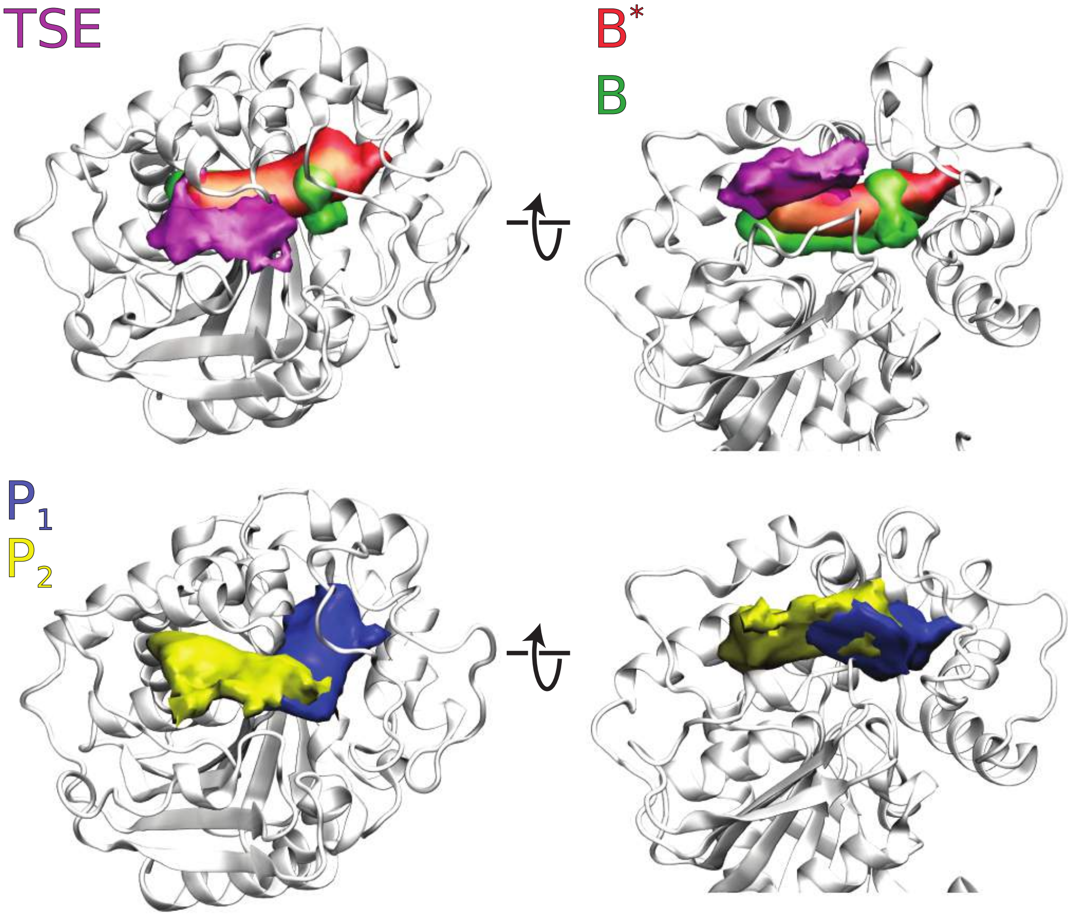


Figure 5.7: Density isosurfaces for various network subsets. Two views are shown for each: the front in the left column, and the bottom in the right column. The upper panel shows two views containing density isosurfaces for the transitions state ensemble (TSE; purple), the deeply bound community ensemble (B; green), and the near-bound community ensemble (B^*; red). The lower panel shows density isosurfaces for the two main exit pathway communities. Path 1 (P1) is shown in blue and Path 2 (P2) is shown in yellow. All surfaces shown at a density of $\rho = 0.2$.

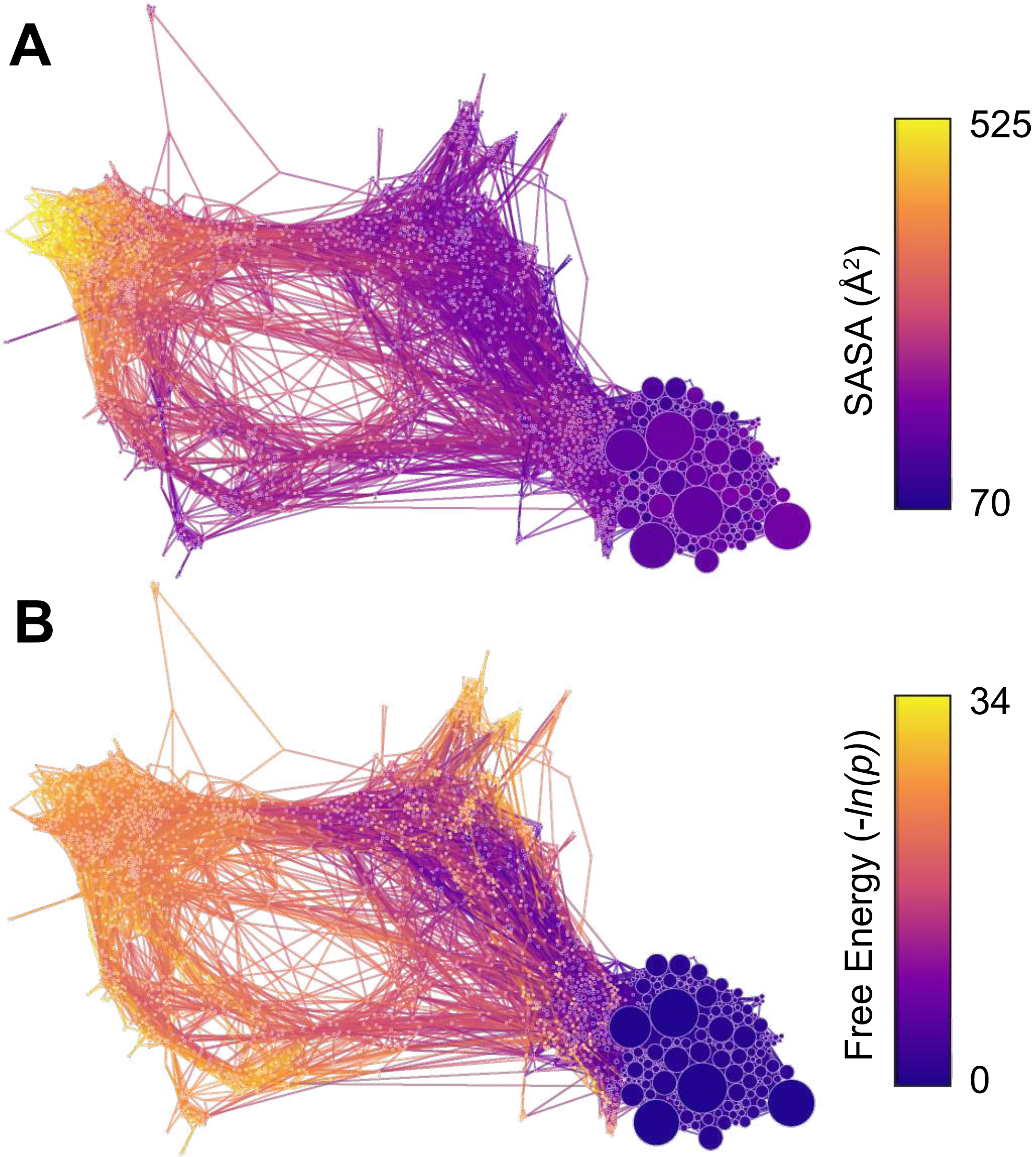


Figure 5.8: Conformation space networks colored according to A) solvent accessible surface area (SASA) of the TPPU ligand ($\mathring{A}^2$) calculated with the Shrake-Rupley method [4], and B) non-equilibrium unbinding ensemble free energy ($-\ln(p)$).

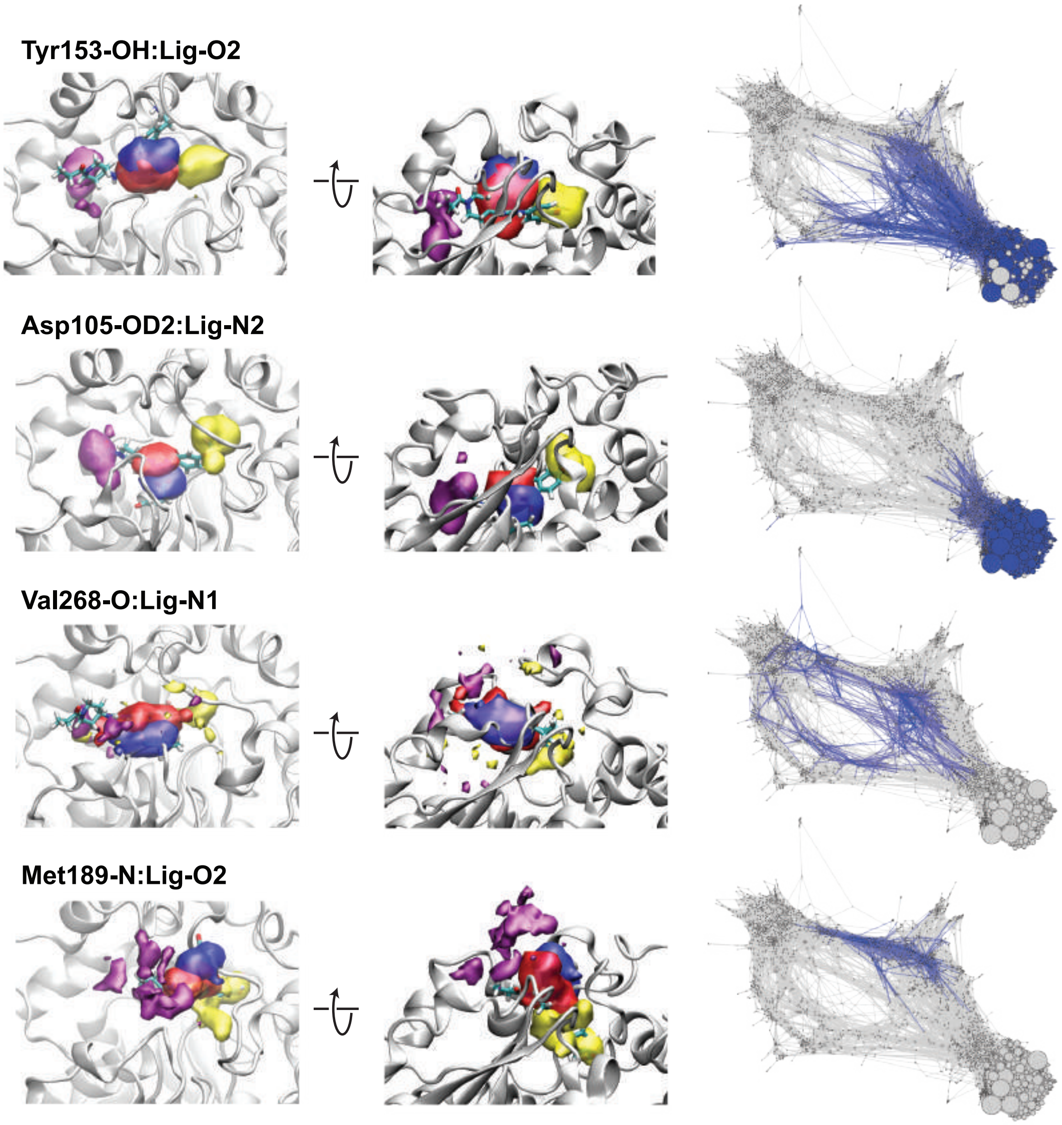


Figure 5.9: Density isosurfaces of the ligand (TPPU) (left and middle) and conformation space network distributions (right) for specific hydrogen bond acceptor-donor pairs. The interaction labels for each row specify the amino acid type and PDB residue index followed by a hyphen (-) followed by the PDB atom type of the acceptor/donor. The ligand (Lig) PDB atom type acceptor/donor follows the colon (:) (see Fig. 5.1(C)). The left column shows the front view of the binding site as shown in Fig. 5.1(A-B). The middle column shows a rotated view of the bottom side of the binding site. The acceptor/donor atom density isosurfaces for the protein are shown in blue and the ligand in red. Two density isosurfaces for distal ligand atoms C16 and C1 are shown in purple and yellow, respectively. The TPPU ligand is shown in licorice representation colored by atom type and the protein backbone is shown as a cartoon in white. Density isosurfaces are calculated from the positions of cluster representatives in which the interaction was detected, and surfaces of density $\rho = 0.025$ are shown. In the right column nodes in the conformation space network (CSN) are colored blue (as well as adjacent edges) by the presence of the interaction in the cluster representative. Nodes in which the interaction was not observed in the cluster representative are in grey.

Similarly, Asp105-OD2:Lig-N2 (Fig. 5.9) and Tyr236 interactions (not shown), are focused in the bound ensemble and part of B^* due to the deep location of those residues. Thus it is likely that the breaking of hydrogen bonds to Asp105 and Tyr236 – which are found in the crystal structure – are necessary first steps in unbinding, while the positioning and flexibility of Tyr153 may stabilize this outward motion.

The role of Tyr153 as a stabilizer of dynamics is supported by the frequency patterns of all hydrogen bond interactions in each community shown in Fig. 5.10. Comparing B_1 and B^* we see that the highest frequency interaction, Tyr153-OH:Lig-O2, is represented in B^* at a higher frequency than in B_1 across network nodes. The other high frequency interactions in B_1 correspond to deep contacts with Asp105 and Tyr236, and are found in B^* at a lower frequency.

The sharp increase in binding site SASA seen at committor probabilities between 0.0 to 0.2 in Fig. 5.12 corresponds to B^* (Fig. 5.15). This suggests that concomitant with the breaking of deep native contacts to Asp105 and Tyr236 the binding site becomes more exposed to solvent. This transition is potentially a high free-energy barrier to unbinding, especially given the hydrophobicity of the binding site [2], and Tyr153 is likely crucial in stabilizing the outward motion of the ligand. Furthermore, B^* also acts as a branching point before the ligand commits to an exit path (*i.e.* P_1, P_2).

5.3.4 Ligand Exit Pathways

Due to the shape and width of the sEH binding pocket [2, 183, 25] there is a range of possible exit trajectories. Here exit trajectories are centered along two dominant pathways, P_1 and P_2, although we also find a fair number of states between the two pathways. P_1 trajectories are closer to the right hand side (RHS) of the pocket and P_2 trajectories are closer to the left hand side (LHS) of the pocket (see Fig. 5.7).

These two exit pathways are not topologically distinct relative to the protein backbone, as we found previously for benzamidine exit pathways from trypsin [47]. The two major

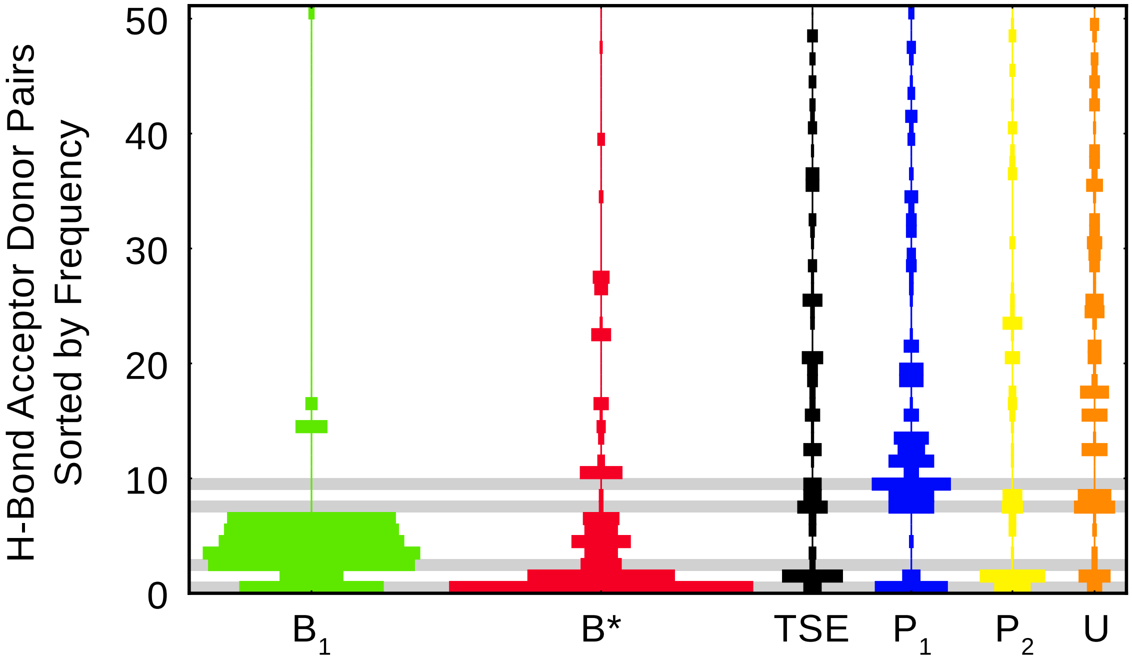
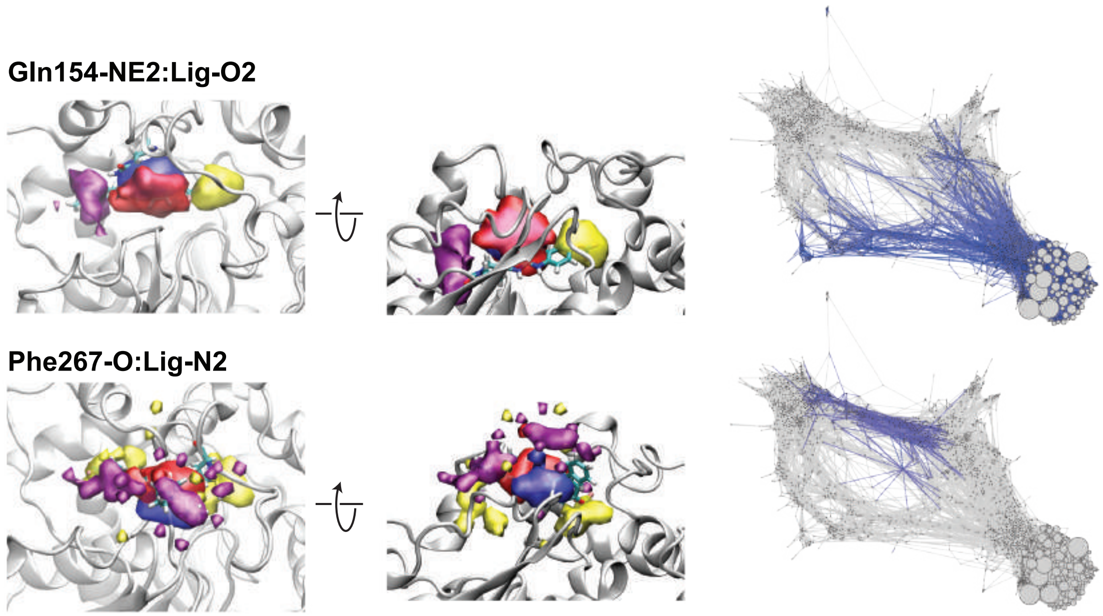


Figure 5.10: Interaction frequency violin bar graphs for the top 50 highest frequency interactions. Horizontal bars are for individual hydrogen bond acceptor-donor pairs, which are indexed by the vertical axis and sorted by total frequency from the bottom to the top. Violins are shown for each major community as labeled on the horizontal axis, and the color corresponds to those in Fig 5.4, except for the Transition State Ensemble (TSE) violin which is taken from Fig. 5.15. The total width of the bar corresponds to the total frequency of the interaction in the community. The gray highlighted rows are, from bottom to top are Tyr153-OH:Lig-O2, Asp105-OD2:Lig-N2, Val268-O:Lig-N1, and Met189-N:Lig-O2, which are detailed in Fig. 5.9 and some others not highlighted are detailed in Fig. 5.11.

modes of exit are likely due to the pinched center (CP) of the "lips" of the binding pocket which likely steers the ligand through the more open RHS and LHS corners of the "mouth" (Fig. 5.1). P_1 is more favorable in terms of free energy (Fig. 5.8), and is the shortest path through 3D space from the crystal structure state. P_1 is also traversed along the highest flux pathway through the CSN according to the transition matrix, as shown in Fig. 5.13 and 5.14. Exit through P_2 first requires a sliding motion in the bound state where a hydrogen bond with Asp105 is shifted between urea moiety nitrogens (N2 to N1) on TPPU. This exchange is not required in P_1, and seems to be associated early on with interactions with Gln154 (see network plot in Fig. 5.11). We note however that analyzing individual pathways


Gln154-NE2:Lig-O2

Phe267-O:Lig-N2

Figure 5.11: Density isosurfaces and network distribution of two additional interaction classes: Gln154-NE2:Lig-O2 (top) and Phe267-O:Lig-N2 (bottom). All other details are the same as in Fig. 5 in the main text.

through the network or from continuous trajectories is likely misleading as can be seen by the low individual contributions of individual paths shown in Fig. 5.14. The bottleneck of the highest flux pathway explains only 0.8624 % of the flux through the network, the next highest being 0.5934 % and a sum of 367 such bottlenecks is needed to explain 80 % of the total flux through the network.

5.3.5 Transition State Stabilizing Interactions

Focusing on P_1 we find that concomitant with the breakage of Tyr153 hydrogen bonds, hydrogen bond interactions are formed between the ligand and the backbone nitrogens of Met189 (Fig. 5.9) and Phe267 (Fig. 5.11). Both of these residues are located on the "lips" of the binding site, whereas Met189 is on the far RHS, Phe267 is more towards the center, and adjacent to Val268 which is located directly at the center pinch (Fig. 5.1). The formation of a P_1-specific interaction pattern that is distinct from the P_2 interactions can be seen in Fig. 5.10. For both P_1 and P_2, the total number of states with specific interactions is much

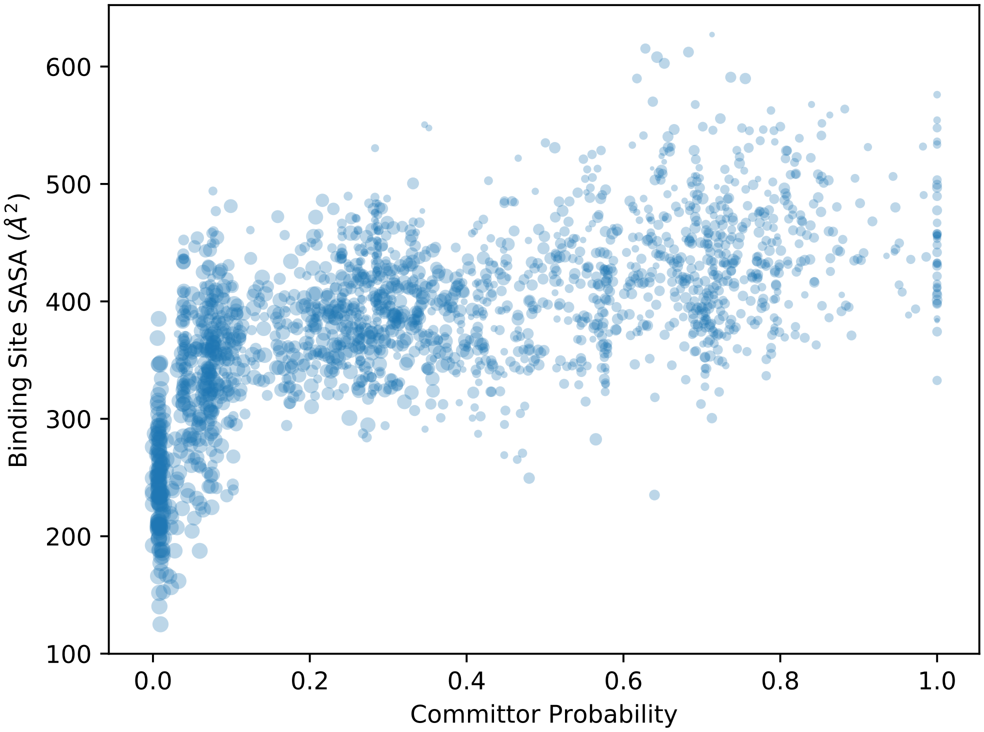


Figure 5.12: Plot of the binding site solvent accessible surface area (SASA) (Å^2) [4] for every cluster representative in the network over their committor probabilities. The size of the points is equal to the negative of the free energy ($-\ln(p)$) of the cluster translated to the positive domain. The binding site was defined to be all protein atoms which were within 4 Å from any ligand atom in the PDB 4OD0 crystal structure. The Shrake-Rupley method was used for calculating SASA. Clusters which were not assigned a committor probability are not included.

less than in B_1 or B^*, likely due to the higher solvent exposure (Fig. 5.8). Interestingly, while the ligand clouds of the residues located near the center pinch, like Val268, are more spread out and heterogeneous, the Met189 ligand cloud is fairly dense and tightly focused on the RHS corner with the ligand turned perpendicular relative to the bound pose. This ligand-protein interaction seems to be a key step in the ultimate release of the ligand as this interaction is found in states directly adjacent to the U ensemble. We summarily observe that for the P_1 pathway, interactions with Tyr153 and Met189 (*e.g.* Tyr153-OH:Lig-O2 and Met189-N:Lig-O2) play major roles in stabilizing the step-wise unbinding of TPPU from

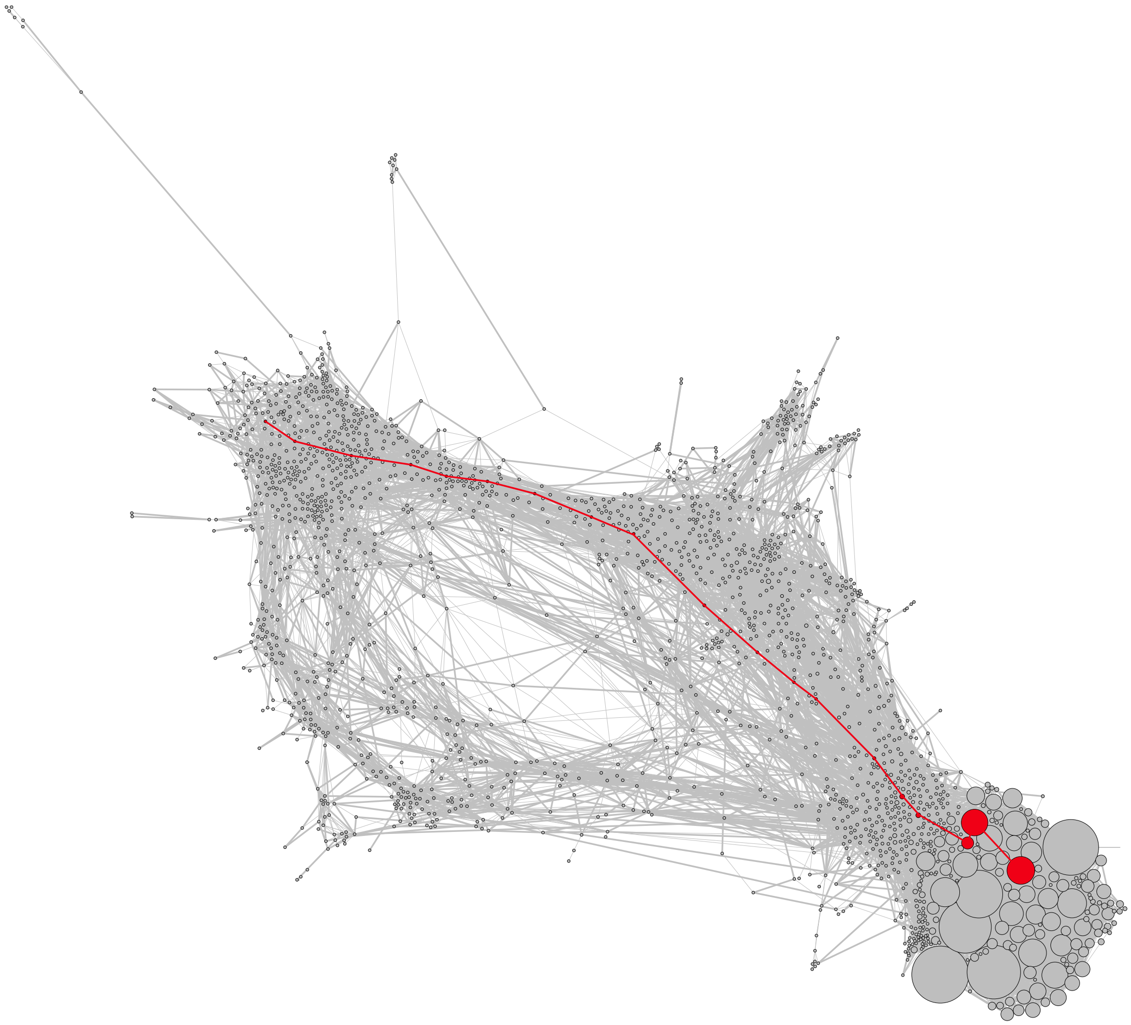


Figure 5.13: The highest flux pathway through the network is shown in red.

sEH.

The implication of Tyr153 or Met189 as the key interactions in stabilizing the overall unbinding transition state is an important one, as destabilization of these interactions could result in longer ligand residence times. Although the positioning and flexibility of Tyr153 seems to be vital for unbinding, it is also a native contact and contributes significantly to ligand affinity, and thus not a favorable target for disruption. Met189 on the other hand is a much more desirable target, because it is not a native contact and could likely be disrupted with little impact on affinity. To ascertain the role these residues in kinetics, we predicted the transition state by first calculating the committor probabilities in a Markov state model

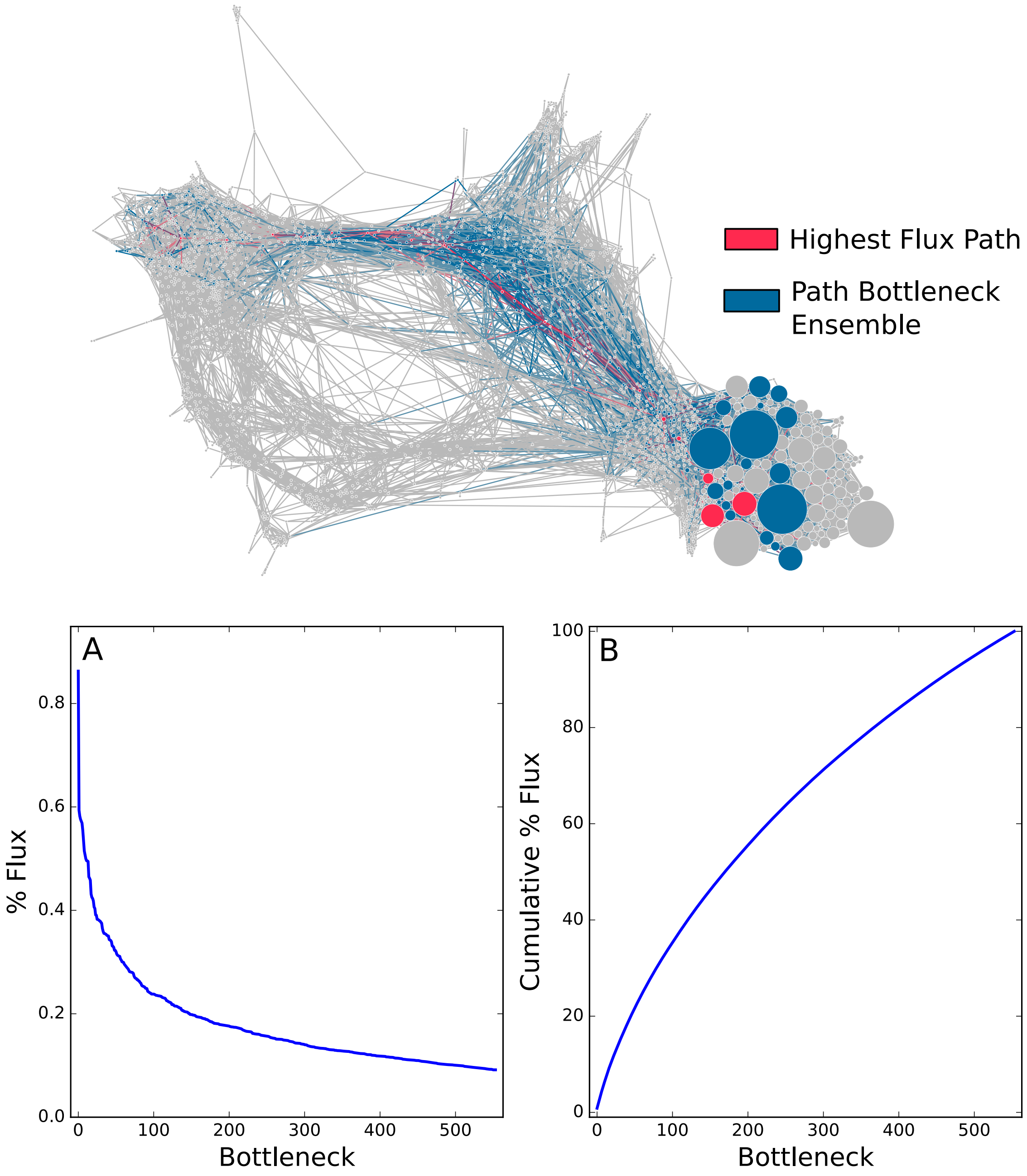


Figure 5.14: The top panel shows the CSN with the highest flux pathway colored in red and the bottleneck ensemble of the highest flux path in blue. In the lower panel A) shows the percent forward flux through the CSN accounted for over the inclusion of more bottlenecks, ordered from highest to lowest flux and shows the cumulative percent flux over the bottlenecks.

framework [187], shown in Fig. 5.15 using the bound source and unbound sink basin nodes shown in Fig. 5.15(B). We then identified states with committor probabilities between 0.4 and 0.6 and use this as our prediction of the transition state ensemble (TSE) shown in Fig. 5.15(B). From this we can immediately tell that the TSE lies closer to U than to B_1 and is colocalized closely with the distribution of Met189 interactions (from Fig. 5.9). This suggests that Met189 is a transition state stabilizer, potentially along with Val268 and Phe267. The identification of transition state stabilizing interactions is important for kinetics-oriented drug design [12] and the disruption of such interactions has already been used to produce slow-onset inhibitors with very long residence times [188, 37, 189].

5.3.6 Unbound States and Unbinding Rates

The near-unbound ensemble (U) has the most varied ligand positions, as expected, and includes fully unbound poses as well poses around the outside of the "lips" of the binding site, potentially highlighting residue interactions (Fig. 5.10) that might be important for the initiation of binding events. While the other communities have mapped reasonably well to distinct phases of unbinding (*e.g.* B_1, P_1, etc.), the U ensemble is not merely the collection of unbound clusters, as can be seen by the protrusion below P_2 in Fig. 5.4. This is a result of the complexity of the CSN topology in the U region and the fact that the communities are determined using only properties of the network as a graph and not properties of the molecular structure. Additionally, as trajectories leave the surface of the protein, they are killed and their weight is recorded along with the time elapsed. Thus states near such an absorbing boundary are not sampled as thoroughly as the other communities and consequently fewer edges connect those states. Accordingly states in the U ensemble are likely short lived and weight is not able to accumulate and lower the free energy, even though, thermodynamically, an unbound solvated ligand in U should have a lower free energy than a transition state pose. This explains the high free energies shown in Fig. 5.8.

The exit times and weights of the exited walkers are used to calculate the rate constant

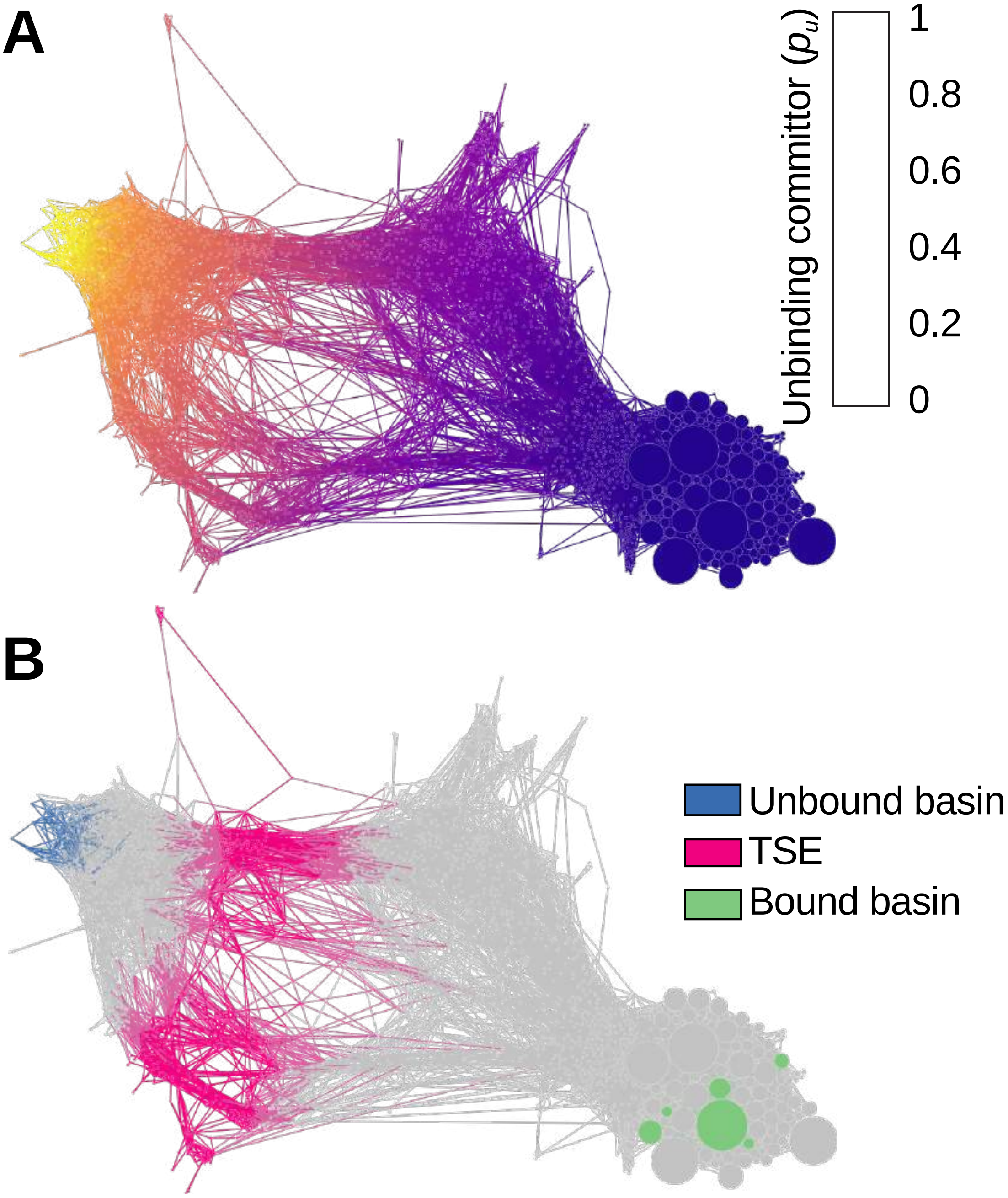


Figure 5.15: CSN showing A) forward committor probabilities (probability of unbinding, $\mathrm{p_u}$) for each node, and B) the bound and unbound basins used to compute the committor probabilities and the transition state ensemble (TSE) where $0.4 \leq p_u \leq 0.6$.

predictions shown in Fig. 5.16 (see Methods section ??). Details of the number of exit points and descriptive statistics of exit point weights in each run are shown in Table 5.1, the 3D positions of the exit points are shown according to run in Fig. 5.17, and the frequency of exit points according to free energy is shown in Fig. 5.18. The spatial distribution of exit points is fairly broad even within runs suggesting not all exit points are highly correlated. The variation in cumulative unbound probability between runs is shown in Fig. 5.16(A) as a function of simulation time. The average cumulative unbound probability is shown as a thick black line, with its standard error in blue. The final total aggregated probability was $3.8 \times 10^{-10} \pm 3.1 \times 10^{-10}$. Aggregate probabilities and passage times are used to calculate the residence time prediction, as well as its standard error, shown in Fig. 5.16(B). From the observed 75 exit points we calculate a residence time of 42 s, compared to the experimentally observed residence time of 660 s (11 min) [2], which agrees to within 1 to 2 orders of magnitude and is just outside the standard error (which ranges from 23 s to 280 s) of our calculation. To assess the convergence of this residence time we performed subsampling on all possible combinations of the different runs and recalculated the resulting residence time (Fig. 5.16(C)). We find that while individually most runs are inaccurate, the average over multiple runs tends to converge towards the experimental value. This is comparable to results from trypsin-benzamidine residence time calculations, which is a 10^3 s faster process [47]. The variation between runs in the number of collected unbinding events, however, varies significantly (Fig. 5.16(A)) and further sampling would be necessary to obtain more precise predictions.

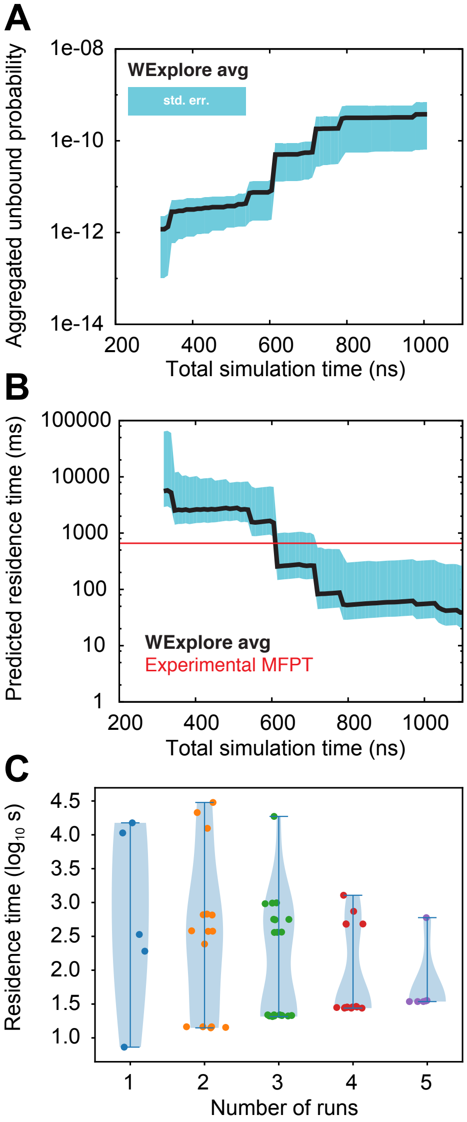


Figure 5.16: A) The aggregated probability of unbound states across all WExplore runs. The average aggregated unbound probability of all runs is shown in the black line, with the standard error in light blue as a function of simulation time. The total aggregated weight was $3.8 \times 10^{-10} \pm 3.1 \times 10^{-10}$, and the final residence time was 42 s with standard error ranging from 23 s to 280 s. The aggregated probability in individual runs are shown in grey. No exits were observed for run 2. B) The predicted residence time of the ligand (black with standard error in light blue) vs. the simulation time passed. This is compared to the experimentally measured residence time shown by the red line. C) Violin plots for residence times calculated for each combination of possible subsamples. Run 2 is excluded from the single run rate calculations because it has no exit points and would have an infinite residence time. Residence time is shown on the vertical axis and the number of runs for each sub-sample is shown on the horizontal axis. Individual residence time values are plotted as dots with a small amount of jitter to make them visible. The violin contour is produced using a Gaussian Kernel Density.

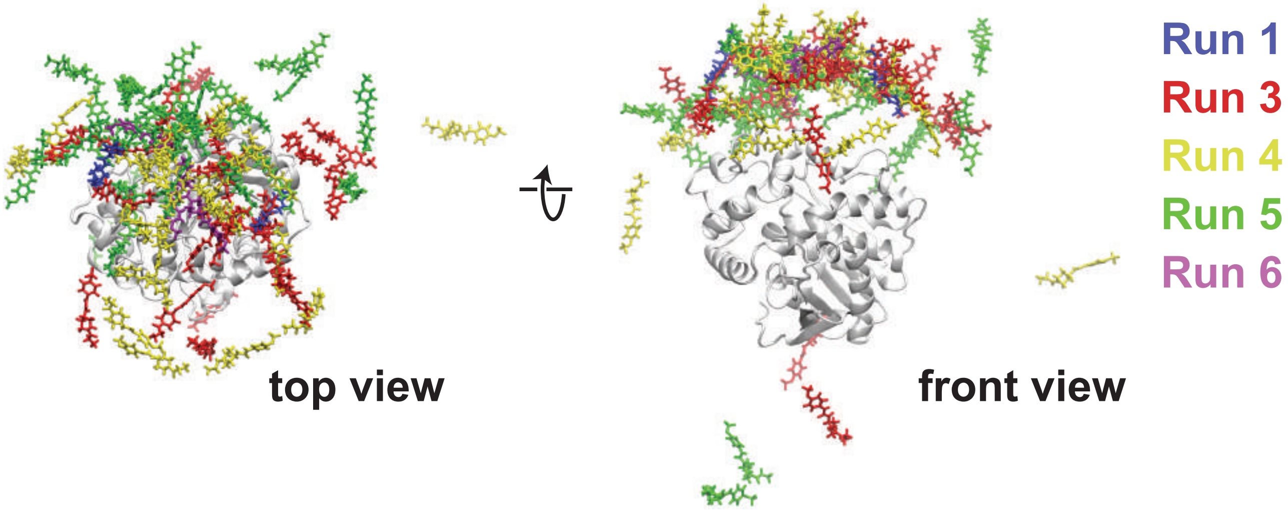


Figure 5.17: Poses of ligands at the exit points colored by run. Ligands are shown relative to the crystal structure in white. This shows a wide variation in geometric positions of the ligands within and among different runs.

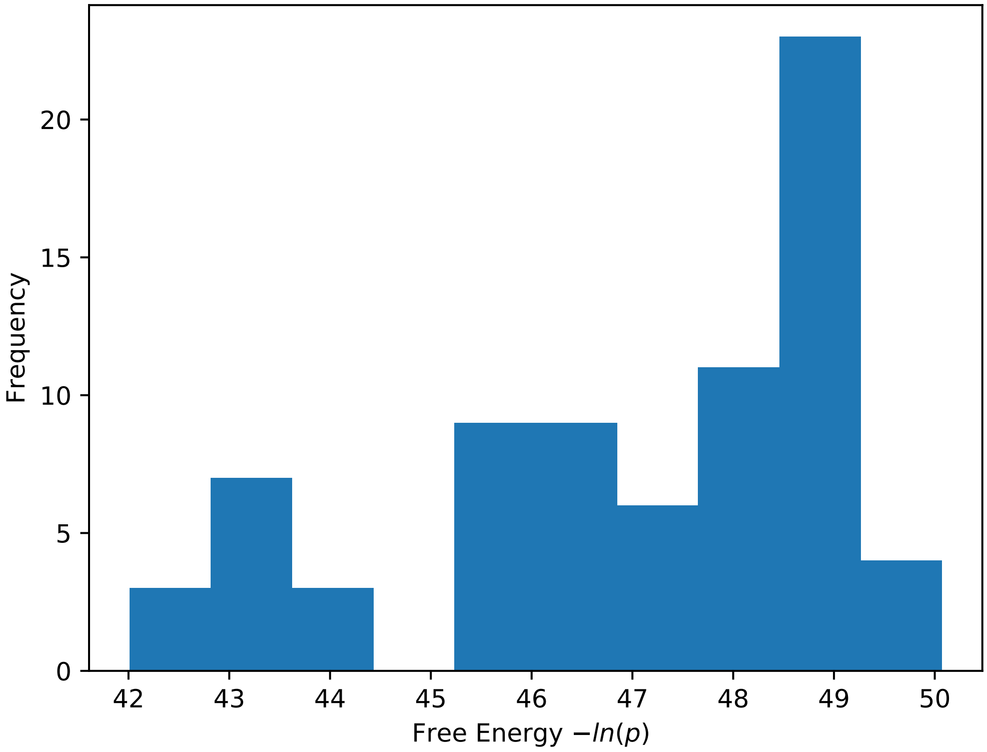


Figure 5.18: Frequency histogram of the exit points binned over their free energies $-\ln(p)$.

Table 5.1: Table of statistics of exit point weights in each run they were observed.

	count	mean	std	min	max
Run Index					
1	3	4.014382e-13	2.818670e-13	2.387022e-13	7.269102e-13
3	20	1.937678e-10	2.180761e-10	7.852805e-12	7.538693e-10
4	22	3.387088e-12	4.199034e-12	5.468285e-13	1.749851e-11
5	27	4.381838e-12	6.320464e-12	3.645440e-13	2.333082e-11
6	3	5.647383e-13	1.362082e-13	4.458460e-13	7.133536e-13

5.4 Discussion

Lee *et al.* [2] examined a series of inhibitors with different substituents for the R2 group of TPPU (see Fig. 5.1). It was observed that larger, more hydrophobic moieties had longer residence times and higher affinities as summarized in Fig. 5.19 (data from [2]). Thus, the increase in hydrophobic bulk certainly lowers the free energy of the bound state relative to the unbound and to the TS. This is intuitive as the R2 group sits in a pocket that is both solvent exposed and hydrophobic in the crystallographic bound state. This does not address the potential (de)stabilizing effects of the hydrophobic modifications to the TS. From our simulation data we may begin to probe whether or not this is plausible by observing correlations between ligand and binding site solvation with committor probabilities around the TS of unbinding.

When we plot the SASA of the entire TPPU ligand as a function of committor probability (Fig. 5.20) we observe a sharp increase in SASA at intermediate committor probabilities ($0.4 \leq p_u \leq 0.6$) around the predicted transition state. This transition is colocalized with the TS more than the binding site SASA (Fig. 5.12), which has a sharp increase only at low committor probabilities ($0.0 \leq p_u \leq 0.2$). In contrast, other potentially useful observables such as the distance between the centers of the two lips of the binding site (Fig. 5.21), show no such correlation. However, when we plot the SASA for only the TPPU-R2 group this trend disappears as shown in Fig. 5.22. This shows that TPPU-R2 solvation is not correlated with the transition state, and from Fig. 5.23 we see that the SASA trend for all non-R2

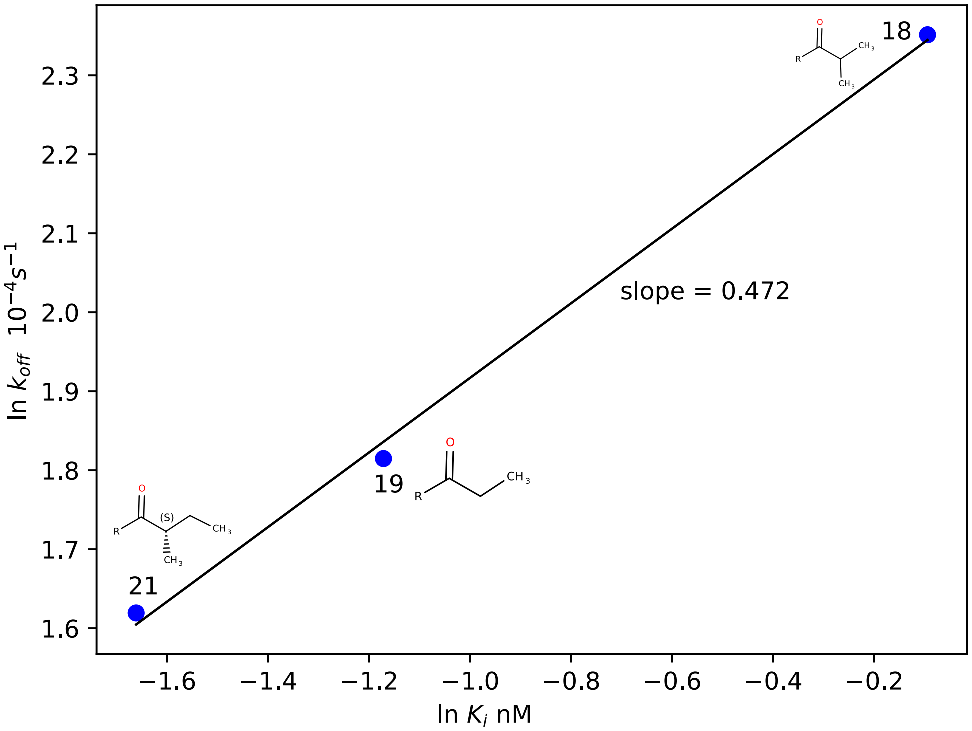


Figure 5.19: Plot of unbinding rates (ln k_{off} (10^{-4} s^{-1})) vs. the inhibition constants (ln K_i (nM)) of inhibitor ligands from Lee *et al.* [2] that have the same scaffold (i.e. same non-R2 structure) but different R2 substituents. The structure of the R2 substituents are pictured next to each point. The result of a linear regression is plotted in black and is shown to have a slope of 0.472. For this series the relationship between k_{off} and K_{i} is linear ($R^2 = 0.998$), at least for this small sample size. When considering all inhibitors of sEH measured in [2] this particular series is a much better fit. $R^2 = 0.999$ for the untransformed data shown in this figure, compared to an $R^2 = 0.440\,21$ when considering all measured inhibitors in the same region of values for k_{off} and K_i.

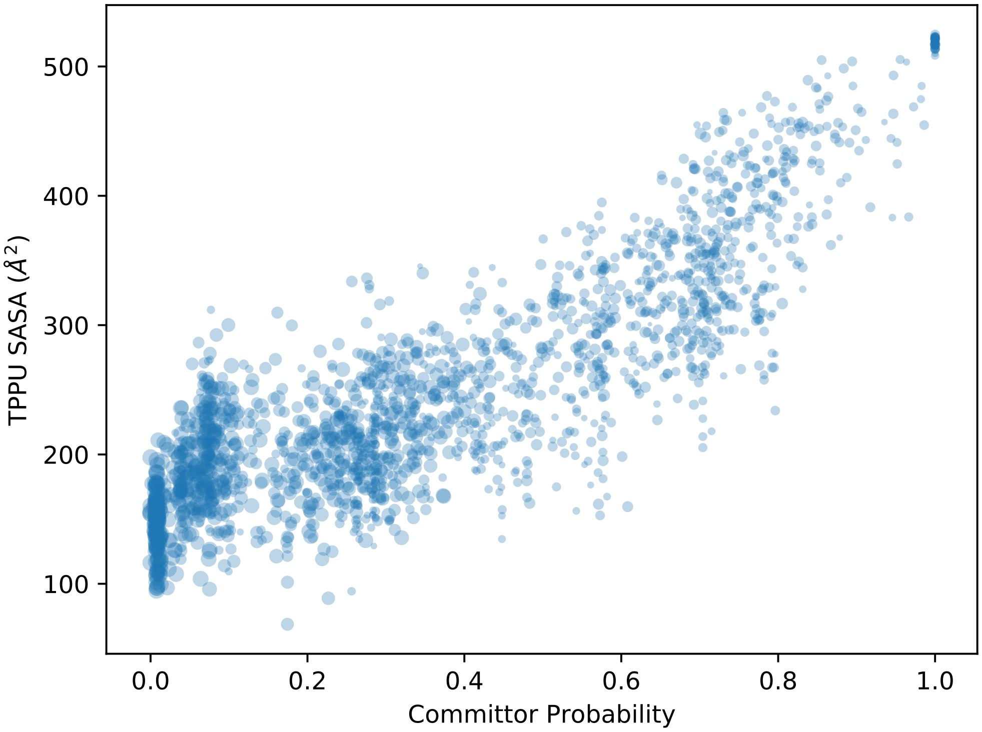


Figure 5.20: Plot of the ligand TPPU solvent accessible surface area (SASA) ($\AA^2$) [4] for every cluster representative in the network over their committor probabilities. The size of the points is equal to the negative of the free energy ($-\ln(p)$) of the cluster translated to the positive domain. Clusters which were not assigned a committor probability are not included.

TPPU atoms recapitulates the trend for the whole ligand shown in Fig. 5.20. Interestingly, when TPPU-R2 SASA is plotted on the CSN (Fig. 5.24) we see a difference between P_1 and P_2 suggesting the role of R2 is heterogeneous across the global transition state, and perhaps plays dual roles in different unbinding pathways.

When compared to other receptors of importance to drug design, the binding site of sEH is similar in some ways to both G-protein coupled receptors (GPCRs) (*e.g.* [49], PDB: 4EIY) and kinases (*e.g.* [52], PDB: 3G5D and 1KV2). GPCRs have a well-defined long, deep binding site with a small orifice and binds the ligand very tightly leaving very low SASA. In contrast, kinases have a more open binding region between two domains with high

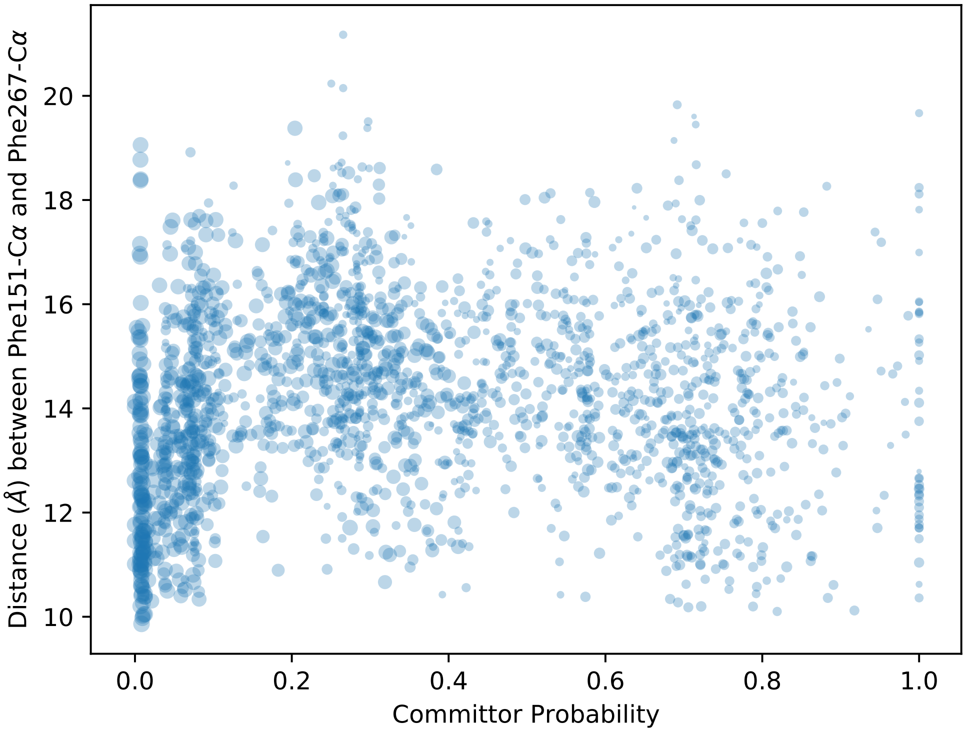


Figure 5.21: Plot of the distance between an atom from the center of both the top lip of the sEH binding site (Phe151-Cα) and the bottom lip (Phe267-Cα) for every cluster representative in the network over their committor probabilities. The size of the points is equal to the negative of the free energy ($-\ln(p)$) of the cluster translated to the positive domain. Clusters which were not assigned a committor probability are not included.

SASA. Comparably sEH has a well defined binding pocket like a GPCR, but is shallower and wider and is not occupied completely by the ligand. Binding pockets and channels which are somewhat restricted, like in GPCRs and sEH, are potentially more amenable to kinetics oriented drug design because the unbinding TS is likely more well defined than that of the open kinases.

In this study we identified two distinct unbinding paths that are not topologically separated, as were those found for trypsin [47, 46, 44] and for type-II kinases [52]. Destabilizations to the TS through ligand modification become more complicated in systems with multiple unbinding pathways as each transition state likely involves different interactions with the

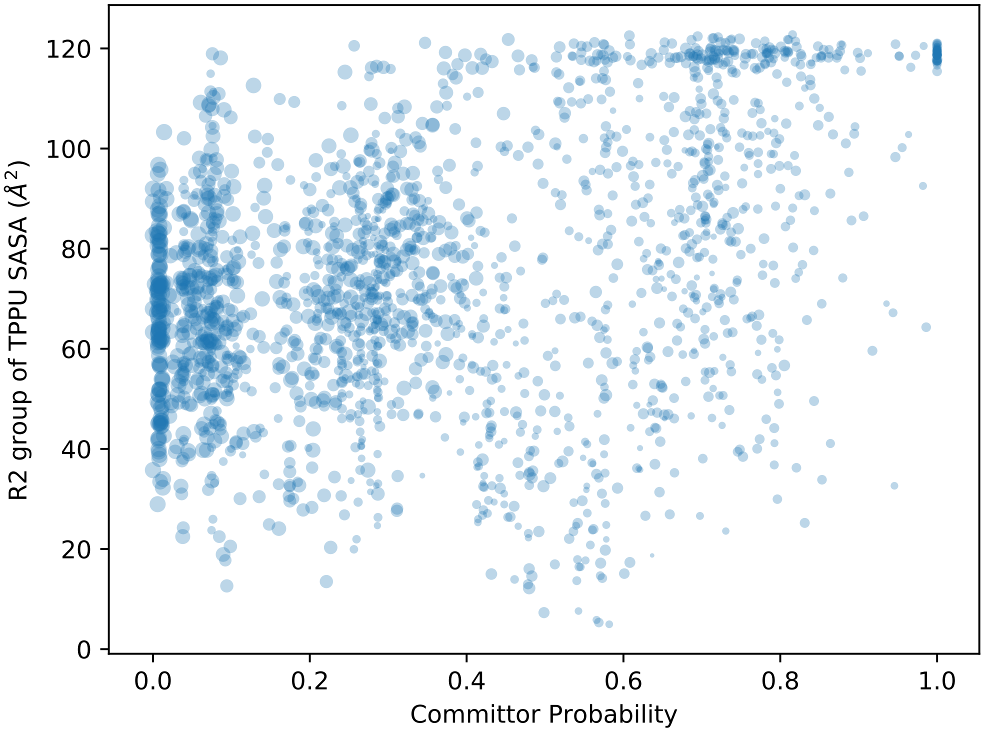


Figure 5.22: Plot of solvent accessible surface area (SASA) (Å^2) [4] for the atoms in the R2 substituent of the TPPU ligand, for every cluster representative in the network over their committor probabilities. The size of the points is equal to the negative of the free energy ($-\ln(p)$) of the cluster translated to the positive domain. Clusters which were not assigned a committor probability are not included.

ligand. Indeed, here we find an intersection of 13 specific hydrogen-bonding interactions between P_1 and P_2, from the total union of 76, in which there is no correlation between pathway ensembles as shown in Fig. 5.25. In systems with multiple major unbinding mechanisms one will be the globally lowest free-energy TS and will contribute the most to the unbinding flux. If the transition states of multiple unbinding pathways are comparable in free energy, then perturbations to drug-TS interactions could potentially shift which pathway is the dominant unbinding path, a phenomenon we name **pathway hopping**. Notably, this phenomenon has been observed previously in the context of protein folding [190]. The complications pathway hopping adds to drug design is an important consideration to studies of drug binding kinet-

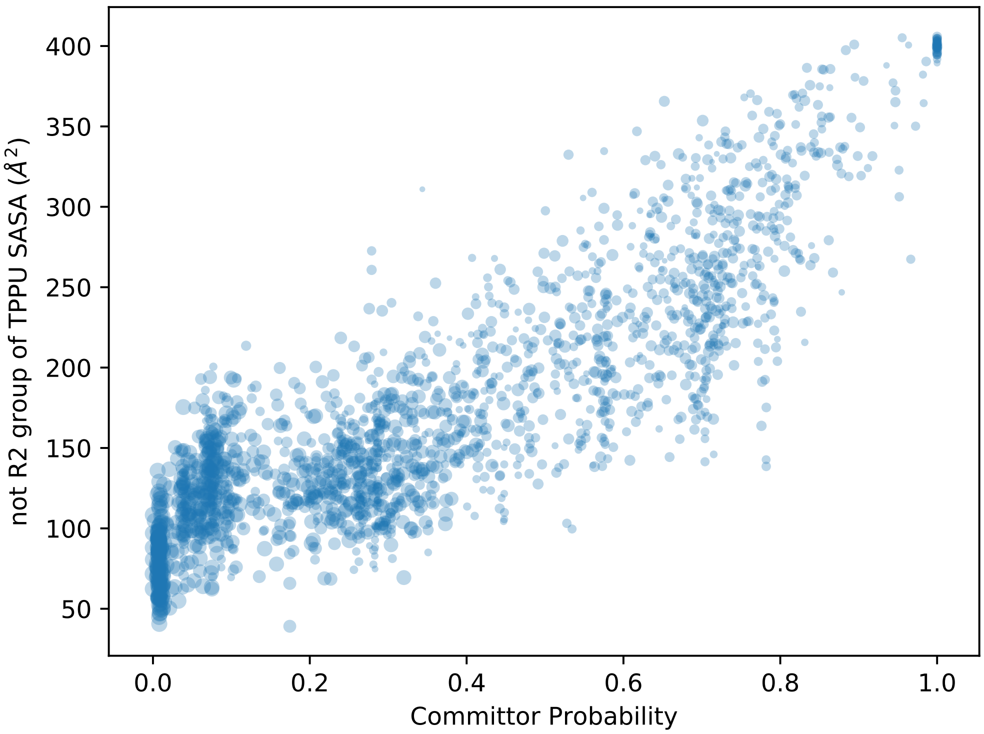


Figure 5.23: Plot of solvent accessible surface area (SASA) ($\AA^2$) [4] for all of the atom not in the R2 substituent of the TPPU ligand, for every cluster representative in the network over their committor probabilities. The size of the points is equal to the negative of the free energy ($-\ln(p)$) of the cluster translated to the positive domain. Clusters which were not assigned a committor probability are not included.

ics, and emphasizes the importance of methods which identify the full spectrum of unbinding pathways. For instance, as suggested above, larger, more hydrophobic R2 groups could stabilize the P2 pathway while destabilizing the P1 pathway, potentially changing the set of TS-stabilizing interactions to consider during drug design. WExplore is particularly useful in this regard because it emphasizes the discovery of these alternative pathways, and may preemptively address drug-design concerns. A discovery-oriented sampling regime will likely improve the robustness of drug-design for kinetics by both anticipating pathway hopping, and by providing unbiased information for the construction of better order parameters as suggested in [191].

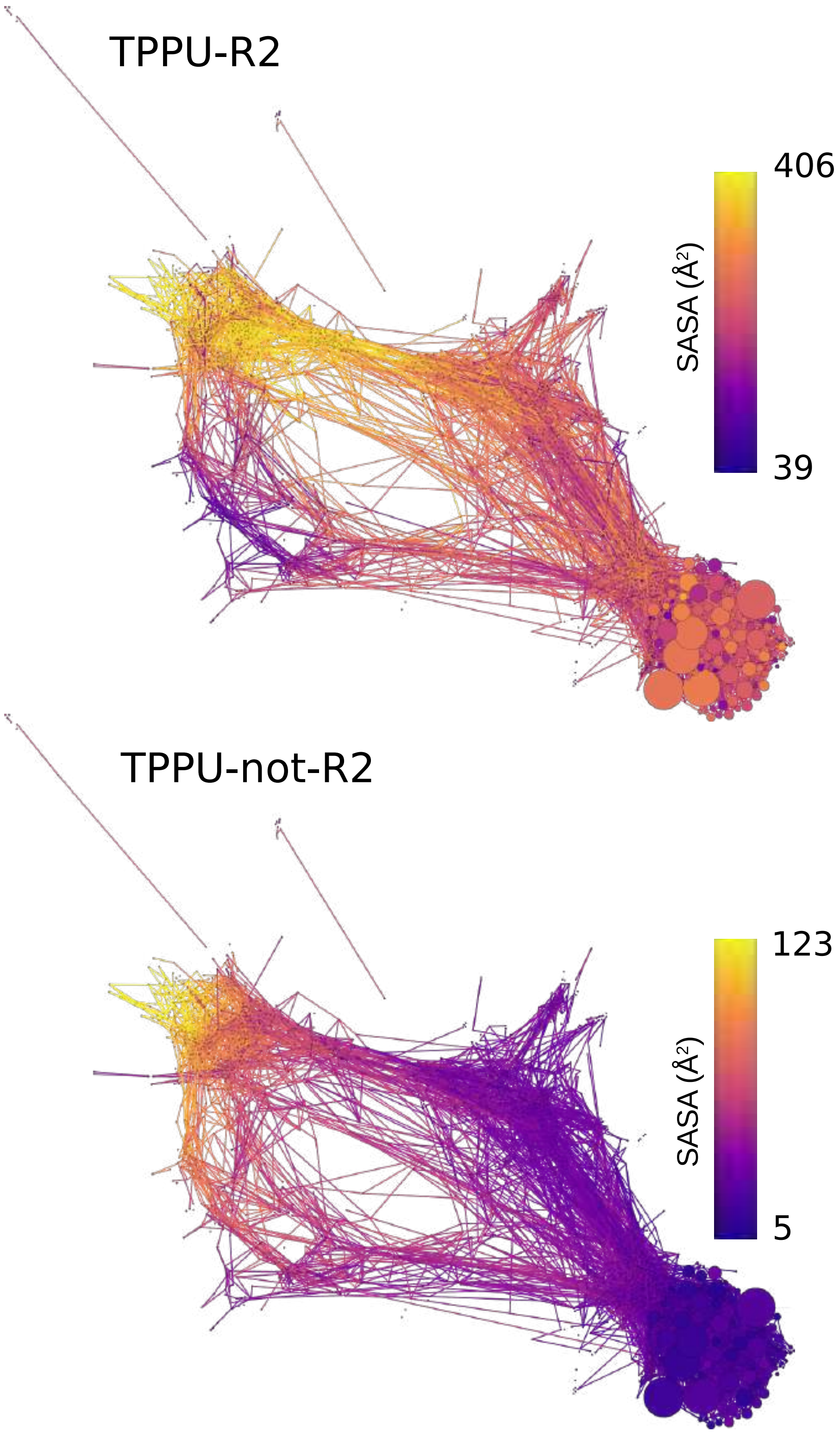


Figure 5.24: Conformation Space Network (CSN) colored according to the solvent accessible surface area (SASA) (Å^2) of only the R2 group of the TPPU ligand (top) and only the non-R2 atoms of the TPPU ligand (bottom).

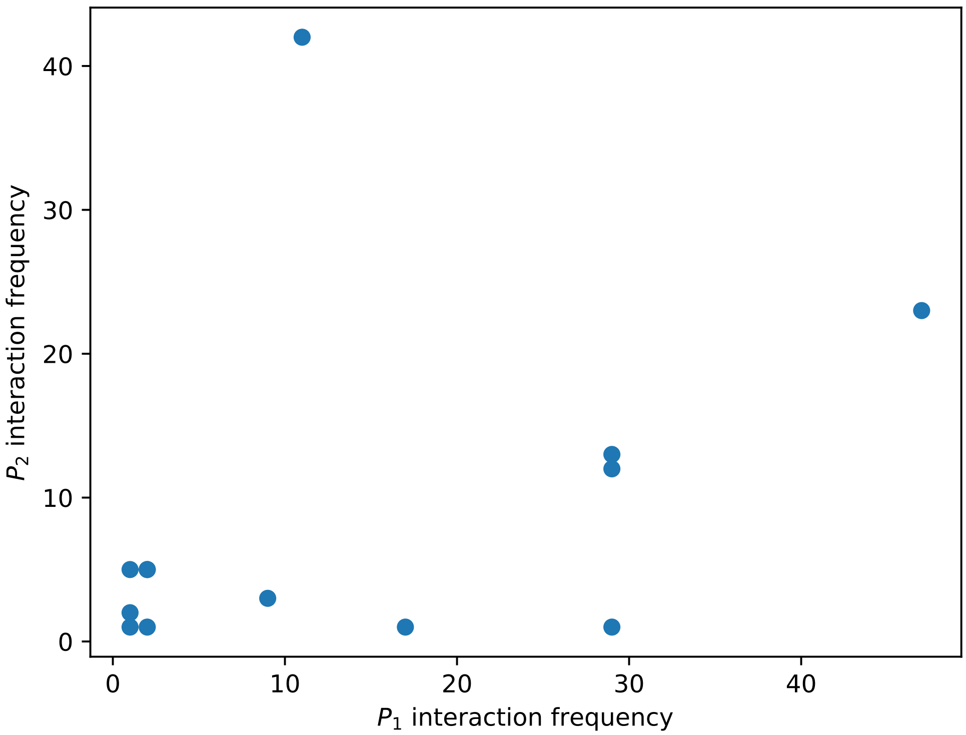


Figure 5.25: Scatter plot showing the frequency of the 13 intersecting interactions between P_1 and P_2 community ensembles. Each point represents a specific acceptor-donor hydrogen bond pair.

We consider order parameter driven methods, such as metadynamics, to fundamentally be in a refinement-oriented regime in which a particular unbinding mechanism is sampled extensively for kinetics estimations. We thus emphasize that discovery- and refinement-oriented enhanced sampling methods can be used in tandem to both elucidate the structural elements of complex unbinding mechanisms as well as their rates. Particularly, refinement-oriented methods control for the sensitivity of discovery-oriented methods like WExplore to initial conditions and subsequent divergence between different runs. The combination of both regimes to solve increasingly difficult processes will require significant collaboration in order to unify enhanced sampling implementations and share data.

5.5 Conclusion

This work describes advances in general methods applicable to the simulation of complex macromolecular processes, such as ligand unbinding, protein folding, and conformational changes. To our knowledge this is the first simulation of drug unbinding that occurs at pharmacologically relevant timescales (11 min) with no specific knowledge of the mechanism and no biasing forces, using only commodity hardware. Furthermore, our results were achieved in only 6 µs of simulation time, which compares favorably to similar simulations using either brute force or Markov state modeling that typically use one to two orders of magnitude more simulation time [44, 48].

We report the simulation of the drug-like inhibitor TPPU unbinding (and rebinding) from the soluble epoxide hydrolase (sEH) enzyme using the WExplore algorithm. We visualize the conformational ensemble as a network which we use to identify specific interactions involved in the unbinding process. Some of these interactions are implicated in the stabilization of the unbinding transition state, which potentially could be used in kinetics-oriented drug design. With these results as a guide, a given ligand-protein system can be examined using about 6 µs of sampling (or 6000 cycles). Using one node with 4 NVIDIA K80 graphics processing units we are able to obtain 120 cycles per day for this system. With a modest cluster of 5 such nodes, additional ligands can be examined at 10 days each. A larger cluster with 36 such nodes could simulate a new ligand every 33 hours. Thus, we have also proposed a general way forward for obtaining and improving the accuracy of kinetic models from MD using both discovery- and refinement-oriented enhanced sampling methods.

Supporting Information. Supplemental figures and tables can be found in the supplemental pdf. Simulation systems, data, and analysis code will be made available on zenodo.org

CHAPTER 6

LIGAND-UNBINDING TRANSITION STATE PLASTICITY IN LEAD OPTIMIZATION

6.1 Introduction

In our previous studies of drug-like inhibitors unbinding from the soluble epoxide hydrolase (sEH) target (in Chapter 5 and [102]) we were able to demonstrate multiple breakthroughs. Firstly, we demonstrated that the WExplore enhanced sampling algorithm is able to sample full unbinding trajectories for molecular processes with timescales on the order of tens of minutes. Secondly, we were able to make a prediction of the 3D atomic structure of the transition state(s). Thirdly, we identified two distinct unbinding pathways in which ligands might travel. This information was useful not only for practitioners interested in enhanced sampling methods and molecular simulations, but also to medicinal chemists that are interested optimizing various kinetic parameters of the drug-like inhibitors. As we have detailed in Section 5.2.6, sEH is a good system to study because it is both tractable to work with in simulations and is also being specifically targeted for kinetics-oriented optimization of inhibitors. By dint of having real clinical interest there is already a fairly large body of experimental data available for both kinetics and affinity [2]. Computational studies should strive to be both verifiable to some degree from experimental data and enrich it with information that is not readily available. For these reasons, in this study we aim to build on our initial findings for sEH and continue to answer questions of importance both for general knowledge of ligand (un)binding mechanisms as well as answer specific questions for a target of real clinical importance.

Parallel to the main outputs of our previous study our new objectives are as follows:

1. obtain full unbinding trajectories and rate estimates for similar inhibitors and evaluate accuracy of simulations,

2. improve predictions of transition states, and
3. investigate the plasticity of the transition state during lead optimization (i.e. chemical perturbation) in terms of both the structure and/or the dominant unbinding pathway.

The first objective aims to connect our computational predictions to experimental determinations of unbinding rates. Rate estimates from simulations are of less importance than the mechanistic detail that is uncovered from them, but provide at least a rudimentary way to assess the accuracy of the simulations. In this study we have chosen to simulate ligands that have experimental rates and affinities measured. With this we can compare the absolute estimate of accuracy (cardinality) as well as a relative estimate (ordinal) through rank ordering. It is expected that that cardinal estimates of rates will be inaccurate not only because of error in the simulations (simulations are undersampled and not converged; i.e. experimental error), but also due to systematic errors introduced by inaccuracies in the force fields and other force evaluation algorithms and parameters used in the simulations. For our scope the latter systematic error cannot be corrected for and as such will contribute to an unknown amount of error in cardinal estimates of rates. We also note that there are also errors in the experimental data itself, both in terms of experimental error and the model error which is used to calculate rates from the raw fluorescence data [192]. Thus it is much more viable to compare the rank-ordering of rate estimates from simulation and to the experimental rank-ordering. Simulation model errors do not necessarily effect each system linearly and rank ordering could still be “correct” from the sampling point of view. Importantly, while comparisons of rate estimates is a natural checkpoint for evaluating simulation results it is still far from being authoritative. This of course is an interesting avenue for investigation, but is not the primary objective of this study. See Section 1.2.7 for a detail of how rate estimates are calculated and evaluated.

The second objective will primarily be addressed by improving the MSMs used for calculating committor probabilities used in transition state prediction, see Section 6.2.4 for full details. The outcome of this objective is that we improve confidence in the models of transition states. This should in turn improve predictions of the molecular mechanisms that determine rates, both for medicinal chemists making intuitive judgments as well as more data driven assessments as in this study.

The third objective (transition state plasticity) effects our workflow at multiple levels and is the primary topic of the current study. It comes from an understanding that as different ligands are developed for a target the effects of chemical modifications (a.k.a. perturbations) are not always easily predictable. That is simply to say that the design space is complex, where small perturbations in one aspect (the side group in a ligand) might cause large perturbations of the entire system (an alternative bound pose is assumed).

This is a recognized problem in drug-design and it is common practice to find ranges of chemical perturbation that do have somewhat predictable outcomes. Such a model is typically called a structure activity relationship (SAR), however for kinetics the term is adapted to structure kinetic relationship (SKR). Models like these typically take the form of simple linear models from experimental data over some projected feature of the chemical perturbation, e.g. molecular mass of an R-group. Not unexpectedly, these models provide little predictive power outside of already sampled ranges.

A small thought-experiment illustrates the issue. A chemist might see that increasing the charge of a ligand increases the affinity having samples of ligands that are of charges -1, 0, and $+1$. The chemist then samples a ligand with a $+2$ charge and finds it to have decreased affinity. For someone with an understanding of chemistry this obviously indicates that the receptor is of charge -1 and a single positive charge on the ligand is highly favorable, but that any additional charge is either ineffective or worse.

This example shows the lack of predictive power of such simple models, and the power of a deeper understanding of the mechanisms in action. However, while electrical charge is both well understood and has a large effect on such systems, most of the causes by which we improve drug molecules are much less well understood and of lower effect size. Simple SAR

& SKR models then are most effective for small incremental optimizations on an already proven modality (i.e. leads and/or scaffolds).

To reiterate our goal, we are interested in the plasticity of ligand unbinding transition states in order to gain in understanding of the mechanisms that *invalidate* simple SKR models. We are less interested in making incremental optimizations for *predicting* kinetic properties. This is easier done simply by synthesizing and experimentally validating ligands, which would be done anyhow in “last-mile” optimizations.

Ligand unbinding simulations for events of this timescale are still computationally very expensive requiring hundred to thousands of GPU hours, with few guarantees it will continue to scale. Furthermore, the actual estimates of the kinetic properties via these simulations is rather poor for even simpler systems (see Chapter 4 and [47]). Atomistic simulations are then much better suited to generating hypotheses as to why simpler SKR models failed. The specifics of the approach taken in this study will be discussed in Section 6.2.1 as well as evaluated in the Results.

6.2 Methods

6.2.1 Predicting Pathway-Hopping from Experimental Data

Given that our simulations are relatively expensive and chemical space is large it is appropriate to first develop a strategy to obtain the most amount of information possible from each simulation. Previous studies in ligand unbinding simulations were primarily “proofs-of-concept” and a positive result for a new receptor-ligand system of slow kinetics was considered a success. Beyond the proof-of-concept stage of investigation a researcher must pay careful attention to experimental design and planning. Fortunately, for the system at hand (sEH) we have existing data from both experiment and simulation that can be used to guide us. In this section I will:

1. detail the experimental data available in terms of how it is measured and for what

ligands it is available (Section 6.2.1.1),

2. describe and prove how experimental rates and affinities for a series of ligands can be used to estimate perturbations specific to TS free energies (Section 6.2.1.2),

3. explain the hypothesis relating TS free energy perturbations to “pathway hopping” (Section 6.2.1.3), and

4. make predictions for TS plasticity for a selection of inhibitors which will be used as the experimental design for further simulation studies (Section 6.2.1.4).

6.2.1.1 Experimental Rate & Affinity Measurements and Data

As mentioned previously there are existing datasets for a wide range of inhibitors for soluble epoxide hydrolase (sEH). For this study we used the data available from Lee *et al.* (2014) [2], as many of the inhibitors described there are variations on the basic scaffold of the TPPU inhibitor. See Section 5.2.6 for more basic information on sEH and the TPPU crystal structure.

First we will introduce some nomenclature used to describe the series of ligands we will discuss. As shown in Fig. 6.1 all of the ligands under consideration are based off of the TPPU prototype and have a core piperidyl-urea scaffold with two variable R-groups: R_1 and R_2. The consistent theme for modifications to the R_2 group are that of a large hydrophobic group with a carbonyl moiety at the base. Modifications to R_1 are much more variable, but commonly are based on the 4-trifluoromethoxyphenyl group of TPPU.

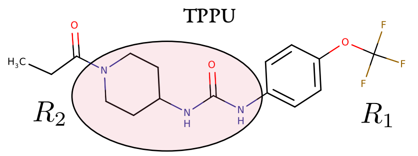


Figure 6.1: Structure of the prototypical TPPU inhibitor with the core piperidyl-urea pharmacophore highlighted. The R_1 and R_2 groups are labelled as such.

In Table 6.1 we have listed all of the inhibitors discussed in [2] with measurements for both rates (k_{off}) and affinities (K_i). Both the k_{off} and K_i are measured using a displacement assay of a fluorescent ligand described in [192]. The structures of all of these inhibitors are shown in Fig. 6.2. We note that TPPU is also known as and may be referred to as inhibitor 17.

We decided to exclude inhibitors 24, 30, and 31 which all have sulfonamide R_2 groups. This was because inhibitors 24 and 30 showed worse or similar properties as 17 (TPPU) and would add an extra complication to comparing results as ligand force field parametrization is known to be less than optimal. Figure 6.3 shows measured rates and affinities for all of the retained inhibitors along with the associated experimental errors.

In the Lee *et al.* study [2] inhibitor 17 (TPPU) was the lead molecule and using the crystallographic structure for the complex (PDB: 4OD0) the rest of the inhibitors were proposed, synthesized, and tested. This was successful as these new inhibitors all had reduced k_{off} values (i.e. longer residence times) and lower K_d values (i.e. higher affinity). Furthermore, we can see that there is a somewhat linear trend between these optimized inhibitors and the starting lead inhibitor (17). Within this linear trend there are some outliers however, such as inhibitor 10.

As explained in Section 6.1 we are less interested in why there is a linear relationship between these and more interested in what causes divergences in the outliers. In the next section we will detail how we can use relationships in this data to estimate differential TS perturbations that potentially contribute to outliers such as inhibitor 10.

Table 6.1: Table of inhibitors from Lee *et al.* [2] having both k_{off} and K_i measurements available. Ligands will be referred to by the values in the "ID" column in this study. Note that ligand 17 is the ID for TPPU. The values in the column "Lee (2014) ID" are the identifiers from the original paper [2].

ID	Lee (2014) ID	$k_{off}(s^{-1})10^{-4}$	K_i (nM) human	k_{off} error	K_i error
3	4	6.57	0.66	0.3	0.1
4	5	7.91	0.49	0.31	0.08
5	6	5.76	0.37	0.26	0.04
6	7	5.19	0.22	0.09	0.04
8	9	4.75	0.43	0.11	0.06
10	11	5.79	1.17	0.43	0.27
12	13	3.13	0.02	0.06	0.01
13	14	5.39	0.36	0.39	0.09
16	17	3.51	0.23	0.2	0.07
17	18	10.5	0.91	0.2	0.13
18	19	6.14	0.31	0.18	0.18
20	21	5.05	0.19	0.02	0.04
23	24	4.39	0.02	0.43	
24	25	23.1	7.16	1.1	0.42
30	31	10.3	0.98	0.1	0.19
31	32	8.9	0.44	0.35	0.11

6.2.1.2 Estimating Specific Transition State Free-Energy Perturbations from Rates & Affinity

The current study differs from many traditional drug design endeavors because we are not solely focused on improving affinities of inhibitors (i.e K_d). Certainly the affinity of an inhibitor has a major impact on the overall potency and efficacy of a drug. However, there is significant interest in whether the optimization of the (un)binding kinetics of inhibitors can improve orthogonal efficacious factors (see Section 1.1 for a detailed introduction).

This change in perspective brings a number of novel challenges to overcome. First was simply the challenge of observing the transition state structures which (partially) govern the kinetics (this challenge was addressed in Chapters 3, 4, & 5). Secondly, even given experimental kinetics and affinity data, as shown in Fig. 6.3, we are left to wonder as to what exactly is changing from a thermodynamical perspective.

This is because the thermodynamic terms that determine these macroscopic observables are not independent. The affinity is a function of the bound state free energy (G_B) and the unbound state free energy (G_U) through Eq. 6.11 and the following definition:

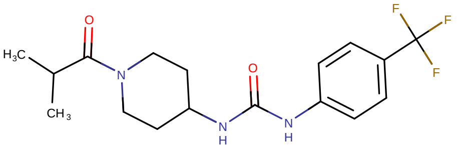
$$\Delta G = G_U - G_B \tag{6.1}$$

Similarly the off rate is a function of the transition state free energy (G_T) and again the bound state free energy (G_B) through 6.10 and the following definition:

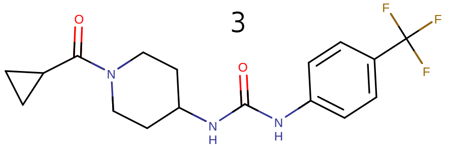
$$\Delta G^{\ddagger} = G_T - G_B \tag{6.2}$$

As we can see the G_B is common to both. This is easy to understand if we consider the change in the off rate (k_{off}) with respect to G_B. Assuming G_U and G_T are both constant, a lowering of G_B would result in both an increased K_d and k_{off}.

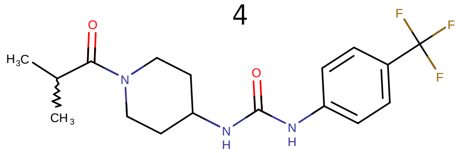
Looking again at the data in Fig. 6.3 we can't immediately say whether changes in k_{off} are due to changes in G_B or G_T. For a drug designer this is very important as they will likely make modifications to lead molecules differently for bound states and transition



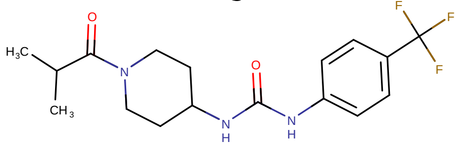
3



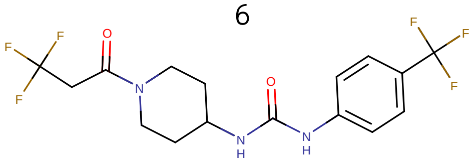
4



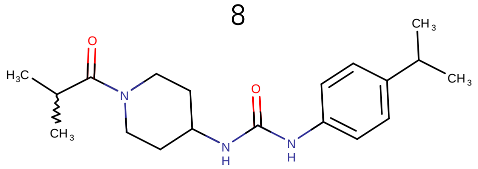
5



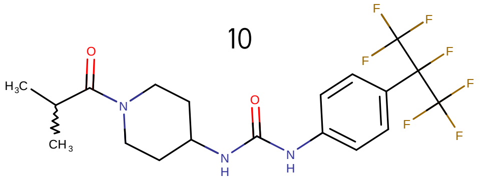
6



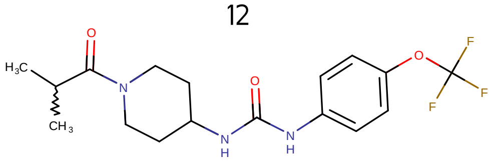
8



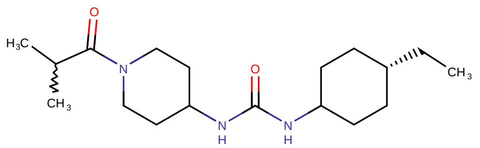
10



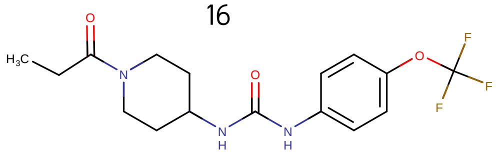
12



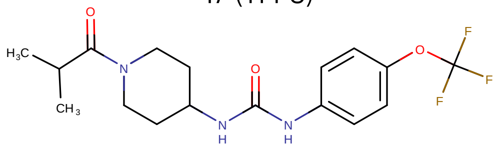
13



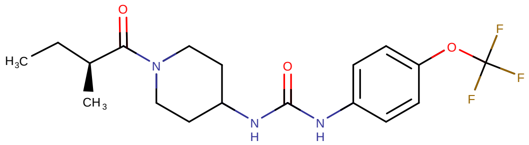
16



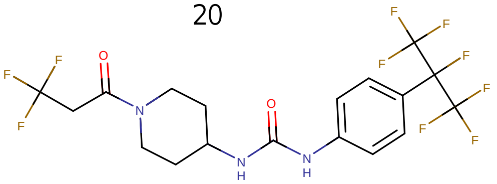
17 (TPPU)



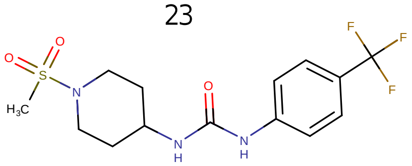
18



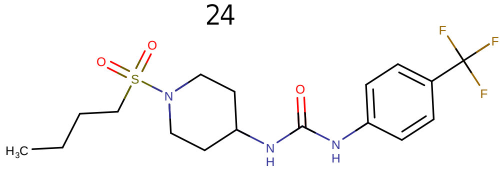
20



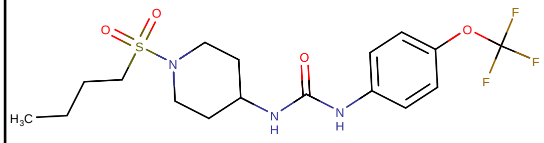
23



24



30



31

Figure 6.2: Chemical structures of all of the inhibitors from [2] that had both unbinding rates and affinities experimentally measured as in Table 6.1. The R_2 groups are oriented to the left and the R_1 groups are oriented to the right. Bonds shown as a zig-zag were measured as a racemic mixture of stereo-isomers.

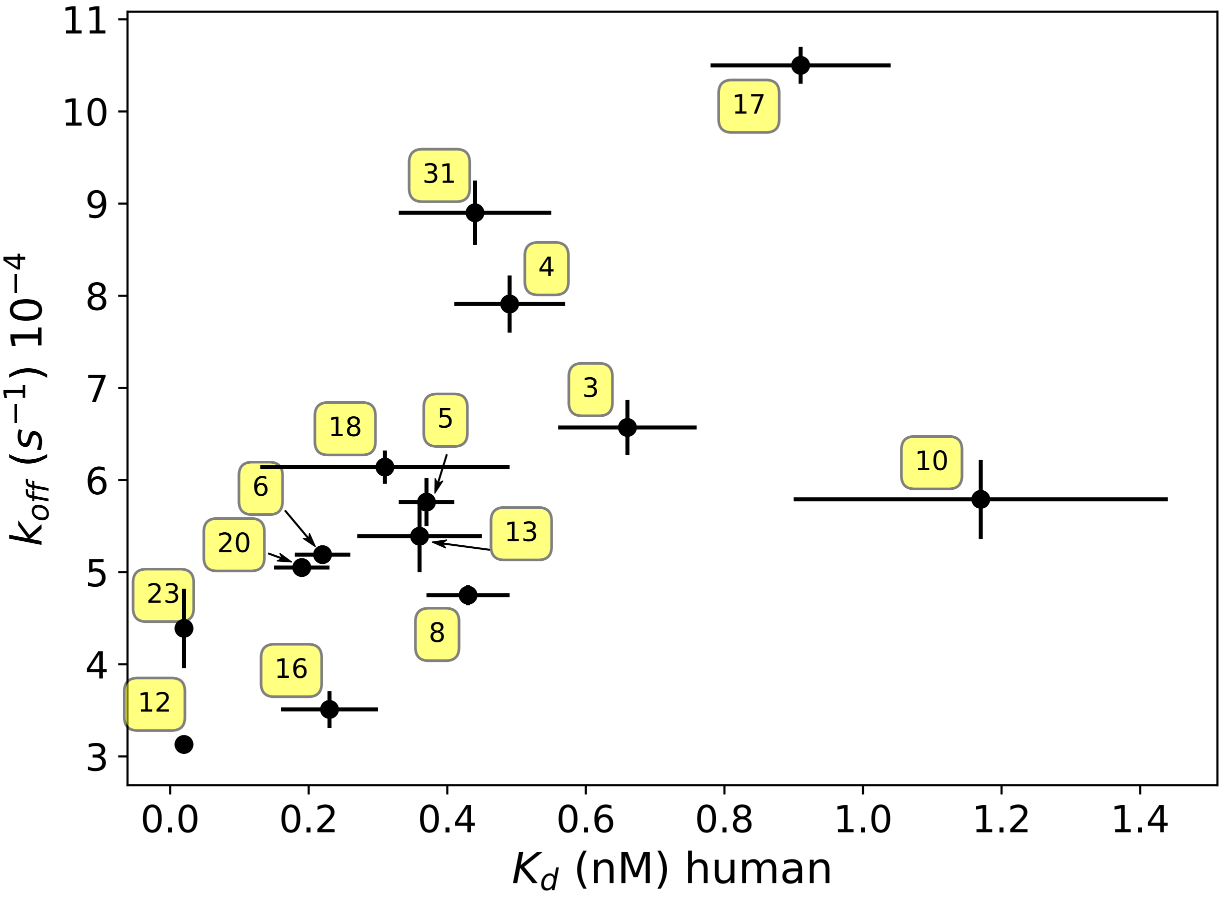


Figure 6.3: A plot of the experimental k_{off} over the K_i values for each ligand except for 24, 30, and 31. Both values are from human sEH. Experimental error is shown for each measurement as the lines intersecting each point.

states. While it is possible that the same molecular feature has utility with respect to both, we assume that this is not general enough to be relied on.

As this conflation of variables is endemic to any kind of macroscopic measurements we need a way to separate or at least estimate the degree of independence. To state the problem explicitly we introduce a chemical perturbation relationship between two inhibitor molecules: X and Y. The perturbation itself is denoted Δ_{XY} and is a function that is simply the difference between variables of Y and X (see Eq. 6.14).

From this we naturally have the following terms for the absolute free energies:

$$\Delta_{XY}(G_B) = G_B^Y - G_B^X \tag{6.3}$$

$$\Delta_{XY}(G_U) = G_U^Y - G_U^X \tag{6.4}$$

$$\Delta_{XY}(G_T) = G_T^Y - G_T^X \tag{6.5}$$

$$\tag{6.6}$$

As we are interested in modulating kinetics we are interested in the perturbations to the transition state ($\Delta_{XY}(G_T)$) that affect the k_{off} independently of perturbations to the bound state ($\Delta_{XY}(G_B)$). Given K_d data we should be able to account for contributions to k_{off} via changes in G_B.

Theorem 1 provides a measure of the relative contributions of the two chemical perturbation free energy changes which we call the perturbation ratio and is denoted as S. Furthermore, the theorem provides a means of estimating this quantity through the measured k_{off} and K_d values. The proof of this (Proof 6.2.1.2) only assumes that the free energy of the unbound state does not change, i.e. $\Delta_{XY}(G_U) = 0$.

Theorem 1. *The perturbation ratio S is estimated by the following expression:*

$$S \propto 1 - Q_{XY} \tag{6.7}$$

where,

$$S = \frac{\Delta_{XY}(G_T)}{\Delta_{XY}(G_B)} \tag{6.8}$$

when $\Delta G_U = 0$; and where

$$Q_{XY} = \frac{\Delta_{XY}(\ln k_{\mathit{off}})}{\Delta_{XY}(\ln K_d)} \tag{6.9}$$

The application of the theorem will be seen in the following sections. Following the proof we describe the direct interpretation S values to series of chemical perturbations and potential improvements and extensions to this model. In Section 6.2.1.3 we will use this interpretation to make a hypothesis regarding the plasticity of transition states to this value. Following that in Section 6.2.1.4 S values will be calculated for the sEH inhibitors introduced in 6.2.1.1 and hypotheses of transition state plasticity will be made.

Proof. Proof of Theorem 1

We start with the knowledge that the relative free energies ($\Delta G^{\ddagger}$ and ΔG) are estimated by the k_{off} and K_d, respectively:

$$\Delta G^{\ddagger} \propto -RT \ln k_{\text{off}} \tag{6.10}$$

$$\Delta G \propto -RT \ln K_d \tag{6.11}$$

Then for a chemical perturbation Δ_{XY} we have similar equations:

$$\Delta_{XY}(\Delta G^{\ddagger}) \propto \Delta_{XY}(-RT \ln k_{\text{off}}) \tag{6.12}$$

$$\Delta_{XY}(\Delta G) \propto \Delta_{XY}(-RT \ln K_d) \tag{6.13}$$

Where,

$$\Delta_{XY}(m) = m^{Y} - m^{X} \tag{6.14}$$

and m_i is a measured quantity m for molecule i.

We start by taking a ratio of the activation energy and the binding free energy:

$$\frac{\Delta_{XY}(\Delta G^{\ddagger})}{\Delta_{XY}(\Delta G)} \propto \frac{\Delta_{XY}(-RT \ln k_{\text{off}})}{\Delta_{XY}(-RT \ln K_d)}$$

We can immediately simplify the right-hand side to get Q_{XY}:

$$\frac{\Delta_{XY}(-RT\ln k_{\text{off}})}{\Delta_{XY}(-RT\ln K_d)} = \frac{\Delta_{XY}(\ln k_{\text{off}})}{\Delta_{XY}(\ln K_d)} = Q_{XY}$$

And for the left-hand side we substitute in the full expressions for each in terms of the energies for each molecule according to Eq. 6.14:

$$\frac{\Delta_{XY}(\Delta G^{\ddagger})}{\Delta_{XY}(\Delta G)} = \frac{\Delta G^{\ddagger}{}_Y - \Delta G^{\ddagger}{}_X}{\Delta G_Y - \Delta G_X}$$

The relative free energies themselves can be expressed in terms of the absolute free energies of the relevant states,

$$\Delta G^{\ddagger} = G_T - G_B \tag{6.15}$$

$$\Delta G = G_B - G_U \tag{6.16}$$

We substitute these in for the values of each molecule:

$$\frac{\Delta_{XY}(\Delta G^{\ddagger})}{\Delta_{XY}(\Delta G)} = \frac{(G_{T,Y} - G_{B,Y}) - (G_{T,X} - G_{B,X})}{(G_{U,Y} - G_{B,Y}) - (G_{U,X} - G_{B,X})}$$

Rearranging the right hand side we come to:

$$\frac{\Delta_{XY}(\Delta G^{\ddagger})}{\Delta_{XY}(\Delta G)} = \frac{(G_{T,Y} - G_{T,X}) - (G_{B,Y} - G_{B,X})}{(G_{U,Y} - G_{U,X}) - (G_{B,Y} - G_{B,X})}$$

We notice that the expressions inside the parentheses are just those of the perturbation free energies $\Delta_{XY}(G_T)$, $\Delta_{XY}(G_B)$, and $\Delta_{XY}(G_U)$ as given in Equation 6.14.

Making this substitution we come to:

$$\frac{\Delta_{XY}(\Delta G^{\ddagger})}{\Delta_{XY}(\Delta G)} = \frac{\Delta_{XY}(G_T) - \Delta_{XY}(G_B)}{\Delta_{XY}(G_U) - \Delta_{XY}(G_B)} \tag{6.17}$$

If we then make use of our assumption $\Delta G_U = 0$:

$$\frac{\Delta_{XY}(\Delta G^{\ddagger})}{\Delta_{XY}(\Delta G)} = \frac{\Delta_{XY}(G_T) - \Delta_{XY}(G_B)}{0 - \Delta_{XY}(G_B)} = 1 - \frac{\Delta_{XY}(G_T)}{\Delta_{XY}(G_B)} \tag{6.18}$$

We notice that the ratio on the right-hand side is just the value S.

$$\frac{\Delta_{XY}(\Delta G^{\ddagger})}{\Delta_{XY}(\Delta G)} = 1 - S \propto Q_{XY} \tag{6.19}$$

Rearranging the proportionality relation we come to what we aimed to:

$$S \propto 1 - Q_{XY} \tag{6.20}$$

QED

This proof is useful because the Q_{XY} can be calculated between any pair of molecules with K_d and k_{off} estimates available, as can be done experimentally.

Explicitly,

$$Q_{XY} = \frac{\ln k_{\text{off},X} - \ln k_{\text{off},Y}}{\ln K_{d,X} - \ln K_{d,Y}} \tag{6.21}$$

To interpret individual S values consider the features of the function plotted as $S(Q)$ over the entire domain, including the root, limits, and when the function crosses the S axis. This is summarized in Table 6.2, different ranges and values for S & Q have different interpretations for the ratio of $\Delta_{XY}(G_T)$ to $\Delta_{XY}(G_B)$. Plots of k_{off} vs K_d can be interpreted directly using this table for the given slopes Q of any relationship.

As might be expected vertical lines ($Q = \mp\infty$) indicate that only perturbations to G_T are occurring, and indicates perfect independence of the transition state to the bound state. The opposite however is not true for horizontal lines ($Q = 0$), where $S = 1$. This indicates that that G_B and G_T change exactly in lockstep which results in no change in k_{off}, but independent changes in K_d. Perfectly independent perturbations to G_B occur when $S = 0$ and $Q = 1$. This makes sense in that for every change in the G_B there is the same response in both the $\Delta G^{\ddagger}$ and ΔG.

The other cases cover the ranges between these extremes where changes are occurring (in both directions) in both the G_B and G_T. High values of $Q > 1$ (corresponding to high $|S|$

Table 6.2: Table showing important ranges and values for both the perturbation ratio S and the associated slope Q on a k_{off} vs. K_d plot. Using a Q slope for a desired chemical perturbation the dominance or independence of perturbation free energy of the bound and transition state can be identified. The final columns shows the directions of the perturbations for the $\Delta_{XY}(G_B)$ and $\Delta_{XY}(G_T)$ respectively.

S values	Q slope	$\Delta_{XY}(G_B)$ vs. $\Delta_{XY}(G_T)$ relation	Direction of Perturbation
$S > 1$	$Q < 0$	$S\Delta_{XY}(G_B) = \Delta_{XY}(G_T)$	↑↑
$S = 1$	$Q = 0$	$\Delta_{XY}(G_B) = \Delta_{XY}(G_T)$	↑↑
$0 < S < 1$	$0 < Q < 1$	$\frac{1}{S}\Delta_{XY}(G_B) = \Delta_{XY}(G_T)$	↑↑
$S = 0$	$Q = 1$	$0 = \Delta_{XY}(G_T)$	− ↑
$S < 0$	$Q > 1$	$-S\Delta_{XY}(G_B) = \Delta_{XY}(G_T)$	↓↑
$S = \pm\infty$	$Q = \mp\infty$	$\Delta_{XY}(G_B) = 0$	↑ −

where $S < 0$) indicate that large perturbations to G_T correspond to large opposite responses in G_B. This is perhaps the most interesting regime to be in for kinetics oriented drug design as both the the k_{off} and K_d will be decreased. Compared to values closer to $S = 0$ we know that the relative contribution of the $\Delta_{XY}(G_T)$ is greater as well. This makes sense when you consider it graphically, the k_{off} is decreasing faster than the K_d. High $|Q|$ perturbation relationships are of primary interest to us in this study because we are interested in inhibitors which are specifically destabilizing to the transition state.

Looking again at Fig. 6.3 we can't directly interpret specific slope values as the units needed for the S equation are log-scaled. The data is re-plotted in Fig. 6.4 with the appropriate units (and eliding the error bars). Here we can see that if we compare all the optimized inhibitors (i.e. Y) to the starting lead inhibitor 17 (i.e. X) they all have $Q > 0$. We could calculate slopes for each $17 \rightarrow Y$ pair or parametrize a linear regression of a series to obtain specific slopes. However, the utility of this is dubious as relative slopes between Y inhibitors are obvious by inspection. From this we can immediately ascertain that for example inhibitors 3 and 10 have high $|Q|$ values relative to the others. At first 10 looks as if there is even a slight negative Q however, given the large error in the K_d measurement this is not clear. If it does indeed have a $Q < 0$ (and $S > 1$) and extrapolation of this holds for a series then we can expect that the K_d would actually increase, which would be undesirable for real lead optimization. However, because $|Q|$ is large for inhibitor 10 it would still make an interesting test subject for analyzing decoupled transition state perturbations. Alternatively, a series that follows the trend of high $|Q|$ where $Q > 0$, such as inhibitor 3 would result in decoupled $\Delta_{XY}(G_T)$ as well as favorable affinity improvements. In the case of inhibitor 3 this seems primarily due to the absence of the methoxy ester in the R_1, whereas for inhibitor 10 the replacement of the R_1 4-trifluoromethoxyphenyl with an aliphatic group.

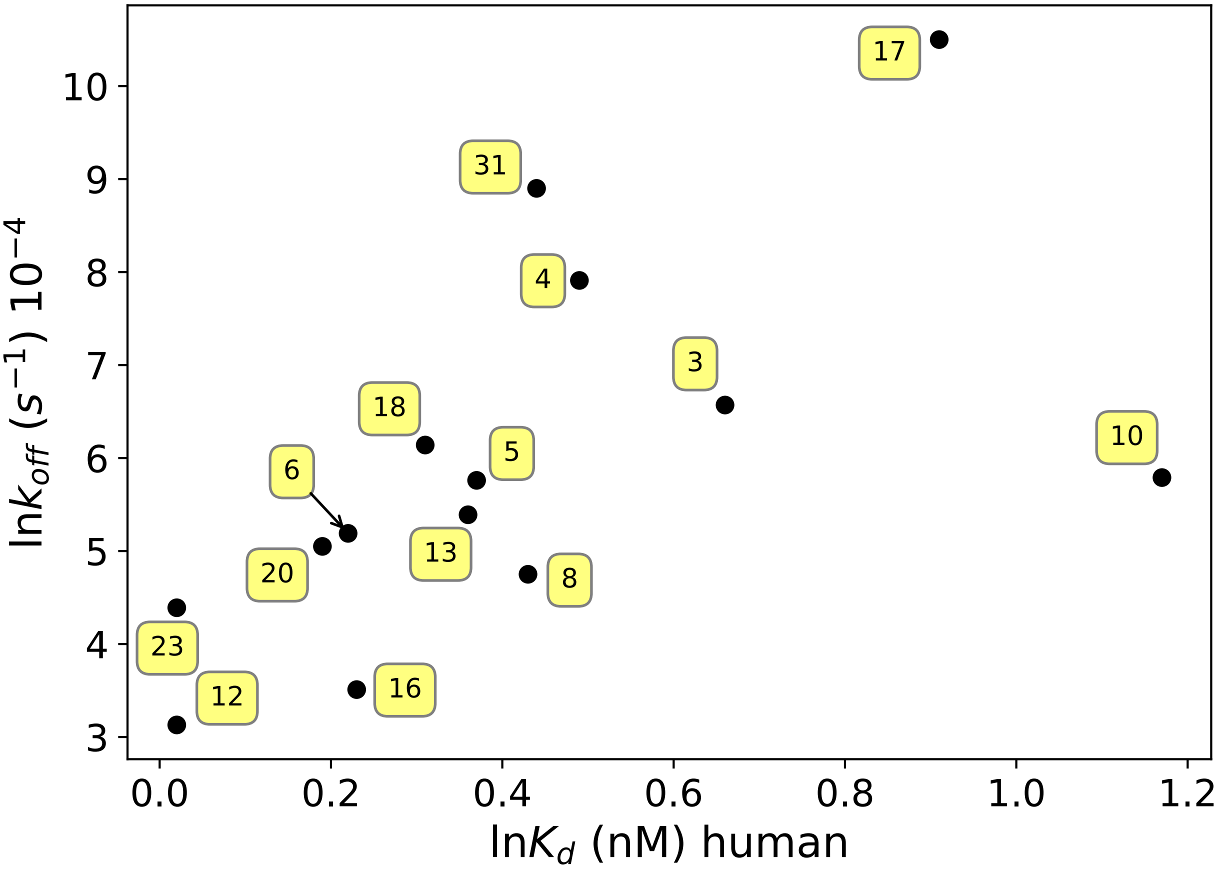


Figure 6.4: A re-plot of the experimental k_{off} over the K_i values from Fig. 6.3 transformed by the natural logarithm (ln).

We also note the possibility of further generalizations of this pairwise relationship between molecules to spaces of chemical perturbation, λ. That is, the integral of $\ln k_{\text{off}}$ with respect to $\ln K_d$:

$$
\begin{gathered}
\Delta_{XY}(\ln k_{\text{off}}) = Q_{XY}\Delta_{XY}(\ln K_d) \\
\int^{\lambda} d(\ln k_{\text{off}}) = \int^{\lambda} Q\, d(\ln K_d) \\
\ln k_{\text{off}} = Q(\ln K_d) + B
\end{gathered}
$$

Where now Q is the slope of a series of molecular perturbations (since samples of λ will be discrete molecules). The main practical implication of this is that the S can be calculated for a particular series of ligands designed during lead optimization after doing some sort of linear fit (i.e. linear regression) and finding the slope.

We note, but do not elaborate here, that the B value in the linear equation is related to the G_U (i.e. solubility) of the ligands and perhaps correction factors could be applied here given solubility data of the ligands.

6.2.1.3 Hypothesis: Pathway-Hopping is Predicted by Specific Free-Energy Perturbations to Unbinding Transition States

In the introduction of this chapter we stated that the present study is interested in investigating transition state plasticity during lead optimization so that we can explain irregularities in structure kinetic relationship (SKR) models used by drug design researchers. Towards this goal we propose the following hypothesis: *SKR irregularities are causes by highly plastic transition states caused by TS free energies that are highly sensitive to chemical perturbations.* Said another way, we predict that for a series of ligands in which the $\Delta_{XY}(G_T)$ is relatively large then it is more probable that the actual structure of the TS will be plastic.

To understand why this is a reasonable hypothesis, consider a hypothetical situation for 3 inhibitors: the starting lead molecule X and two optimized versions of it A & B. For all three we have experimental data of k_{off} and K_d. As shown in the previous section (Sec. 6.2.1.2) we can estimate the $\Delta_{XY}(G_T)$ contribution for each of these pairs (where Y is either A or B) by calculating the S ratio. We find that $S(\Delta_{X,A}) = -1$ $(Q = 2)$ and $S(\Delta_{X,B}) = -10$

($Q = 11$). This indicates that for every N units of *destabilization* to the G_T there is only $1/N$ units of *stabilization* to G_B. The effect is multiplied by 10 for the $\Delta_{X,B}$ perturbation compared to $\Delta_{X,A}$. If the chemical perturbation (i.e. the change in inhibitor structure) for A & B are relatively similar then we can infer that the G_T is much more sensitive to modifications like in B compared to A. The possible reasons for this sensitivity is manifold. It could be caused by a dramatic decrease in solubility such that the free energy difference is entropically driven. Or perhaps it could be that a specific interaction that stabilized the TS was lost in B but retained in A.

Here we make clear the distinction between **sensitivity** and **plasticity**. Sensitivity merely describes the magnitude of change in G_T begotten by any chemical perturbation, whereas plasticity indicates a mechanistic change in the identity of the TS. Pretend we successfully observe the actual unbinding transition states of all 3 inhibitors. We find that X & A both are rate limited by solvation (called TS_{solv}), but that B is limited not by solvation but by a conformational change in the protein binding site that allows for its release (called TS_{conf}). It could be said then that the $\Delta_{X,B}$ perturbation is both more sensitive (i.e. $S(\Delta_{X,A}) < S(\Delta_{X,B})$) as well as more plastic, since the rate limiting mechanism of unbinding is wholly different.

While it is possible that for $\Delta_{X,B}$ it also has TS_{solv} we predict that higher sensitivities are correlated with higher plasticity. This makes sense as there will be a behavior similar to a switch in an electronic circuit at some point. You cannot arbitrarily target a single transition state and destabilize this by an infinite amount. Using the same example, if we know that inhibitor X has TS_{solv} and in inhibitor B we substitute a group with a massive hydrophobic one this will impose steric limitations which may cause the TS_{conf} barrier to become the limiting factor. We call this putative behavior **pathway hopping** and hypothesize that this accounts for many of the irregularities in SKR models.

The issue however is more complex than the “circuit” analogy above. We are describing microscopic systems that *in situ* are well described as statistical ensembles, and as such there is no one single unbinding “path” that all inhibitors must take. Even in our earlier studies we observed multiple distinct pathways by which ligands might unbind [47, 102], or even less well defined exit distributions for nearly solvent sized fragments [39]. So in reality we are talking about an ensemble of transitions states (the transition state ensemble (TSE)) with varying degrees of coherence. Even so, we find it likely that for inhibitors of adequate size and binding sites with sufficient specificity there will be very few dominant “pathways” for unbinding. Given our preliminary findings for sEH-TPPU pathways this seems to be a reasonable assumption when looking at similar inhibitors.

To summarize, we predict that chemical perturbations (Δ_{XY}) that have high-sensitivity in the transition state free energy (G_T) are more probable to exhibit TS plasticity and be liable to **pathway hopping** behavior, thus reducing the utility of SKR models. In the next section (6.2.1.4) we use this hypothesis to make predictions for sEH inhibitors described previously (Sec. 6.2.1.1).

6.2.1.4 Experimental Design: Identifying Potential Pathway-Hopping Inhibitors

Thus far we have introduced the nature of and availability of experimental data, described a method for estimating transition state free energy perturbations from this data, and made a hypothesis that this predicts pathway hopping behavior. Now we apply our method and hypothesis to the experimental data for a series of inhibitors with the intention of testing this hypothesis through computational means.

Figure 6.5 is an annotated version of Fig. 6.4 which summarizes our hypothesis. As already mentioned above, the obvious outliers of the major trend (represented as the left-most trend line) are inhibitors 3 and 10. These inhibitors have relatively vertical lines (high Q_{XY}) and thus similarly high S values indicating decoupled perturbations to G_B and G_T.

Following our hypothesis that sensitivity in specific perturbations to the G_T predicts pathway hopping we predict that inhibitors 3 and 10 will have significantly different transition state structures compared to inhibitor 17 (TPPU). By extension we also predict that

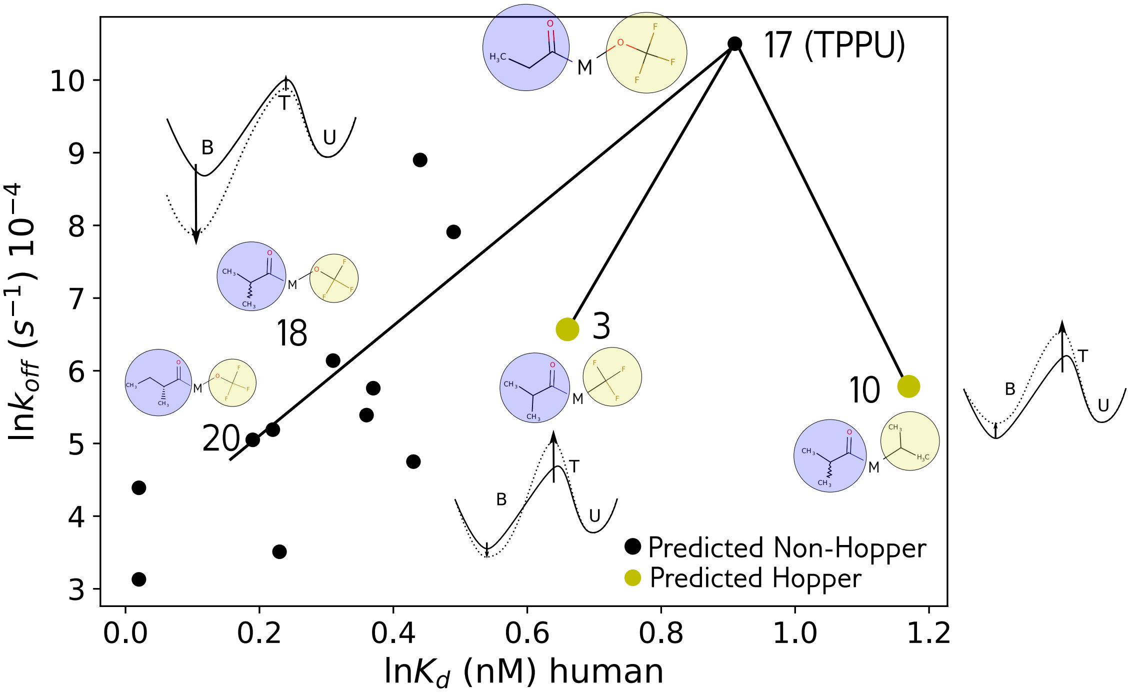


Figure 6.5: An annotated plot of the same data as in Figure 6.4. The five ligands which were chosen for simulation are labelled with their numerical identifier (as in Table 6.1 and Figure 6.2), whereas all others have been elided. Additionally, the relevant side groups of these ligands are represented in a reduced form. The 'M' symbol represents the common scaffold structure; the purple (left-side) highlights the R_2 side group and the yellow (right-side) highlights the R_1 group as defined in Fig. 6.1. The yellow colored points indicate inhibitors which we predict to be pathway hoppers, whereas black dots indicate an opposite hypothesis. The black lines emphasize the relationships that characterize the S. The cartoon diagrams associated with these trend lines provide a rough interpretation of these lines with respect to the perturbation free energies of the bound and transition states.

inhibitors in the cluster around inhibitors 18 and 20 will have similar transition state structures to TPPU.

This choice of inhibitors sets up a small experiment where we can test whether the degree to which differences in S values impacts transition state structure plasticity. There are a few possible outcomes we should consider:

- Predicted non-hoppers are actually pathway hoppers, as well as the predicted hoppers. This indicates that transition states for this system are extremely plastic. This is bad for drug design as there is no reference transition state to design against.

- Some predicted hoppers are found to not be. This is quite likely and simply indicates that transition state structures are fairly stable in these series. This is actually favorable for drug design.

- Results are the exact opposite of our predictions. This would probably be due to the fact that it is impossible to carry completely matched experiments at the quanta of molecules and side group substitutions. In short we can't match values of K_i and k_{off} in our design. If this is the case some more sophisticated corrections would be needed. A second reason could be that our assumption of similar G_U values has been invalidated. This would also be informative as we have data on the solubilities that could be used to begin to calibrate the effectiveness of the method.

It is worth noting that the issues raised in the last point could also be confounding factors in any outcome. In any case valuable information would have been obtained for both kinetics oriented drug design for sEH and at large.

In the following sections we provide specific protocol methods for calculations including how transition state models will be obtained (Section 6.2.4).

6.2.2 Simulation Details

The MD system preparation process is overall similar to the one given in Section 5.2.1, except here we first docked the query ligands before system preparation as follows. Ligand molecules were prepared by first drawing the 2D structures with the MarvinSketch software [193], then embedding these to 3D coordinates and optimizing them with the UFF force field in [178]. Charges were assigned to both the ligand and protein using the OpenBabel software [194]. The PDB 4OD0 was used as a docking template after removing the TPPU molecule from the crystal structure. Docking of ligands to 4OD0 was done with the `smina` program [195] using the "autobox" option with the original crystal structure pose of TPPU as the template. The top ten docking results for each ligand were manually assessed and the highest scoring model that had a similar orientation to the 4Od0 crystal structure pose of TPPU was chosen for simulation.

Following this the systems were constructed following the methods in Section 5.2.1. To summarize the sEH protein domain unrelated to the binding site was truncated, the system was solvated with a 12 Å square box with TIP3P water molecules, and parametrized for the CHARMM36 MD force-field [176] along with CGENFF for the small molecule ligands [177]. The systems were then heated up and equilibrated to 300 K and 1 atm following the protocol given in 5.2.1.

WE simulations used identical parameters as those given in Section 5.2.2 except for three things. First, the Particle Mesh Ewald non-bonded interactions were set with a simple cutoff of 10 Å. Second, due to observed instabilities in some of the simulations the MD timestep was reduced from 2 fs to 2 fs for ligands 10, 18, and 20. Third, the minimum allowed probability for ligands 10, 18, and 20 was lowered to 10^{-16}.

The distance metric used was the same and is the RMSD of the ligand molecules after aligning the protein binding sites to the starting structure protein binding site. The non-equilibrium boundary conditions were also the same as in Section 5.2.2, where trajectories were restarted in the starting states after the shortest protein-ligand distance is greater than 10 Å.

6.2.3 Clustering & Network Visualization

The clustering and network visualization is similar to the methods for the previous sEH ligand unbinding study (Section 5.2.3).

Simulation states were featurized as a vector of individual distances between specific protein-ligand atoms. The feature vectors are intended to be as similar as possible across each different ligand and so the atoms chosen for each ligand were chosen such that they are similar to every other ligand. The specific choices of these "homologous" atoms are shown in Figure 6.6. Atoms on the receptor were chosen as the $C - \alpha$ atoms of a selection of amino acid residues from around the binding pocket as shown in Figure 5.1. These residues are (as labelled in Fig 5.1): Tyr236, Asp105, Tyr153, Met189, and Ala134 (not shown and lies on the left side of the pocket opposite Met189). The feature vector was created by taking the distances for all pairs of the Cartesian product between the set of ligand atoms and binding site residue atoms. This results in a feature vector of length 30 for each ligand.

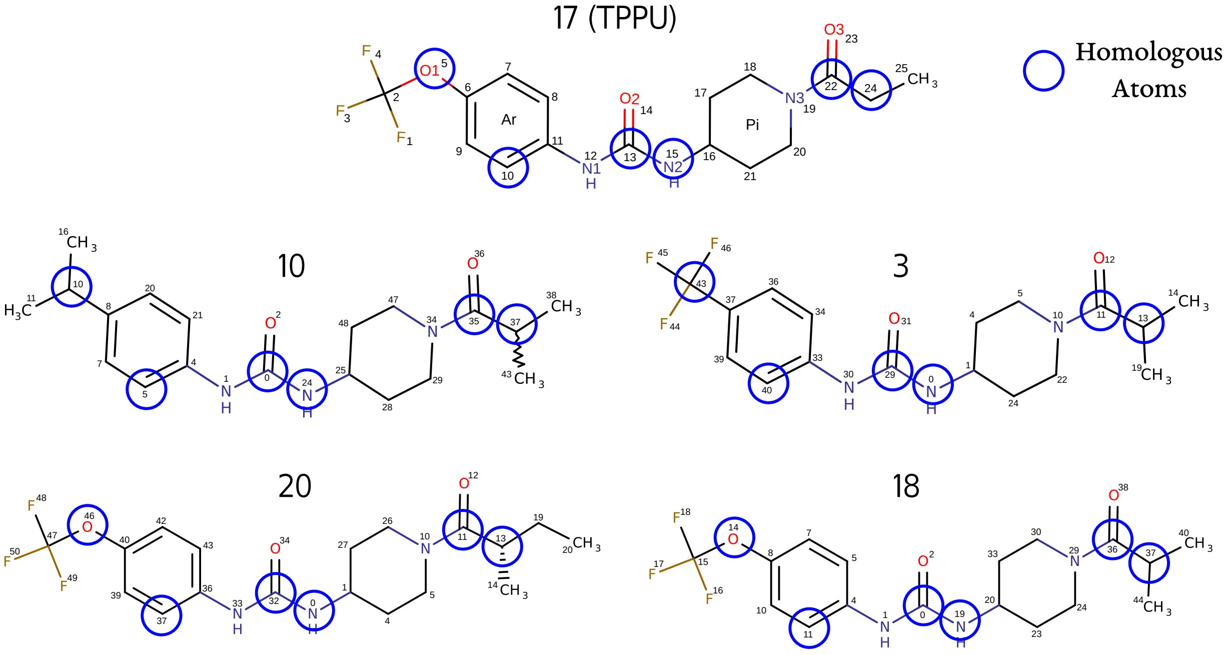


Figure 6.6: The specific homologous ligand atoms that were chosen for the clustering feature vector generation. Each ligand is labelled and the chosen atoms are circled in blue.

Clustering was done using the `KCenters` algorithm from MSMBuilder [157] using 1000 clusters. The Canberra distance metric which normalizes the magnitude between large distances and differences between small distances. This ensures that the distances between states with solvated ligands do not dominate the clusters, as these are essentially degenerate, and emphasizes small differences in states which are closer to the bound state.

The details of network visualizations are covered in Section 5.2.3. Other analyses of networks were performed in Python [196] with the aid of the analysis tools provided in `wepy` [108].

6.2.4 Committor Probabilities & Transition State Prediction

The protocol here is somewhat similar to that given in Section 5.2.5 with a few key differences. To compute the committor probabilities we first need to construct a proper MSM. First, "ergodic trimming" of the whole CSN is performed. This was done by considering the absolute number of microstate trajectory transitions between nodes without regarding the weight of the walkers (the "transition counts"). From these transition counts we set a cutoff for the number of transitions (both to and from) for the nodes to be considered "strongly connected". All nodes which are not in the strongly connected subgraph of the entire CSN were excluded from the MSM. This was accomplished using the `strongly connected subgraph` algorithm in MSMBuilder [157] from the network data-structure `MacroStateNetwork` in `wepy` [108]. Secondly, we calculated the transition probability matrix by first normalizing the outgoing transition counts, weighted by the weight of the walker at the time of transition, to a probability distribution summing to unity. This transition probability matrix is not reversible, which is appropriate for the non-equilibrium steady-state simulations which were performed.

After the MSM is constructed we then calculate the committor probabilities. To do this we must choose a set of "source" (B) and "sink" (U) "basin" nodes to compute committor probabilities ($p_{B \to U}$) between. We then used the `committors` algorithm in MSMBuilder [157] to calculate the forward committors ($p_{B \to U}$). See the results section (Section 6.3.4) for the specific criteria for basins. To estimate the TSE we then choose all clusters which have committor probabilities $0.4 \leq p_{B \to U} \leq 0.6$ as the TSE.

6.3 Results

6.3.1 Unbinding Events & Rate Estimates

We first report the overall results of the simulations in terms of unbinding events observed and total extent of sampling. For all simulations we consider an aggregate sampling time of 2 μs (that is the total of sampling across all walkers). Of the four ligands simulated only ligands 3 and 10 had full unbinding trajectories observed (i.e. max-min distance of ligand to sEH protein greater than 10 Å). All 6 of the ligand 3 replicate simulations had unbinding events while only 1 of the ligand 10 simulations observed any unbinding events.

Figure 6.7 shows the details of these unbinding events. In Panel B the aggregate probability over the course the simulation is shown for each replicate as well as the average and standard error (if calculated). The aggregate probability over time is used to make the rate predictions (via the Hill relation) which are shown in Panel A as the residence time $1/k_{\text{off}}$. Additionally, Panel A shows the experimentally determined value for both ligands compared to the simulation estimate.

From this we see that we have not improved our ability to estimate rates from our studies with TPPU (see Chapter 5). Of these the single simulation in which unbinding was seen for ligand 10 we cannot expect this to be reliable for lack of independent samples from other replicates. That being said the estimate is not much worse than that of either ligand 3 or previous TPPU estimates and differs only by a single order of magnitude. The variance in the estimates from ligand 3 reinforces our suspicion that good estimates can be obtained from single replicates as estimates from different simulations differ by around 6 orders of magnitude. The final average estimate is off by more than 2 orders of magnitude, however the experimental value still lies within the sample range. We can also see that most of the error was contributed by a single replicate in Run 1 from a single unbinding event of a highly weighted walker, which can be seen in the large vertical jump in both panels.

6.3.2 Sampling Quality & Extent

Now we evaluate the degree to which our simulations have converged along the unbinding pathways. This will contextualize the quality of our transition state estimates later. In Figure 6.8 the non-equilibrium free-energies are plotted for each replicate simulation for all ligands. Each panel shows superimposed free energies at fractions of the total sampling time for each replicate. The horizontal axis is shown the RMSD of the ligand after aligning the

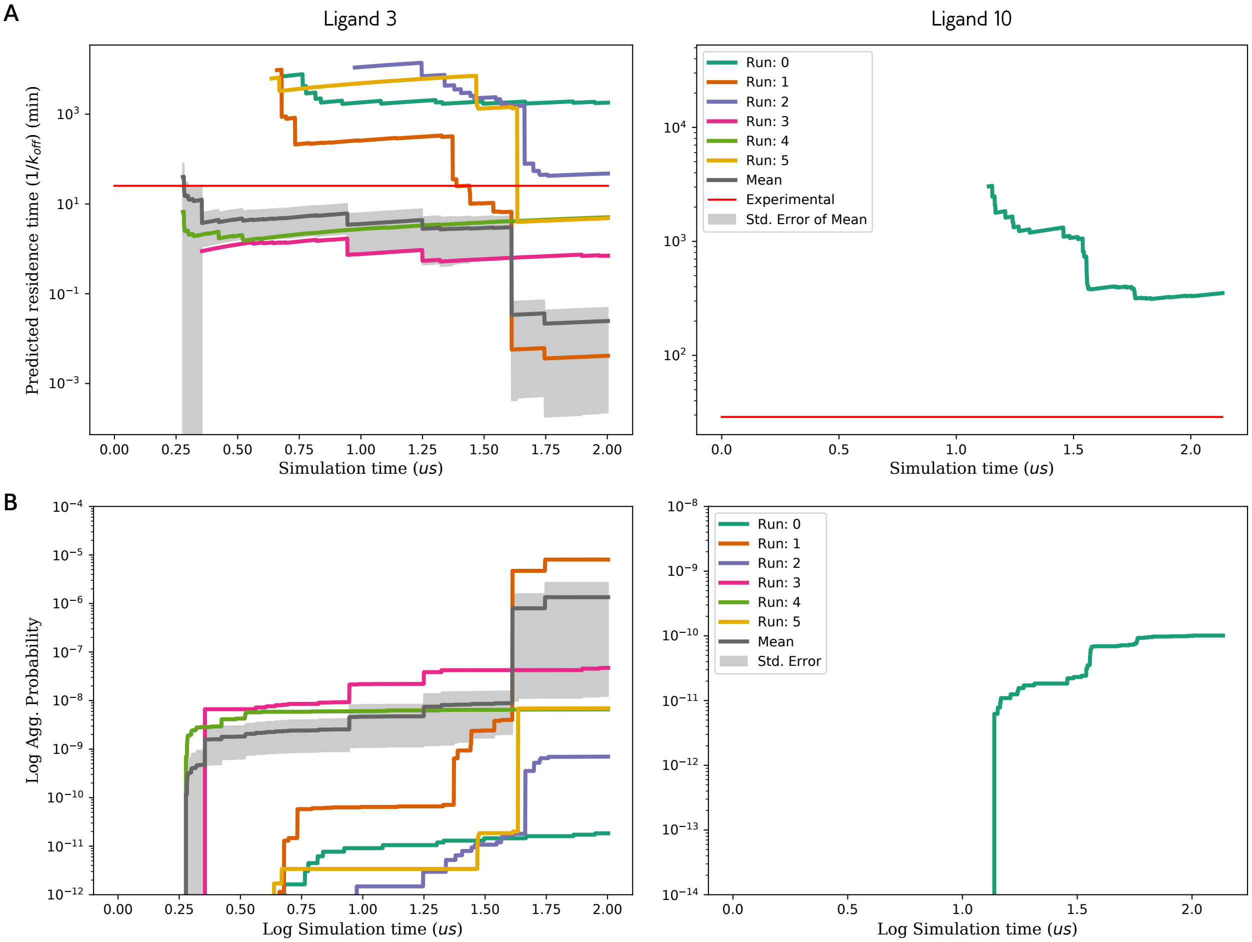


Figure 6.7: Plots of the rate estimates (**Panel A**) and aggregate probability flux through the unbinding boundary condition **Panel B**. The total aggregate simulation time across all walkers is given on all of the horizontal axes for each plot in microseconds. For all plots the individual replicate runs for which there were unbinding events observed are plotted in color lines (see the legend in plots) and the average of these is plotted in the dark grey line with the standard error of the mean shown as the light grey envelope. Plots in Panel A show the predicted residence time $RT = 1/k_{\text{off}}$ over time starting with the first unbinding event observed for a simulation. The horizontal red line is the experimentally determined rate as determined by Lee *et al.* (2014) [2]. Plots in Panel B show the cumulative probability over time that had crossed the boundary. Ligand 10 only had a single replicate observe unbinding.

sEH binding site to the initial structure close to the crystal structure. The ligand RMSD is a good projection as it captures the degrees of freedom that we are trying to characterize, and is also the same as that used in the distance metric in the WExplore resampling algorithm.

An individual panel shows a number of things. First, the extent of a curve to higher values of the ligand RMSD shows the overall progress of the simulation. This is useful for understanding just how far along simulations that did not observe unbinding events were. Second, it gives some idea of the convergence of the simulations. Obtaining full convergence is very difficult for such long timescale processes, and it is highly unlikely we have reached convergence. If the curves for the final fractions of the simulation do not show large changes then we have some evidence that there were was no active discovery occurring. For instance looking at the top panel of ligand 3 we can see that the first 3/5 of the simulation there was large changes in the free energy estimates, but that between 4/5 and the end there was very little change. Compared to panel 3 for ligand 3 we see that very early on that discovery had ended. For ligand 3 (of which we will focus our analysis more heavily on later) we see that while there is a lot of variance in the free-energy estimates in early portions of sampling, but in the final stages they are metastable (i.e. in a local minimum).

As for the other ligands many of them are obviously not converged as the free energy profiles are shifting even in the last fractions. Replicate 0 (first row) of ligand 10 however does look to be pretty stable in terms of convergence.

Now that we have analyzed the convergence of the individual replicate simulations we should compare the final results from each replicate to each other to assess variance across free energy estimates. Figure 6.9 the final free energies plots for each replicate superimposed by ligand. Here we see that there exists fairly significant variance between replicates. This is unsurprisingly prominent for ligands 10, 18, and 20 as none of these had convergence across most of the replicates. What is interesting is that even for ligand 3 in which each replicate looked to have metastable convergence, there is significant variance in the replicates. While all of the replicate curves however have roughly the same trends, they differ on the fine

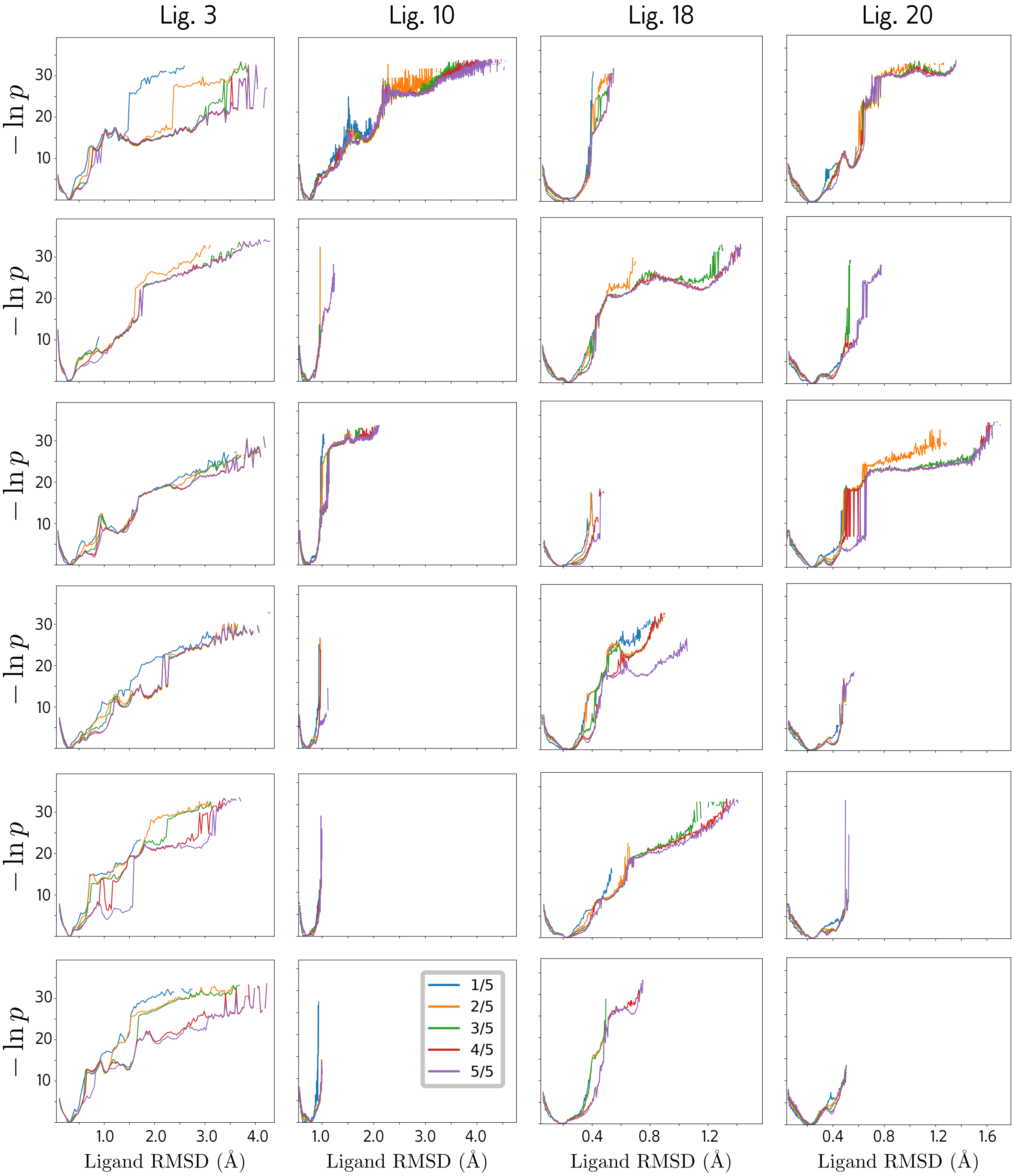


Figure 6.8: Convergence of replicate simulations shown as the free energy ($-\ln p$) on a ligand RMSD (Å) projection. Columns are for each ligand and rows are for each replicate. Curves in each are divided into five even fractions of each replicate simulation; in order blue, orange, green, red, purple.

details of the "landscape" and free energies at large RMSD values. This is not entirely unexpected given that the WExplore algorithm and ligand RMSD distance metric are designed to enhance exploration of state space, rather than exploiting discovered information to refine detailed free energies.

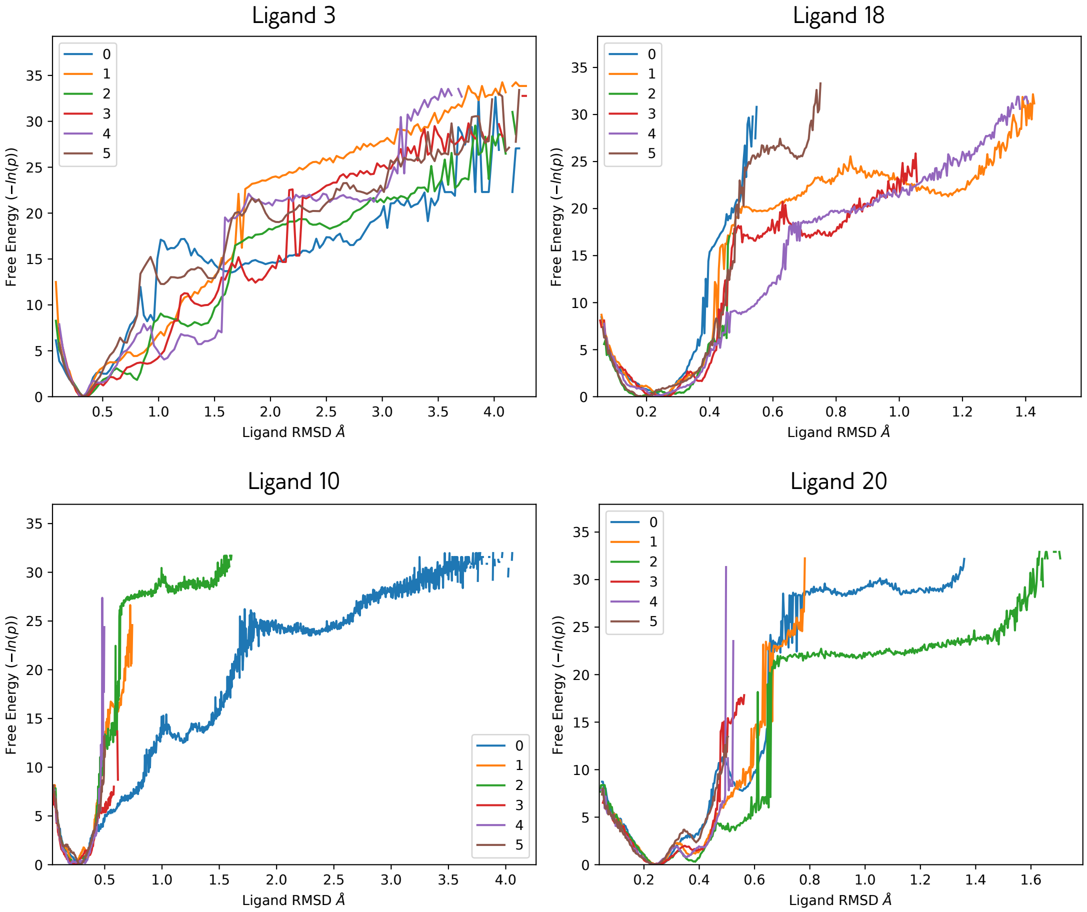


Figure 6.9: Final free energy $(-\ln p)$ of each replicate for each ligand over the ligand RMSD (Å).

One important question that is addressed by these simulations is the limit to which the current resampling algorithms, distance metrics, and other hyperparameters are able to sample ligand unbinding processes of increasingly longer timescales. From these results it is fairly clear that there is a steep increase in difficulty in the range of k_{off} sampled by the ligands in this study. The TPPU ligand (17; $k_{\text{off}} = 10.5(s^{-1})10^{-4}$) and ligand 3 ($k_{\text{off}} = 6.57(s^{-1})10^{-4}$) the simulations were able to readily sample unbinding events.

The relatively small difference between ligand 3 and ligand 10 ($k_{\text{off}} = 5.79(s^{-1})10^{-4}$) was concomitant with a large decrease in the reliability and quality of sampling. Given that ligand 10 was more difficult to sample than ligand 3 it is obvious that ligand 20 would be even more difficult ($k_{\text{off}} = 5.05(s^{-1})10^{-4}$), but it is curious that ligand 18 ($k_{\text{off}} = 6.14(s^{-1})10^{-4}$) has a greater k_{off} (shorter residence time) than ligand 10 and had no unbinding events. This could be either due to simply not having enough samples or actual experimental error (see Table 6.1).

One other possible explanation for the sharp difference is that slope of the free energy surface preceding the transition state (in unbinding) is more difficult to incrementally sample. This could be caused by low enhancement of some slow process in the orthogonal degrees of freedom to those captured in the distance metric. In Figure 6.10 we see the average free energy profiles from each of the ligand simulation groups. We can use this to directly compare the roughness of the landscapes. The curvatures of ligand 3 and 10 look fairly similar except for the higher free energy of ligand 10, but this is expected. This would indicate that the algorithm and hyperparameters are simply being saturated in terms of performance and that additional sampling (e.g. more walkers, longer simulations) would drive improvements in actual flux along the unbinding path.

However, if we look at the range of RMSDs close to the stable native state we can see some difference when comparing to ligands 3 and 10 to ligands 18 and 20. It is reasonable to look at this region in terms of the quality of sampling for ligands 18 and 20 as these were well sampled and seen to converge in Fig. 6.8 and have low variance between replicates, Fig. 6.9. For ligand 18 the slope of the free energy curve is much steeper, and there seems to be a few deep local minima close to the starting state for ligand 20. It is not clear exactly how these kinds of features would effect the resampling algorithm, but should be kept in mind for future investigations. Possible explanations could be that a protein conformational changes that must occur before the ligand can be released and this is poorly enhanced by the ligand RMSD metric. This could also be caused by more metastable local minima along the pathways which could use up a discovery "budget" or stall simulation progress.

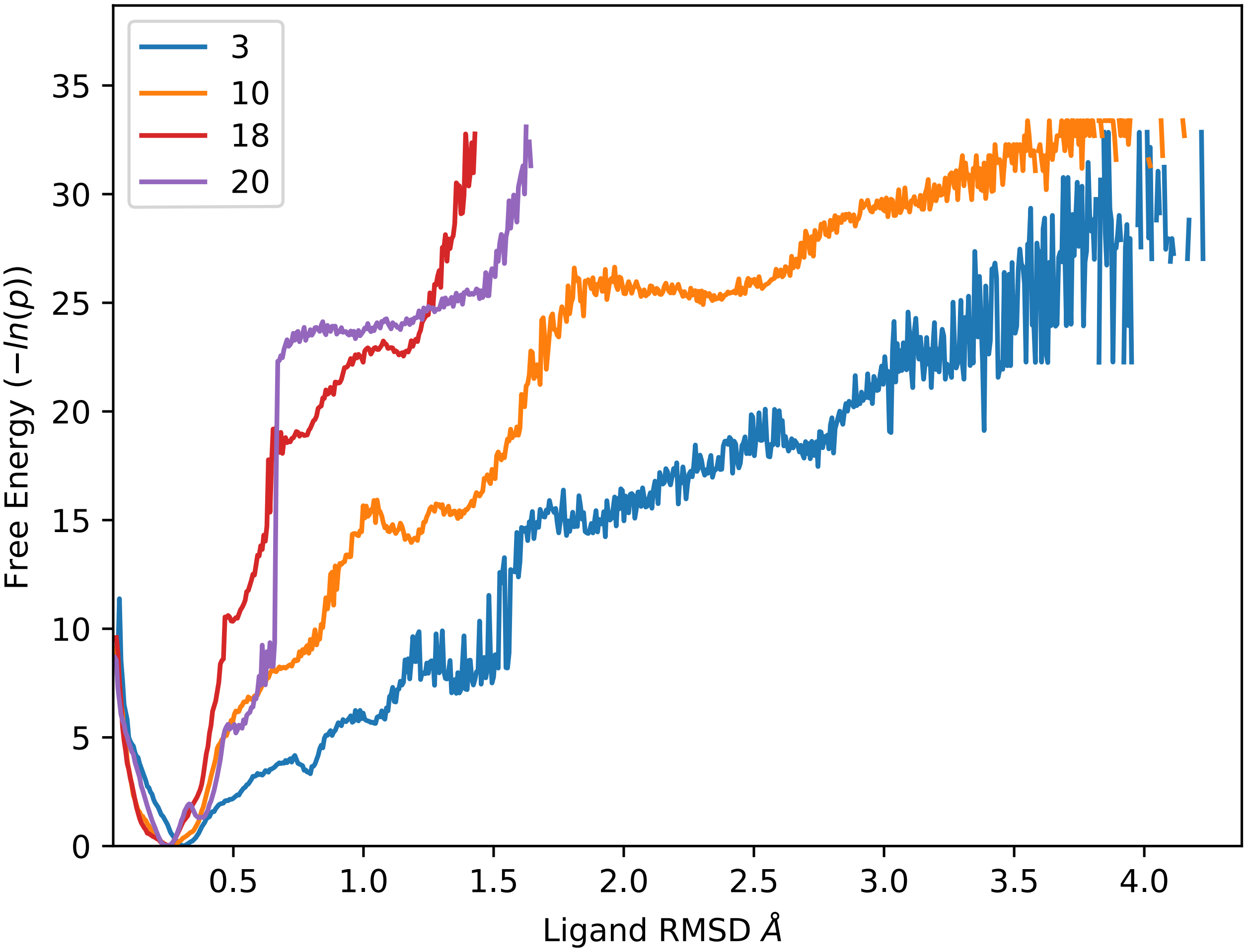


Figure 6.10: Average between each replicate final free energies ($-\ln p$) for each ligand over the ligand RMSD (Å).

6.3.3 Conformation State Networks

Simulations such as those in this study which have large amounts of diverse sampling of inherently high-dimensional state variables (i.e. atomic positions of proteins and ligands) can be difficult to visualize and obtain high-level understanding of their contents. The one-dimensional projections shown in the previous section have their uses, but they do not give information about diverse pathways and other complex motions without clever projections. For this reason we choose to create conformation space network (CSN) representations as has been done in previous investigations. The protocol for generating the clustering model and generating network depictions for Figures 6.11, 6.12, and 6.13 is given in Section 6.2.3.

We first look at the free energies ($-\ln p$) over the clustering projection which is computed the same as the ligand RMSD projections except over a discrete domain of clusters in Figure 6.11. These CSN depictions immediately show us that while for ligand 3 we were able to sample at least two distinct pathways between the bound states (lower right with low free-energy) to the unbound states, all of the other ligand simulations only discovered a single pathways. The overall shape of the ligand 3 CSN is very similar to that of TPPU (Figure 5.8). For the other ligands this difference is not unsurprising given that no more than one simulation observed full unbinding and from previous results for TPPU a single replicate often only unbinds via a single pathway (see Figure 5.3).

In addition to visualizing the free energies over the network, which tells us the probability of the states, we color the same networks according to different physical observables. The same ligand RMSD from the one-dimensional free-energy profiles is shown in Figure 6.12, and the solvent accessible surface area (SASA) is shown in Figure 6.13. For each of these the color is determined from the average of all the frames in each cluster. For these two observables there are no interesting observations to make as they do not seem to be different for the pathways for ligand 3 and seem to correlate fairly well with the free-energy.

6.3.4 Predicting Transition State Ensembles

In addition to the utility of CSNs as used in the previous section, they are also useful for predicting transition states by way Markov state model (MSM) theory. The ultimate use for MSMs in this study is to use them to predict the TSE for the simulations. The general protocol for doing this is given in Section 6.2.4. In this study we followed a data-driven strategy for choosing the basins used in computing committor probabilities as this might have an impact on the accuracy on the ultimate choice of TSE. There are three variables to consider:

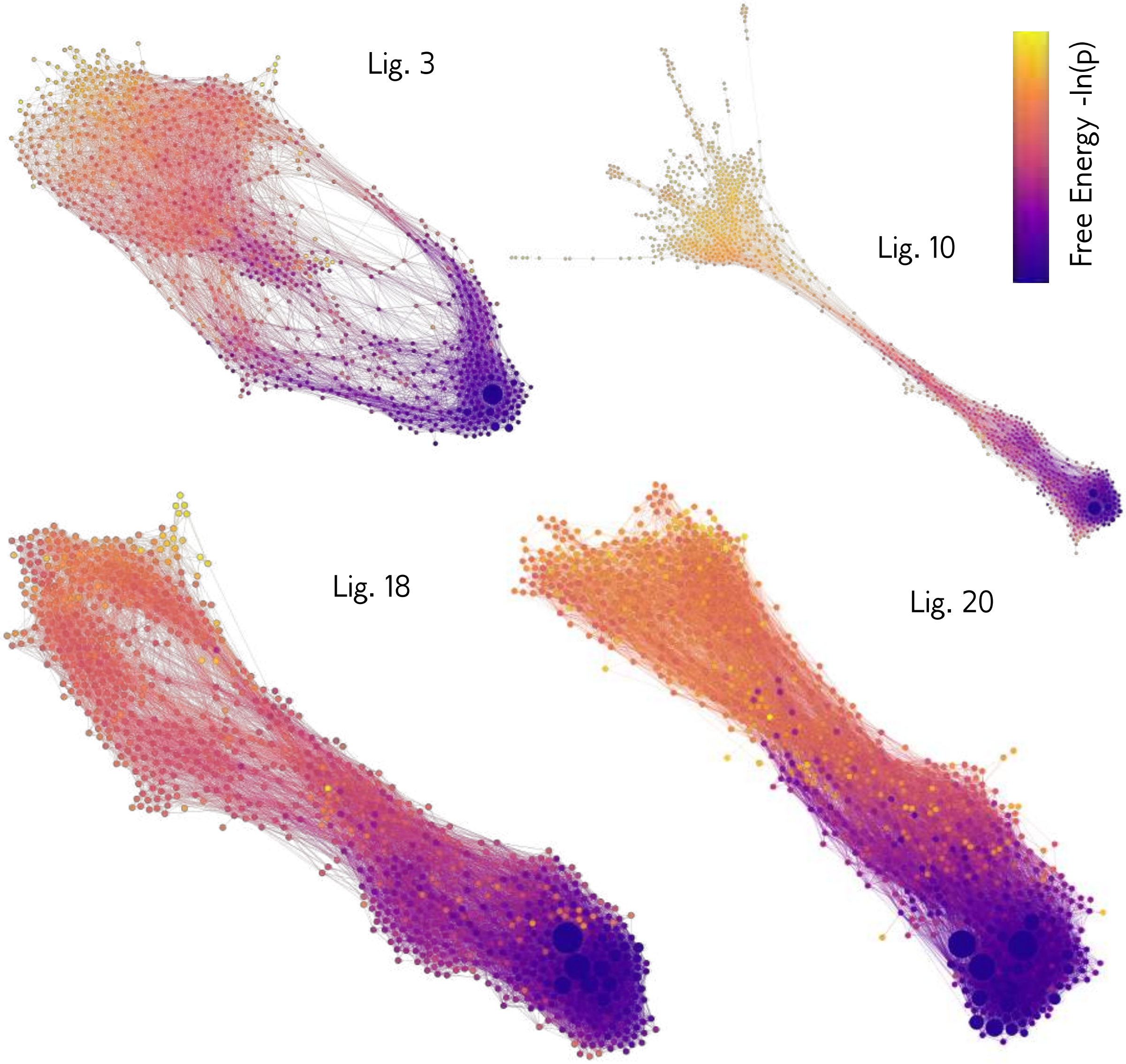


Figure 6.11: The free energies of the aggregate simulations over the conformation state network projections of each ligand. The color values are internal to each network. The size of the nodes is indicative (but not quantitatively accurate as there is a minimum node size cutoff) of the sum of the walker weights in each cluster.

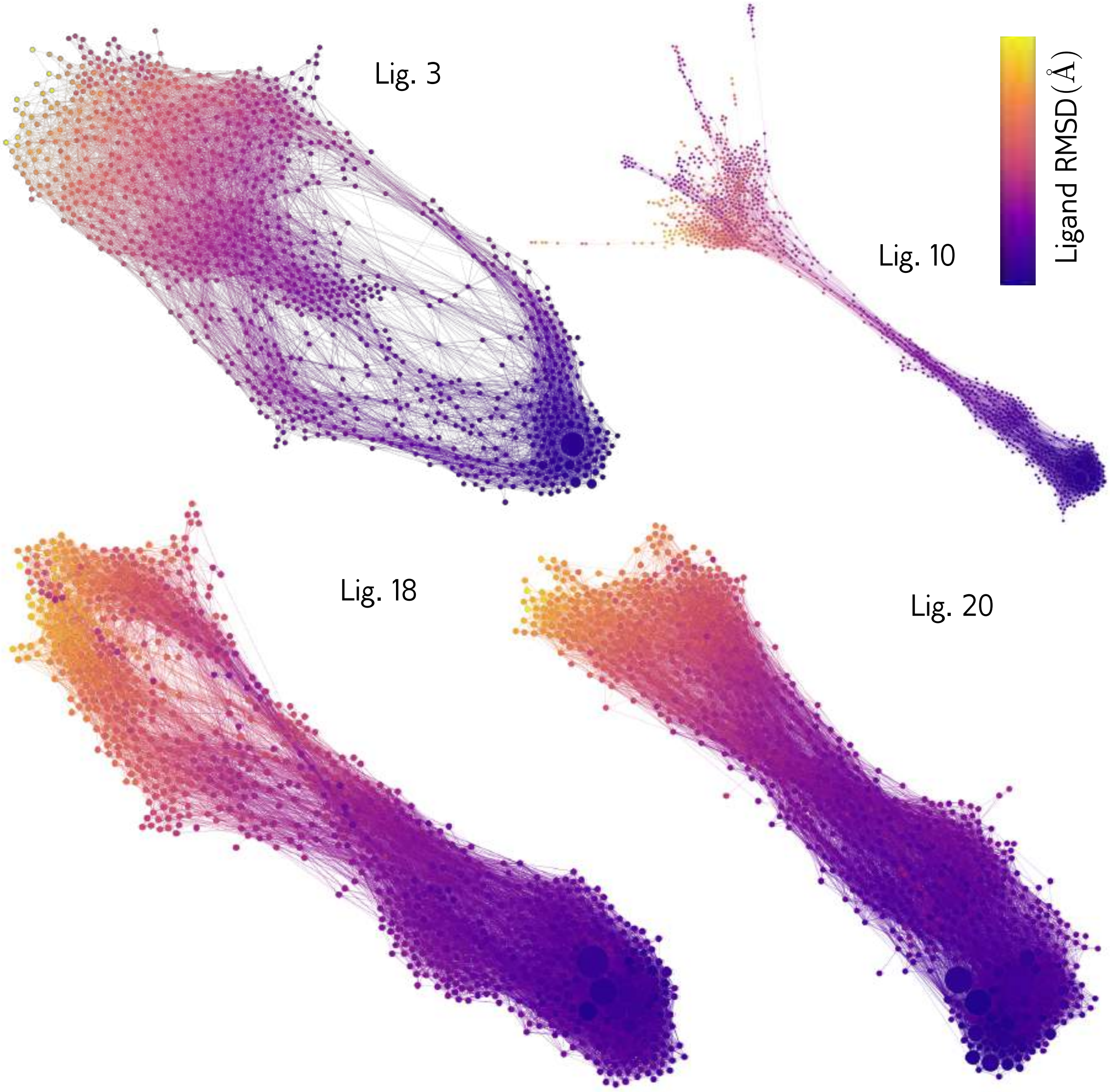


Figure 6.12: The ligand RMSDs (Å) of the aggregate simulations over the conformation state network projections of each ligand. The color values are internal to each network.

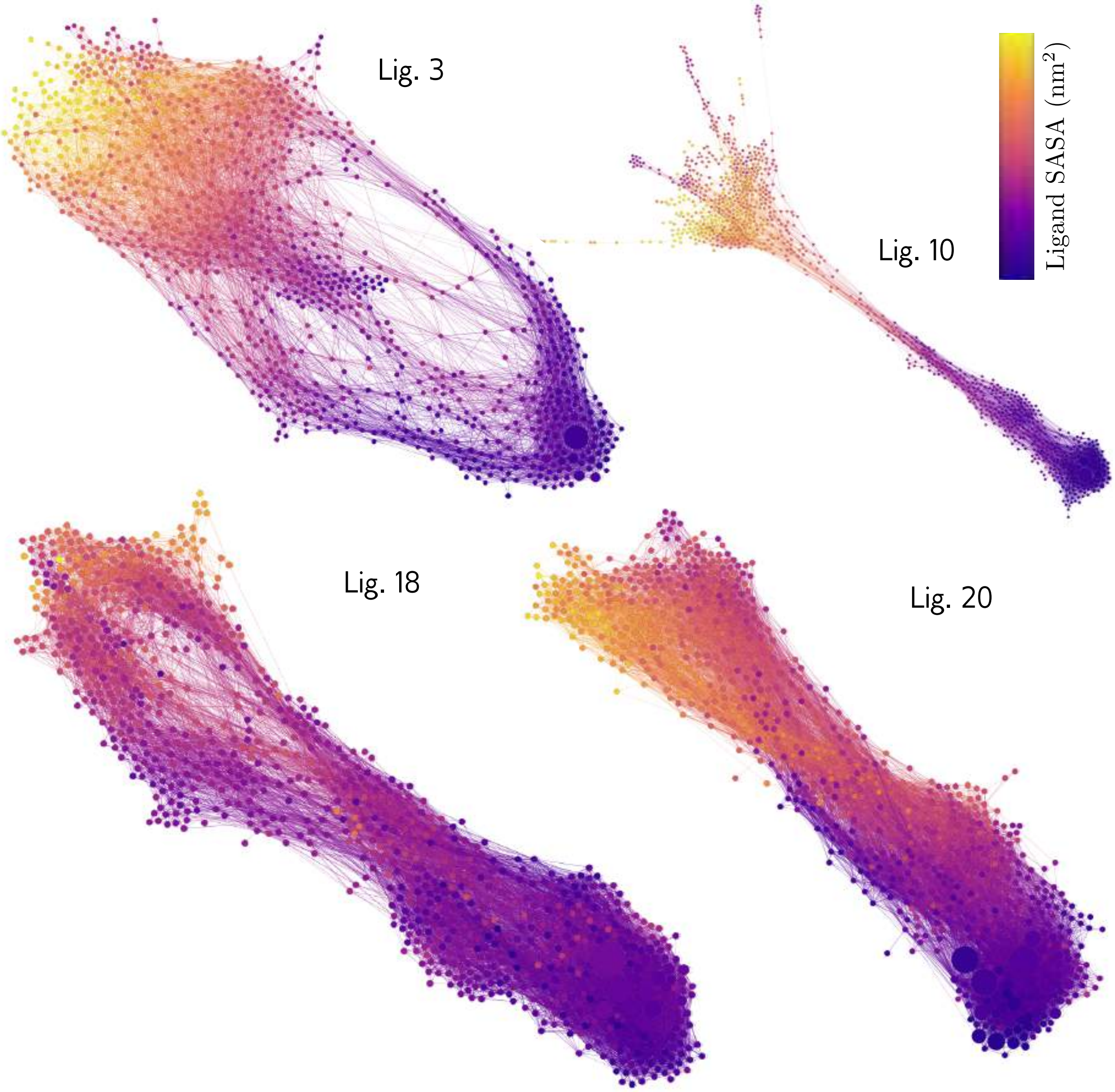


Figure 6.13: The ligand SASAs (nm^2) of the aggregate simulations over the conformation state network projections of each ligand. The color values are internal to each network.

1. The cutoff for the number of trimmed nodes.

2. The nodes included for the bound (source) basin.

3. The nodes included for the unbound (sink) basin.

The cutoff value for the trim nodes is the number of transitions (regardless of weight) that must be present between a node and the largest strongly connected subgraph for it to be included in the MSM. For the constraints of the MSM theory to be satisfied there must be at least one transition to and from the main subgraph. If there is not then the Markov chain is not ergodic as any transition to a weakly connected component would be irreversible. While a single transition is enough to satisfy the hard constraints of the theory, it may not provide adequate statistical power and could result in inaccurate transition probabilities. On the other hand the MSM nodes that are very near the steady-state boundary condition have relatively few samples and throwing too many away would exclude the actual unbound nodes from acting as sinks. For this parameter we chose to look at the following cutoff values: 1, 2, 5 and 10.

The other two parameters concern the choice of the basin nodes given an ergodically trimmed MSM and are chosen to correspond to the physical definitions of bound and unbound. Unbound basin nodes U are chosen based upon the max-min all atom distance between the ligand and protein of the cluster center. Given that the steady-state boundary conditions in the simulation (using the same metric) were set to 10 Å we chose values close to this where the ligand is still fully solvated (but may still be effected by long range interactions in the simulation): 7 Å, 6 Å and 5 Å. The bound basin nodes B were chosen based on an the same RMSD metric used in the definition of the distance metric for the WExplore resampler (see Section 6.2.2). As was shown in the free-energy plots of this RMSD observable 6.10 there is a deep free-energy for small RMSDs, as such the following parameters chosen to test are: 1 Å, 2 Å and 3 Å.

Figures 6.14 and 6.15 show the results of a parameter sweep over a subset of these variables. In the first (Fig. 6.14) the effect ergodic trimming cutoff is examined (given a constant choice of bound basin) and in the second (Fig. 6.15) the effect of the choice of both bound and unbound basins (for a constant choice of trim cutoff). In both figures two metrics are used to evaluate the parameter choices:

- the distribution of nodes across the range of committor probabilities (top quadrants), and
- the distribution of total node weights (sum of microstate walker weights) across the committor probabilities near $p_{B \to U} = 0.5$ (bottom quadrants).

The full possible grid of parameters is not shown as this would be very difficult to visualize, and as it turns out the bound cutoff has little effect on any of the distributions (for a trim value of 1) (see Fig. 6.15).

In Figure 6.14 the choice of the trimming cutoff is examined. The trim cutoff was varied along with the unbound cutoff as these are related. If too much of the CSN is trimmed there will be fewer nodes to choose from for the unbound basin. We see that the distribution of node numbers for ligand 3 (TL) is fairly robust with respect to these changes. This is likely due to a large number of unbinding events and sample density near the unbound basin compared to ligand 10. As the TR quadrant shows the node count distribution is much more sensitive to the cutoff value especially at 10 which obviously has too few nodes in the unbound basin. For ligand 10 it seems a choice of 1 or 2 is necessary.

The total weight distributions around the transition states shown in the bottom row of Fig. 6.14 are much more sensitive to the choice of parameters. Again for ligand 10 (BR) we see relatively miniscule populations in this range (when compared to ligand 3 in BL) for all choices. It is worth noting here that the choice of trimming has a much larger effect than the choice of the unbound basin criterion for both ligand 3 and 10. The choice of the trimming changes the shape of the distribution (which is roughly bimodal), while the choice of the unbound basin is evident mostly as a shifting of the populations along $p_{B \to U}$. The total lack of probability density in the range for a trim cutoff of 10 in ligand 10 excludes it, and it appears that there is a more uniform distribution of density for trim 1. For ligand 3 it appears that either trim 1 or 2 makes the most sense as the minimum of density is co-occurring with TSE $p_{B \to U} = 0.5$, which in theory should be the least populated region given it is the peak of free-energy between the two basins. Particularly the distributions for $(\text{trim} = 2, \text{unbound} = 5)$ and $(\text{trim} = 1, \text{unbound} = 6)$ are the best. Given the choice between the two we would choose the one with more samples at trim = 1. Thus, we use this value in the rest of the analysis.

With our choice of trimming set, we turn our attention to only the choice of the basins. In Fig. 6.14 we ignored the unbound cutoff as there is no dependency between it and the trimming criterion. Figure 6.15 thus compares the range of choices for either basin. We can immediately see from this that the choice of bound basin has almost no effect on our metrics. This indicates that at least near the bound state the clustering was effective. That is, adding or removing a few nodes from the basin has little effect due to the high connectedness and few bottlenecks in the region. It is likely given that free-energy landscape from Fig. 6.10 that we would see greater sensitivity of committor probabilities in the range of 0.5 Å to 1.0 Å. This choice is largely arbitrary given the insensitivity, but we choose the one with the larger number of samples in the absence of any other criteria: 0.3 Å.

We also can easily tell that unbound criterion has little effect on the overall node count distributions (TL & TR). Using the same decision process for the trim cutoff we observe the probability densities near $p_{B \to U} = 0.5$ across the parameter grid. First we see that ligand 10 (BR) is not effected much and provides little useful information by which to make a decision. From the ligand 3 results (BL) however we see that there are a few strong co-occurrences of local minima to $p_{B \to U} = 0.5$ for unbound values of 6 Å.

Following this parameter sweep we have identified a single set of parameters for both constructing our MSM and calculating committor probabilities: $(\text{trim} = 1, \text{unbound} =$

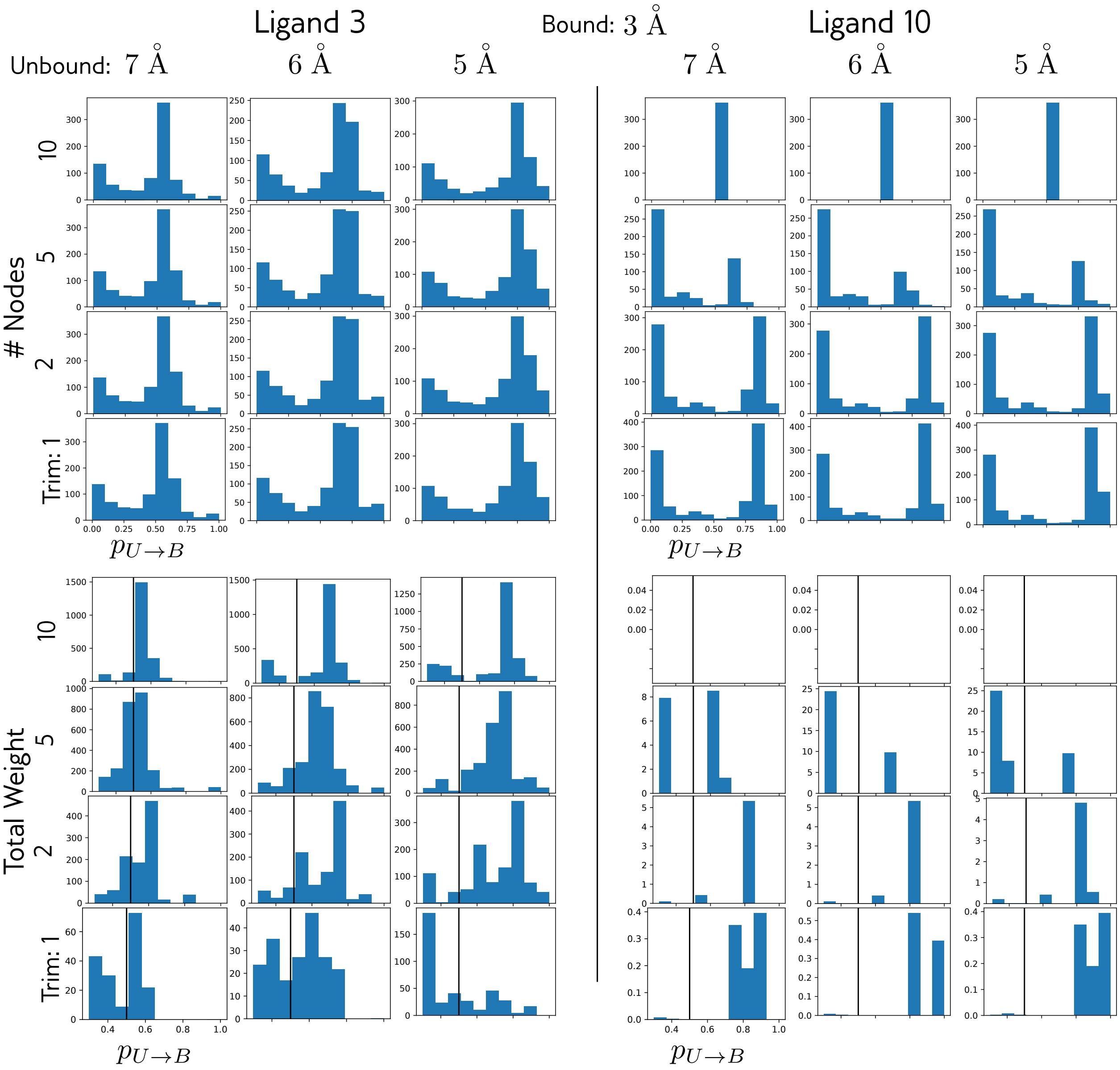


Figure 6.14: First part of parameter sweep for choosing committor probability basins. The plots are arranged as 4 quadrants: top left (TL), top right (TR), bottom left (BL), and bottom right (BR). Within each quadrant is a grid of histograms for different hyperparameters. The dependent variables are binned over the committor probabilities. The top row quadrants (TL and TR) show the number of nodes within each bin and the bottom row (BL and BR) shows the sum of the microstate walker weights in each bin. Additionally the bottom row shows a truncated range of $p_{B \to U}$ to highlight the probabilities around the transition state estimate (marked by the vertical black line at $p_{B \to U} = 0.5$). In each quadrant there are 4 rows and columns. The columns correspond to the different values of the metric used for making the unbound basin and the rows correspond to the values of the trimming cutoff used. All plots were made with a bound cutoff value of 3 as indicated at the top center of the figure. The unbound metric is the max-min distance between any atom of the ligand and the protein and the bound metric is the ligand-RMSD to the starting structure after aligning the binding site to the starting structure. The left quadrants (TL and BL) are plots for Ligand 3 and the right quadrants (TR and BR) are for Ligand 10.

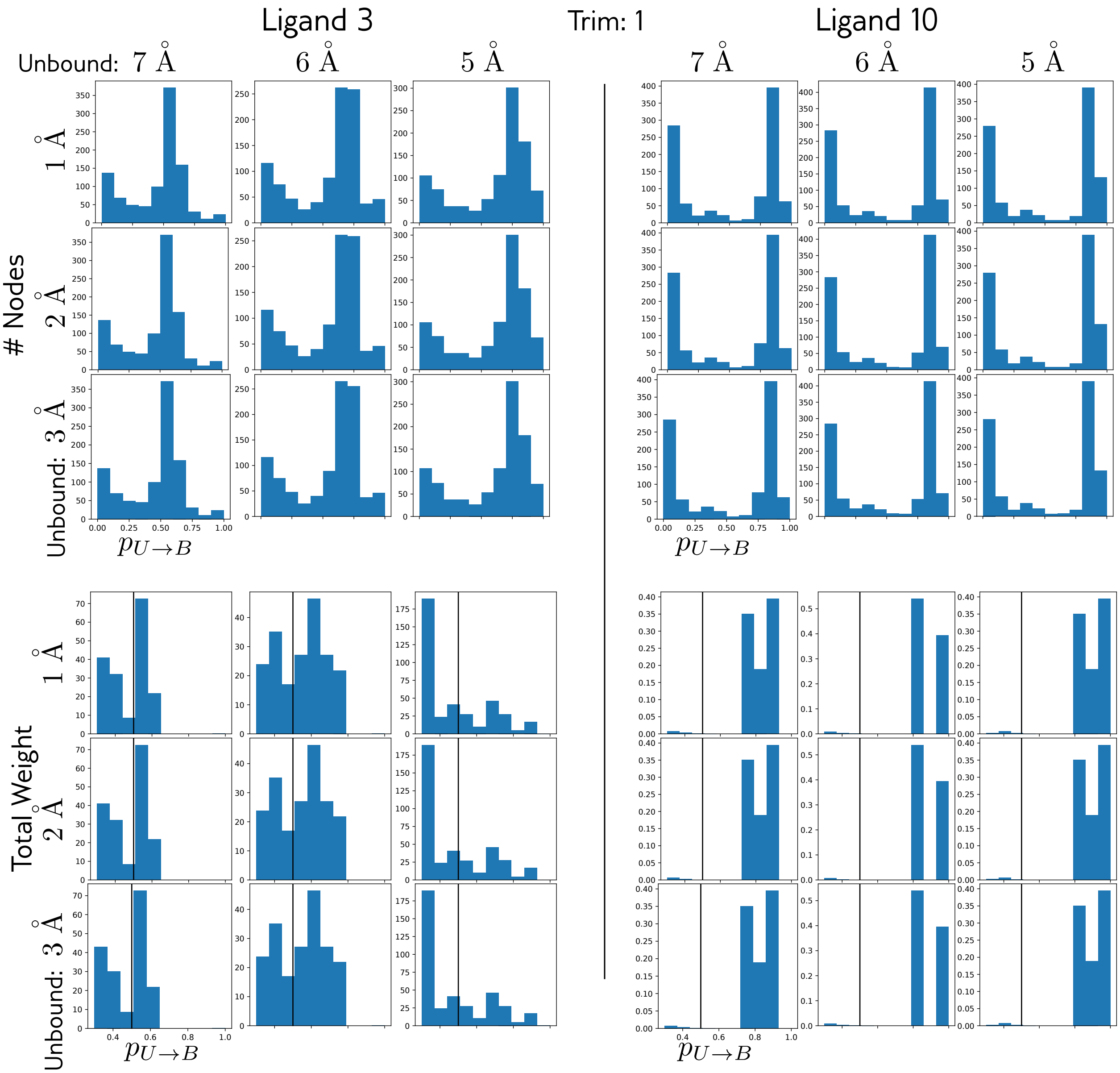


Figure 6.15: Second part of parameter sweep for choosing committor probability basins. The plots are arranged as 4 quadrants: top left (TL), top right (TR), bottom left (BL), and bottom right (BR). Within each quadrant is a grid of histograms for different hyperparameters. The dependent variables are binned over the committor probabilities. The top row quadrants (TL and TR) show the number of nodes within each bin and the bottom row (BL and BR) shows the sum of the microstate walker weights in each bin. Additionally the bottom row shows a truncated range of $p_{B\to U}$ to highlight the probabilities around the transition state estimate (marked by the vertical black line at $p_{B\to U} = 0.5$). In each quadrant there are 4 rows and columns. The columns correspond to the different values of the metric used for making the unbound basin and the rows correspond to the values for the bound cutoff used. All plots were made with a trim value of 1 as indicated at the top center of the figure. The unbound metric is the max-min distance between any atom of the ligand and the protein and the bound metric is the ligand-RMSD to the starting structure after aligning the binding site to the starting structure. The left quadrants (TL and BL) are plots for Ligand 3 and the right quadrants (TR and BR) are for Ligand 10.

$6, \text{bound} = 0.3$). Using these parameters we calculate committor probabilities for all nodes and display them on the CSN depictions of Figure 6.16. There is nothing surprising here to report as the committor probability looks to correlate well with all of the other observables examined previously. Figure 6.17 shows the specific choice of nodes for the bound and unbound basins, in green and blue respectively.

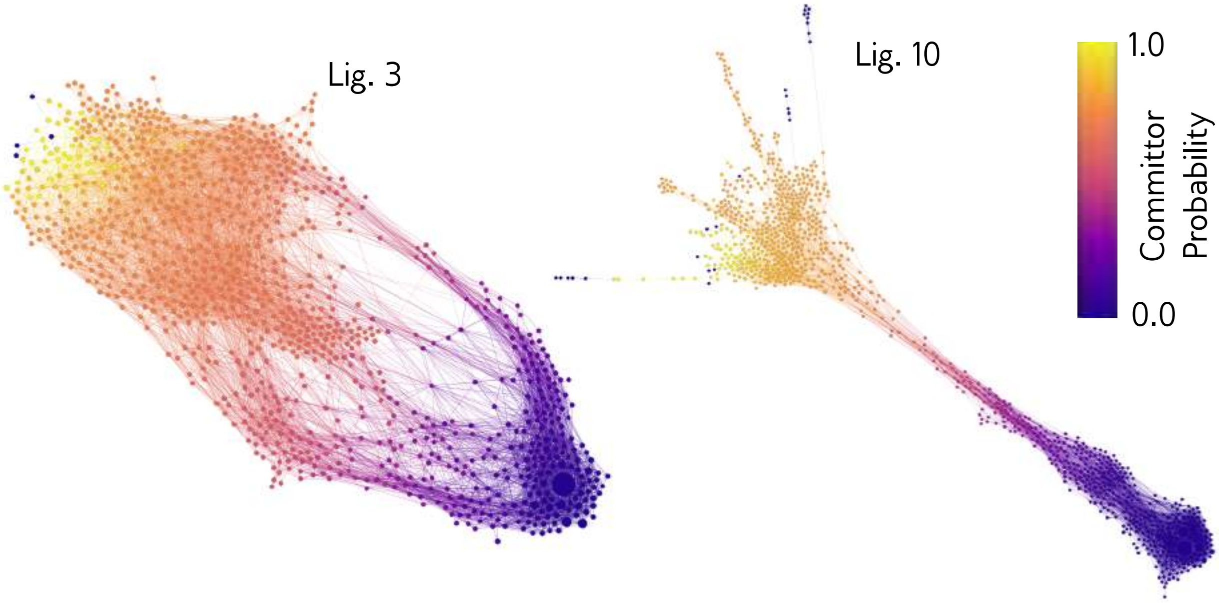


Figure 6.16: Committor probabilities for conformation state networks of ligands that had unbinding events. See basin definitions in Fig. 6.17. The black nodes were trimmed from the network when calculating the committor probabilities.

The estimate of the TSE is also shown in Figure 6.17 in pink. In this figure we compare the TSE for our new ligands to that identified for TPPU from Section 5.3.5. In this regard ligand 3 looks fairly similar where there are roughly two separate transition states for each pathway. Ligand 10 is similar in that the TSE seems to be located at the sparsest point in the network.

6.3.5 Characterizing Transition State Ensembles

Given the TSE model from Figure 6.17 we are now interested in the molecular structure of the substituent states. Our first interest is in understanding the gross geometric positioning

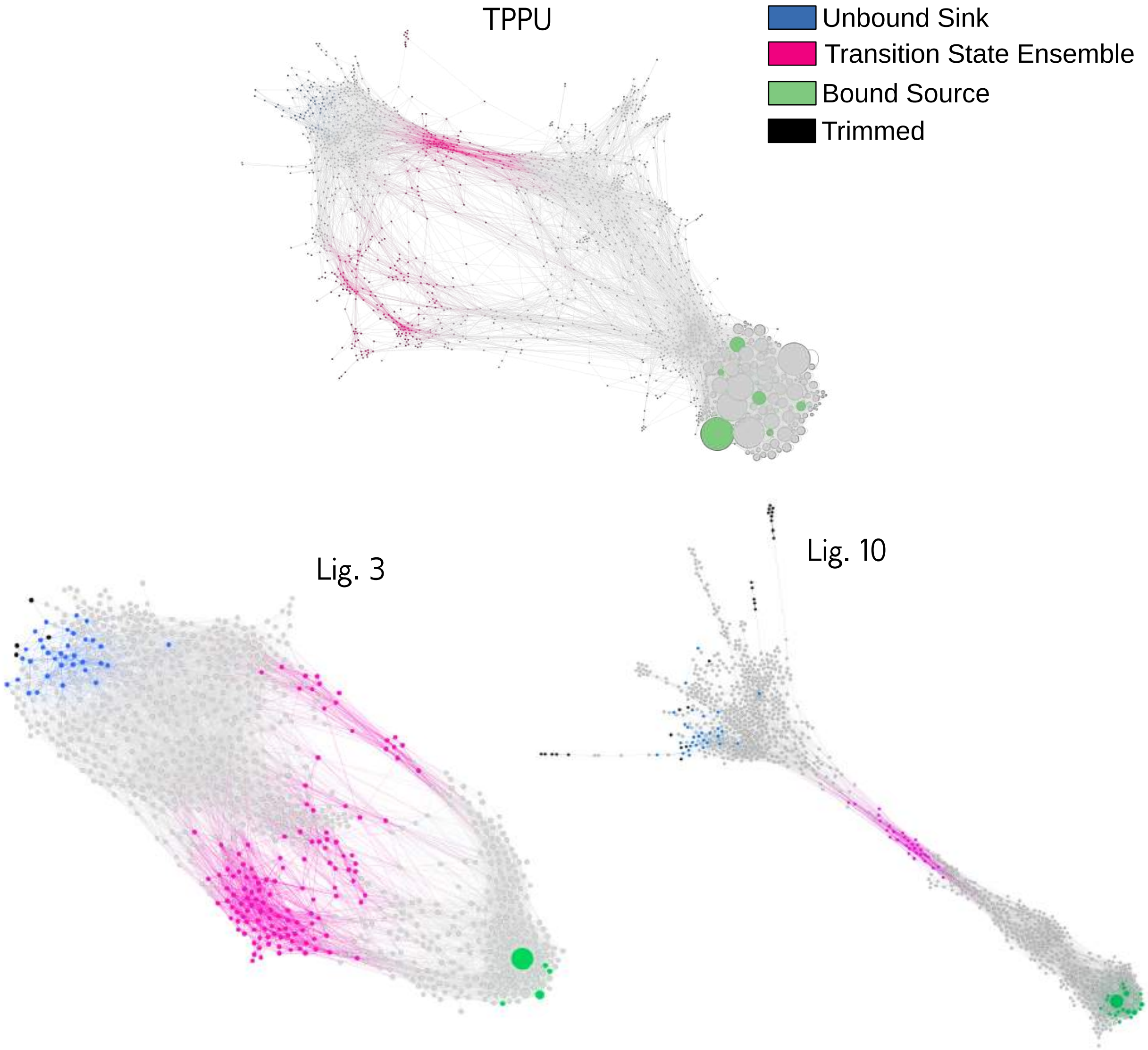


Figure 6.17: Conformation state networks colored to show the choice of source and sink basins (green and blue respectively) and the transition state ensemble prediction (committor probabilities between 0.4 and 0.6 inclusive) in pink. Other nodes are colored grey and trimmed nodes are shown in black.

of the ligand with respect to the protein binding site. This will be useful in characterizing differences in the different unbinding pathways.

In Figure 6.18 we start by showing the centers of geometry (“centroids”) of the homologous atoms of each ligand (see Fig. 6.6) for each cluster center in the TSE. These are represented as the colored spheres superimposed onto the starting structure. Looking at the “Top” view for each ligand we can see that the “cloud” of centroids is roughly an oblong shape that has two dense clusters at each end. To quantify these intuitive descriptions we performed principle component analysis (PCA) over the 3-dimensional coordinates of centroids. This is used as a convenient coordinate transformation that will capture the shape of the cloud and so the resulting principle component vectors (PCs) are 3-dimensional. For each ligand a separate model was trained. We then use projections of the centroids onto the PCs to color them in Fig. 6.18. The first two columns show two views of the centroids colored by the PC that explains the greatest amount of the variance in centroid positions (PC-0), while the third column by the PC that explains the next most amount of variance (PC-1). The PC-0 vector roughly corresponds to the left-to-right variation (relative to Fig. 5.1), where red corresponds to the left hand side (LHS) and blue to the right hand side (RHS). The in-out variation is captured well by PC-1. It appears that there are more centroids for TPPU because the CSN from these simulations was created with 2000 states compared to the 1000 states used in this study.

From this figure it appears that the gross geometric position of the ligand in the TSE is very similar between TPPU and ligand 3. Here we have elided analysis of ligand 10 both for brevity as well as a lack of confidence in the accuracy of our transition state estimate due to the small amount of data collected. The only notable differences between ligand 3 and TPPU seem to be that the RHS centroids in ligand 3 hook backwards into the binding site and that the from the “Side” view looks to not extend as far out of the binding pocket.

According Figure 6.18 centroids seemed to cluster in the LHS and RHS of the binding pocket. Naturally we might expect that these correspond to the two distinct pathways

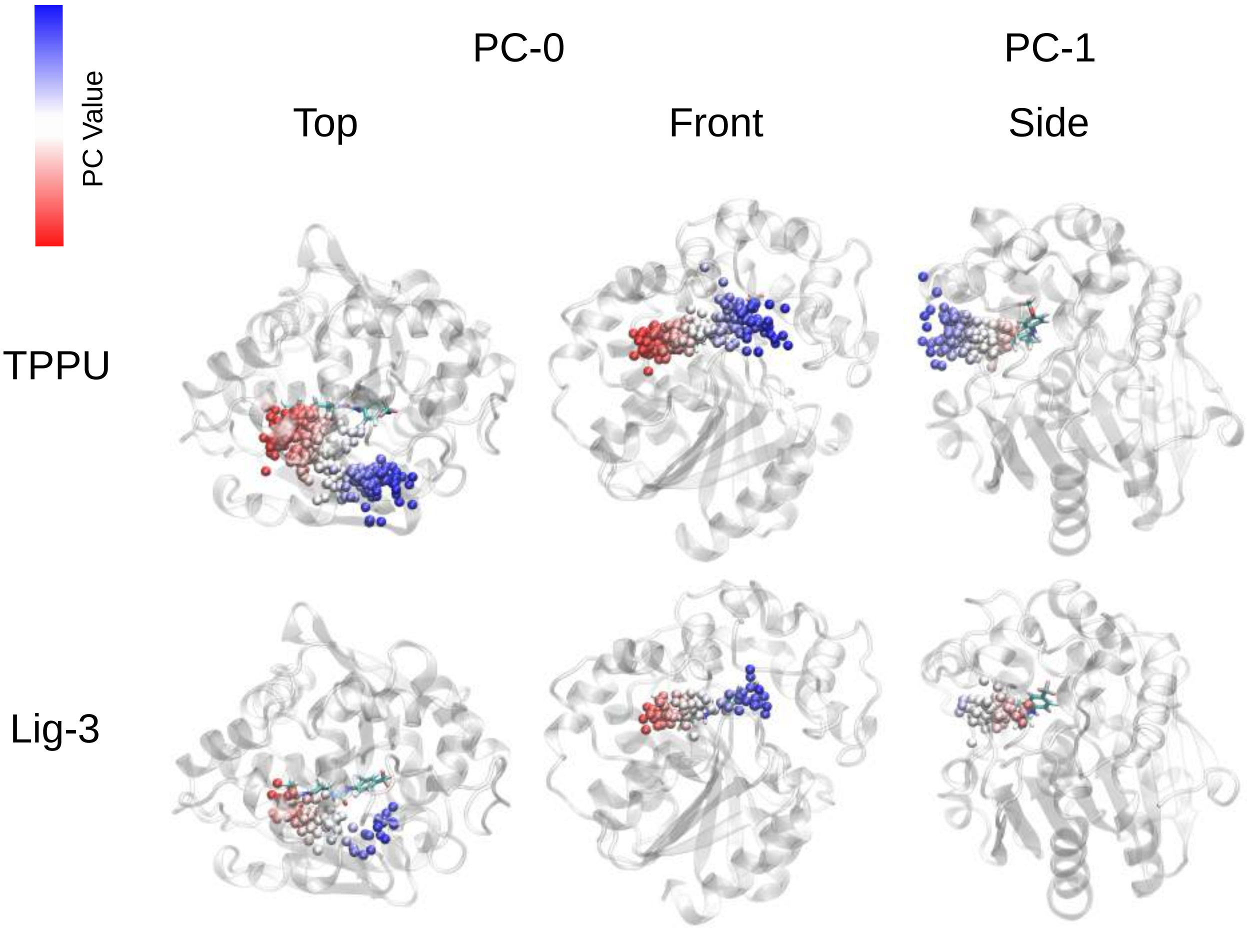


Figure 6.18: 3D structures with the centers of geometry (centroids) of homologous ligand atoms in the transition state ensemble shown as small spheres. The cartoon representation of the protein is of the initial state of the simulations, and includes the ligand in a cyan colored licorice representation. The first two columns show the "top" and "front" view of the protein and the third column shows the "side" view. The colors of the spheres is based on their projected value onto the relevant component (PC) from the PCA analysis of the spheres. PC-0 is the mode that explains the most variance in the PCA model, and PC-1 the second most. The first two columns are colored by PC-0 projections and the third column for PC-1.

observed in the CSNs. To verify this we projected the cluster center of each node in the networks onto PC-0 and colored the network according to this value, which is shown in Figure 6.19. Indeed, from this visualization we can see a strong correspondence between path bundle to PC-0 projection values for both ligand 3 and TPPU.

Finally, we calculate the free-energies for both ligands over the top two PCs in Figure 6.20. The landscape for PC-0 is interesting because the both the positive and negative branches for ligand 3 are roughly symmetric whereas for TPPU it is asymmetric and skewed towards the positive end (blue; RHS). This can also be observed if we compare the free-energies on the CSNs of ligand 3 in Fig. 6.11 (symmetric) and TPPU in Fig. 5.8 (skewed towards Path 1).

The projection onto PC-1 is also interesting in that we observe a well defined local minima on the slope of the positive branch, whereas the slope for ligand 3 is smoother. The negative branch of the curves correspond to the ligand getting even more deeply bound in the protein, which is relatively much higher energy than moving through the binding site. Nonetheless some very high energy states were observed. We find it unlikely these are uninteresting given the high free-energy of the states, but we do find it as an opportunity to improve our distance metrics in resampling to avoid this unwanted discovery in simulations.

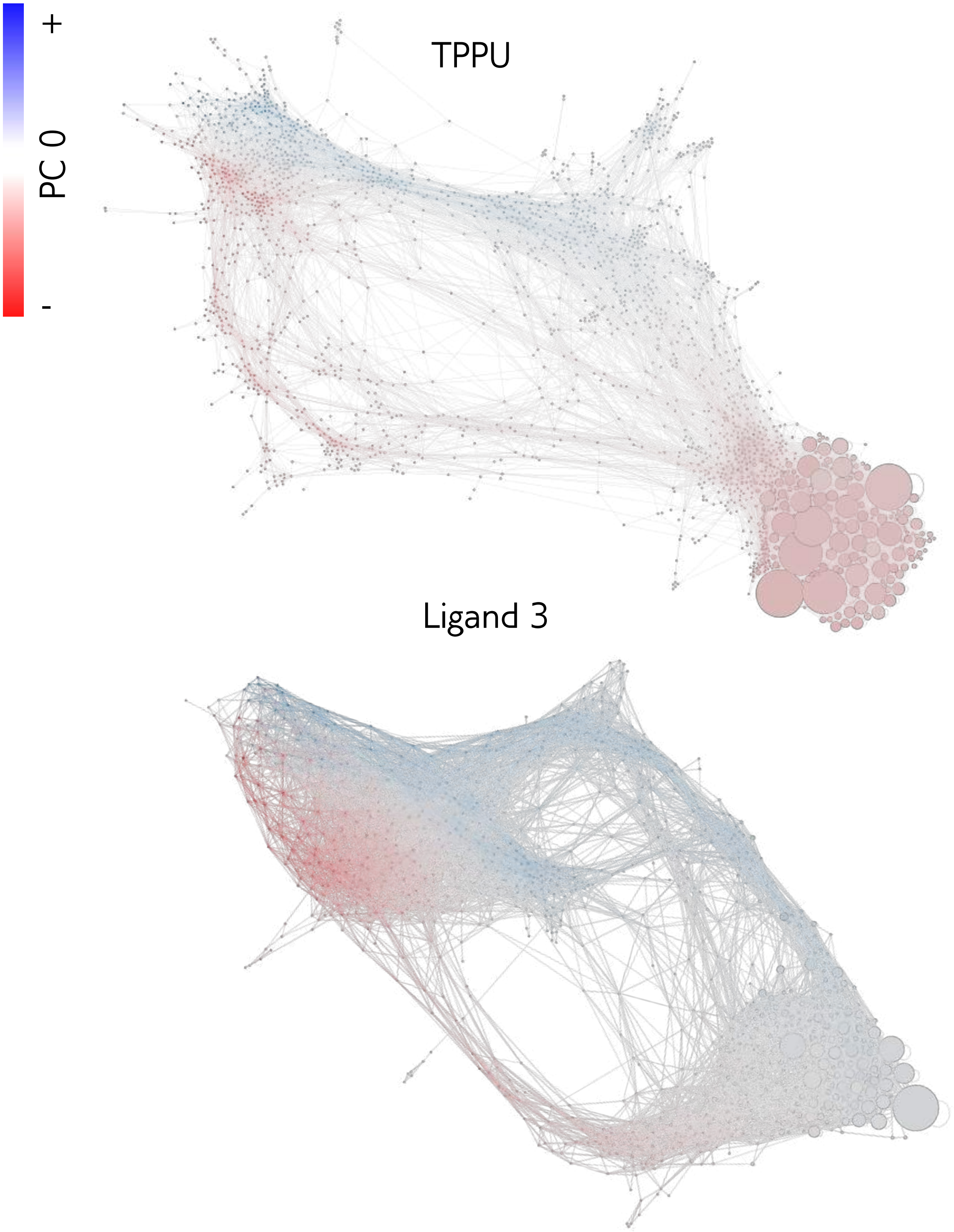


Figure 6.19: Conformation state networks of each ligand where each node is colored according to the projection of the center of geometry of the homologous atoms of the ligand from the cluster representative (here the cluster center). Color scales are internal to each network. Note that the specific layout for ligand 3 is different than before.

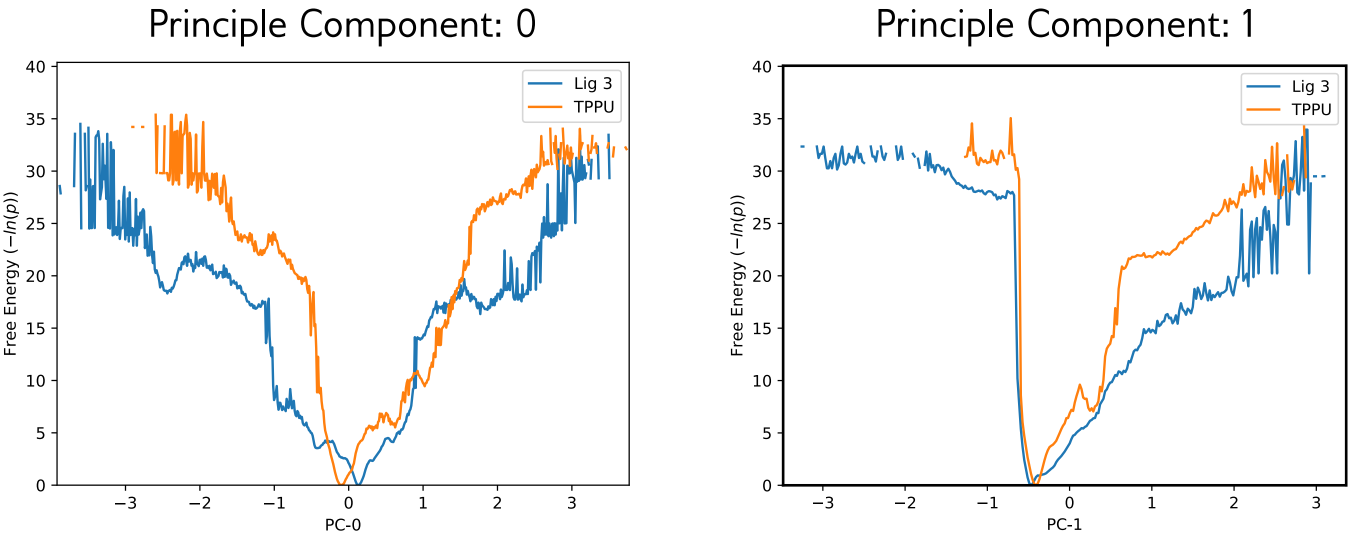


Figure 6.20: Free energy profiles ($-\ln p$) over the top two principle components of the PCA model for each ligand. The projection values are calculated from the centers of geometry of the homologous ligand atoms.

6.4 Discussion

This study is a partial success in that only two out of the four chosen ligands had unbinding simulations obtained, and the results for ligand 10 were obviously undersampled. Disappointingly, ligand 18 and 20 were both of the predicted non-pathway hoppers. This does not render the current results useless however as we are still able to do a comparative analysis of ligand 3 results to our previous TPPU simulations from Chapter 5.

We found that the rough "anatomy" of the CSN models for each are very similar in terms of topology. Similarly the location of the TSE on the CSN was also similar between the two. This is recapitulated in the both the positions of the TSE ligand centroids relative to the binding site as well as the PCA vectors describing this point cloud.

The primary difference between these two ligands is evident from the visualization of free-energy on the network and over the projection onto PC-0 from the TSE ligand centroid PCA model. From this we observed that the difference in free-energies between the two pathways was more uniform in ligand 3 and significantly skewed towards path 1 for TPPU (using the nomenclature from Figure 5.4). From the differences in the ligands we hypothesize that the larger hydrophobic R_2 group leads to an increased favorability for the hydrophobic LHS of the binding pocket. This potentially helps overcomes steric hindrance or balances interactions in the center and RHS of the binding pocket, and thus increasing favorability for the LHS that is associated with path 2.

So while it is obvious from this result that the overall characteristics of the pathways between the two are preserved, there is still an observable perturbation of the transition state in terms of the free-energy balance. We hesitate to qualify ligand 3 as having “hopped” to a different pathway. However, in this situation it would be difficult to target a specific pathway’s transition state for design given that either are equally as likely. Furthermore, if one is destabilized then the other will just become the dominant pathway nullifying any perturbation. In reality it is not clear whether design towards one or the other pathway is actually possible (e.g. specific targetable interactions) or whether the bottleneck is something like ligand solvation. From this analysis the main takeaway is that the possibility of pathway hopping complicating targeted design against specific transition states is relatively high.

Returning to our original hypothesis we can neither accept or reject that our S score is predictive of pathway hopping as we do not have any control samples (ligands 18 and 20). However, that our initial data point for ligand 3 did show some evidence of transition state plasticity is significant, as it was a necessary (but not sufficient) condition for accepting our hypothesis.

CHAPTER 7

SUMMARY, IMPACTS, & OUTLOOK

In Section 1.3 we outlined the goals of this thesis. In this chapter we revisit these points and assess progress towards these goals, the overall impact for the field, and challenges that remain.

The first goal was to characterize ligand unbinding pathways for the systems of interest, which was successful in almost all of the studied systems. While there was only partial success for very long residence time inhibitors of sEH (Chapter 6) our results are nonetheless encouraging for future investigations. Encouragingly, for all of the simulations roughly the same set of parameters and algorithms were utilized throughout the simulation. This was surprising as typically some degree of parameter optimization is necessary in general. As it stands there are many different approaches which could be taken to obtain successful simulations of the long timescale inhibitors including:

- increasing the amount of sampling for each replicate simulations (i.e. run simulations for longer);
- increasing the number of replicate simulations;
- improving MD force field parametrizations and system setup to improve system stability allowing for longer force integration timesteps thus improving sampling efficiency;
- changing of general WE simulation parameters such as the cycle sampling time and walker minimum & maximum probabilities;
- changing the WExplore resampler parameters such as the number of allowed regions, region sizes, or total number of walkers;
- utilizing or developing more efficient resampling algorithms (e.g. REVO); and/or

- utilizing different distance metrics other than the ligand RMSD used in all simulations.

The continued use of a single simulation protocol to a number of increasingly challenging systems was useful in this context to evaluate the robustness of the method. Furthermore, positive results from our initial study of a real drug system (sEH-TPPU) were attempted to be capitalized on for the benefit of medicinal chemists interested in it for real drug design. In the light of the challenges encountered in Chapter 6 we would suggest returning to some of the model systems in order to do more thorough hyperparameter optimization of these algorithms. Towards this we suggest the development of a battery of common metrics by which to score hyperparameters and algorithms

Included in these metrics should be the unbinding rate estimates. While we believe that the calculation of accurate rates is in general a very difficult problem not entirely solvable through improving enhanced sampling methods, i.e. accuracy of force fields. However, as we have seen the precision of our simulations is generally very poor for estimating rates. Even without accurate absolute rate estimates it should be possible to do non-parametric (i.e. rank ordering) comparisons of ligands given rate estimates from simulation. Given the large variances in estimates we have low confidence that these analyses would be valid. Thus, one major issue that should be addressed is understanding the sample sizes (in terms of entire replicate simulations) that are needed to generate precise estimates. We also suggest that there may not be a single set of parameters that is optimal for both discovering unbinding pathways *ab initio* (exploration) as well as refining for properties of interest (i.e. rates and transition states). Further, use of complementary enhanced sampling algorithms other than WE such as metadynamics or replica exchange may ultimately provide the best results in an efficient manner.

The large variance between replicates not only affects rate estimates but ultimately other important characteristics such as our transition state models. In our studies we were able to generate very detailed models of TSEs which could be utilized for drug design. In this sense it was a success since these models are largely unique outputs of WE simulations (in fact we are unaware of any studies that sample transition states with a similar amount of MD sampling). That is there are many other methods that are capable of providing rate predictions, but few that can provide full atomistic unbiased trajectories of unbinding needed for resolving transition states (see Section 1.2.3). Having merely produced *a model* does not necessarily mean it is *a good model.* If we provide false positive results to drug designers there is the potential for a large amount of wasted resources being invested, which is something we should aim avoid as much as possible.

Given that sampling of transition states has remained extremely challenging and we have exposed protocols which lower this barrier to discovery we can now turn our attention to improving them. This is where the importance of having accurate rate estimates is involved, as any *in silico* prediction should be tied to experimental results. In short, improving rate estimates should improve transition state models. Beyond this however there are a number of modeling steps that go into transition state prediction which also need to be evaluated. We have utilized both MSMs and TPT to make our transition state predictions. For MSMs we can leverage the large amount of literature on them to improve these models. Furthermore, as we discussed in Chapter 6 the choice of parameters for computing TPT committor probabilities should also be addressed. Beyond this single method we should also attempt to utilize and develop multiple methods for estimating transition states, which could provide multiple sources of validation.

In addition to unique contributions for unbinding pathways and transition state modeling we have also developed a model for predicting transition state plasticity and some preliminary results testing this model (see Chapter 6). The heart of this question is critical to the success of kinetics oriented drug design. Designing against unstable transition states are fundamentally different than designing against stable bound states. As we saw in Chapter 5 a “transition state” is actually an ensemble of transition states across one or more (perhaps only vaguely distinguishable) pathways. In that chapter we described the increased potential for plasticity in transition states for a series of chemically distinct ligands. Unfortunately, due to the challenges in sampling many of the test cases in this initial study we were unable to address this concern satisfactorily. From the available data (for ligand 3 and TPPU) we were able to get some suggestive results that pathway hopping is a considerable issue to be addressed.

Regardless of whether the specific S metric introduced in 6 for TSE plasticity is ultimately predictive it provides a solid starting point for future improvements. Critically, it also introduces another avenue for comparison to experimental data through the integration of dissociation constant (and potentially solvation free-energy) measurements.

Ultimately, this thesis was successful on multiple fronts. First, the transition states of multiple ligand unbinding systems were resolved at atomic accuracy. This included a challenging drug target of clinical relevance, the results of which have directly influenced the future design of inhibitors. In terms of methods development it is a significant breakthrough as well as there are few other studies that have sampled such long timescale phenomenon without considerably more computational resources.

Secondly, we have provided strong evidence that ligand (un)binding occurs along potentially many different pathways. This is critical not only for concerns of pathway hopping (which we have also begun to address), but also methods development. As many methods *a priori* assume some form of (un)binding pathways (such as funnel shapes [41]) in order to obtain results. We find this approach fundamentally flawed when attempting to objectively study a system for which almost nothing is known about. That is predictions should be made with priors with maximal entropy and only later updated to incorporated sampled information.

Thirdly, we have made important first steps in assessing the potential for transition state plasticity to thwart kinetics oriented drug design efforts. Towards this a mathematical model that incorporates experimental data was derived, as well as a specific strategy and hypothesis for testing this model was designed and partially executed leading to promising preliminary results.

Lastly, in addition to these scientific achievements we have also provided multiple software packages designed for general usage by the field. While not of direct interest to the scientific understanding of kinetics oriented drug design, we believe that the availability of powerful tools serves to enable and accelerate scientific discoveries. Thus, of general interest to any researcher engaged in similar scientific questions.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Danzhi Huang and Amedeo Caflisch. The free energy landscape of small molecule unbinding. *PLoS Computational Biology*, 7:1–12, February 2011.

[2] K S Lee, J Y Liu, K M Wagner, S Pakhomova, H Dong, C Morisseau, S H Fu, J Yang, P Wang, A Ulu, C A Mate, L V Nguyen, S H Hwang, M L Edin, A A Mara, H Wulff, M E Newcomer, D C Zeldin, and B D Hammock. Optimized inhibitors of soluble epoxide hydrolase improve in vitro target residence time and in vivo efficacy. *Journal of Medicinal Chemistry*, 57(16):7016–7030, aug 2014.

[3] Florent Guillain and Darwin Thusius. Use of proflavine as an indicator in temperature-jump studies of the binding of a competitive inhibitor to trypsin. *Journal of the American Chemical Society*, 92(18):5534–5536, 1970.

[4] A. Shrake and J.A. Rupley. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *Journal of Molecular Biology*, 79(2):351 – 371, 1973.

[5] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann, and Mathieu Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLOS ONE*, 9(6):1–12, 06 2014.

[6] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: an open source software for exploring and manipulating networks. volume 8, pages 361–362, 2009.

[7] Robert A. Copeland, David L. Pompliano, and Thomas D. Meek. Drug-target residence time and its implications for lead optimization. *Nature Reviews Drug Discovery*, 5(9):730–739, sep 2006.

[8] Hideaki Fujitani, Yoshiaki Tanida, Masakatsu Ito, Guha Jayachandran, Christopher D. Snow, Michael R. Shirts, Eric J. Sorin, and Vijay S. Pande. Direct calculation of the binding free energies of fkbp ligands. *Journal of Chemical Physics*, 123(084108), 2005.

[9] James C. Gumbart, Benoît Roux, and Christophe Chipot. Standard binding free energies from computer simulations: What is the best strategy? *Journal of Chemical Theory and Computation*, 9(1):794–802, 2013. PMID: 23794960.

[10] Yuqing Deng and Benoît Roux. Computations of standard binding free energies with molecular dynamics simulations. *Journal of Physical Chemistry B*, 113(8):2234–2246, 2009. PMID: 19146384.

[11] Ryoji Takahashi, Víctor A. Gil, and Victor Guallar. Monte carlo free ligand diffusion with markov state model analysis and absolute binding free energy calculations. *Journal of Chemical Theory and Computation*, 10(1):282–288, 2014. PMID: 26579911.

[12] Doris A. Schuetz, Wilhelmus Egbertus Arnout de Witte, Yin Cheong Wong, Bernhard Knasmueller, Lars Richter, Daria B. Kokh, S. Kashif Sadiq, Reggie Bosma, Indira Nederpelt, Laura H. Heitman, Elena Segala, Marta Amaral, Dong Guo, Dorothee Andres, Victoria Georgi, Leigh A. Stoddart, Steve Hill, Robert M. Cooke, Chris De Graaf, Rob Leurs, Matthias Frech, Rebecca C. Wade, Elizabeth Cunera Maria de Lange, Adriaan P. IJzerman, Anke Müller-Fahrnow, and Gerhard F. Ecker. Kinetics for drug discovery: an industry-driven effort to target drug residence time. *Drug Discovery Today*, 22(6):896 – 911, 2017.

[13] Robert A Copeland. The dynamics of drug-target interactions: drug-target residence time and its impact on efficacy and safety. *Expert Opinion on Drug Discovery*, 5(4):305–310, 2010. PMID: 22823083.

[14] Robert A. Copeland. The drug-target residence time model: a 10-year retrospective. *Nature Reviews. Drug Discovery*, 15(2):87–95, feb 2016.

[15] Dong Guo, Thea Mulder-Krieger, Adriaan P IJzerman, and Laura H Heitman. Functional efficacy of adenosine a2a receptor agonists is positively correlated to their receptor residence time. *British Journal of Pharmacology*, 166(6):1846–1859, 2012.

[16] Hao Lu and Peter J Tonge. Drug-target residence time: critical information for lead optimization. *Current Opinion in Chemical Biology*, 14(4):467–474, 2010.

[17] Hao Lu, Kathleen England, Christopher am Ende, James J. Truglio, Sylvia Luckner, B. Gopal Reddy, Nicole L. Marlenee, Susan E. Knudson, Dennis L. Knudson, Richard A. Bowen, Caroline Kisker, Richard A. Slayden, and Peter J. Tonge. Slow-onset inhibition of the fabi enoyl reductase from francisella tularensis: Residence time and in vivo activity. *ACS Chemical Biology*, 4(3):221–231, 2009.

[18] B Costa, E Da Pozzo, C Giacomelli, E Barresi, S Taliani, F Da Settimo, and C Martini. Tspo ligand residence time: a new parameter to predict compound neurosteroidogenic efficacy. *Scientific Reports*, 6:18164–18164, 2016.

[19] Georges Vauquelin and Isabelle Van Liefde. Slow antagonist dissociation and long-lasting in vivo receptor protection. *Trends in Pharmacological Sciences*, 27(7):355 – 359, 2006.

[20] Johan Gabrielsson, Hugues Dolgos, Per-Göran Gillberg, Ulf Bredberg, Bert Benthem, and Göran Duker. Early integration of pharmacokinetic and dynamic reasoning is essential for optimal development of lead compounds: strategic considerations. *Drug Discovery Today*, 14(7–8):358 – 372, 2009.

[21] Donald G Truhlar, Bruce C Garrett, and Stephen J Klippenstein. Current status of transition-state theory. *The Journal of physical chemistry*, 100(31):12771–12800, 1996.

[22] John Comley. Progress in the implementation of label-free detection. Technical report, 2008.

[23] Wilma Keighley. The need for high throughput kinetics early in the drug discovery process. Technical report, Drug Discovery World, 2011.

[24] H J Motulsky and L C Mahan. The kinetics of competitive radioligand binding predicted by the law of mass action. *Molecular Pharmacology*, 25(1):1–9, 1984.

[25] Martin Kotev, Robert Soliva, and Modesto Orozco. Challenges of docking in large, flexible and promiscuous binding sites. *Bioorganic and Medicinal Chemistry*, 24(20):4961 – 4969, 2016.

[26] S. Muff and A. Caflisch. Identification of the protein folding transition state from molecular dynamics trajectories. *The Journal of Chemical Physics*, 130(12), 2009.

[27] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1):141 – 151, 1999.

[28] Michael Levitt and Arieh Warshel. Computer simulation of protein folding. *Nature*, 253:694–, feb 1975.

[29] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. *Methods in enzymology*, 383:66–93, 2004.

[30] David W Borhani and David E Shaw. The future of molecular dynamics simulations in drug discovery. *Journal of computer-aided molecular design*, 26(1):15–26, 2012.

[31] Rhiju Das. Four small puzzles that rosetta doesn't solve. *PLoS One*, 6(5):e20044, 2011.

[32] J. Andrew McCammon, Bruce R. Gelin, and Martin Karplus. Dynamics of folded proteins. *Nature*, 267:585–, jun 1977.

[33] Gordon E Moore. Cramming more components onto integrated circuits, 1965.

[34] John E Stone, David J Hardy, Ivan S Ufimtsev, and Klaus Schulten. Gpu-accelerated molecular modeling coming of age. *Journal of Molecular Graphics and Modelling*, 29(2):116–125, 2010.

[35] David E. Shaw, Martin M. Deneroff, Ron O. Dror, Jeffrey S. Kuskin, Richard H. Larson, John K. Salmon, Cliff Young, Brannon Batson, Kevin J. Bowers, Jack C. Chao, Michael P. Eastwood, Joseph Gagliardo, J. P. Grossman, C. Richard Ho, Douglas J. Ierardi, István Kolossváry, John L. Klepeis, Timothy Layman, Christine McLeavey, Mark A. Moraes, Rolf Mueller, Edward C. Priest, Yibing Shan, Jochen Spengler, Michael Theobald, Brian Towles, and Stanley C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM*, 51(7):91–97, jul 2008.

[36] David E. Shaw, J. P. Grossman, Joseph A. Bank, Brannon Batson, J. Adam Butts, Jack C. Chao, Martin M. Deneroff, Ron O. Dror, Amos Even, Christopher H. Fenton, Anthony Forte, Joseph Gagliardo, Gennette Gill, Brian Greskamp, C. Richard Ho, Douglas J. Ierardi, Lev Iserovich, Jeffrey S. Kuskin, Richard H. Larson, Timothy Layman, Li-Siang Lee, Adam K. Lerer, Chester Li, Daniel Killebrew, Kenneth M.

Mackenzie, Shark Yeuk-Hai Mok, Mark A. Moraes, Rolf Mueller, Lawrence J. Nociolo, Jon L. Peticolas, Terry Quan, Daniel Ramot, John K. Salmon, Daniele P. Scarpazza, U. Ben Schafer, Naseer Siddique, Christopher W. Snyder, Jochen Spengler, Ping Tak Peter Tang, Michael Theobald, Horia Toma, Brian Towles, Benjamin Vitale, Stanley C. Wang, and Cliff Young. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '14, pages 41–53, Piscataway, NJ, USA, 2014. IEEE Press.

[37] Huei-Jiun Li, Cheng-Tsung Lai, Pan Pan, Weixuan Yu, Nina Liu, Gopal R. Bommineni, Miguel Garcia-Diaz, Carlos Simmerling, and Peter J. Tonge. A structural and energetic model for the slow-onset inhibition of the mycobacterium tuberculosis enoyl-acp reductase inha. *ACS Chemical Biology*, 9(4):986–993, 2014. PMID: 24527857.

[38] Min Xu, Amedeo Caflisch, and Peter Hamm. Protein structural memory influences ligand binding mode(s) and unbinding rates. *Journal of Chemical Theory and Computation*, 12(3):1393–1399, 2016. PMID: 26799675.

[39] Alex Dickson and Samuel D. Lotz. Ligand release pathways obtained with wexplore: Residence times and mechanisms. *Journal of Physical Chemistry B*, 120(24):5377–5385, 2016. PMID: 27231969.

[40] Albert C. Pan, Huafeng Xu, Timothy Palpant, and David E. Shaw. Quantitative characterization of the binding and unbinding of millimolar drug fragments with molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 13(7):3372–3377, 2017. PMID: 28582625.

[41] Vittorio Limongelli, Massimiliano Bonomi, and Michele Parrinello. Funnel metadynamics as accurate binding free-energy method. *Proceedings of the National academy of Sciences of the United States of America*, 110(16):6358–6363, 2013.

[42] Ivan Teo, Christopher G. Mayne, Klaus Schulten, and Tony Lelièvre. Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time. *Journal of Chemical Theory and Computation*, 12(6):2983–2989, 2016. PMID: 27159059.

[43] Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proceedings of the National academy of Sciences of the United States of America*, 108(25):10184–10189, 2011.

[44] Nuria Plattner and Frank Noe. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and markov models. *Nature Communications*, 6, jul 2015.

[45] S. Doerr and G. De Fabritiis. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. *Journal of Chemical Theory and Computation*, 10(5):2064–2069, 2014. PMID: 26580533.

[46] Pratyush Tiwary, Vittorio Limongelli, Matteo Salvalaglio, and Michele Parrinello. Kinetics of protein–ligand unbinding: Predicting pathways, rates, and rate-limiting steps. *Proceedings of the National academy of Sciences of the United States of America*, 112(5):E386–E391, 2015.

[47] Alex Dickson and Samuel D. Lotz. Multiple ligand unbinding pathways and ligand-induced destabilization revealed by wexplore. *Biophysical Journal*, 112(4):620–629, February 2017.

[48] Yibing Shan, Eric T. Kim, Michael P. Eastwood, Ron O. Dror, Markus A. Seeliger, and David E. Shaw. How does a drug molecule find its target binding site? *Journal of the American Chemical Society*, 133(24):9181–9183, 2011. PMID: 21545110.

[49] Dong Guo, Albert C. Pan, Ron O. Dror, Tamara Mocking, Rongfang Liu, Laura H. Heitman, David E. Shaw, and Adriaan P. IJzerman. Molecular basis of ligand dissociation from the adenosine a2a receptor. *Molecular Pharmacology*, 89(5):485–491, 2016.

[50] Pratyush Tiwary, Jagannath Mondal, and B. J. Berne. How and when does an anticancer drug leave its binding site? *Science Advances*, 3(5), 2017.

[51] Rodrigo Casasnovas, Vittorio Limongelli, Pratyush Tiwary, Paolo Carloni, and Michele Parrinello. Unbinding kinetics of a p38 map kinase type ii inhibitor from metadynamics simulations. *Journal of the American Chemical Society*, 139:4780–4788, 2017. PMID: 28290199.

[52] Huiyong Sun, Sheng Tian, Shunye Zhou, Youyong Li, Dan Li, Lei Xu, Mingyun Shen, Peichan Pan, and Tingjun Hou. Revealing the favorable dissociation pathway of type ii kinase inhibitors via enhanced sampling simulations and two-end-state calculations. *Scientific Reports*, 5(8457), 2015.

[53] Cameron F. Abrams and Eric Vanden-Eijnden. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proceedings of the National academy of Sciences of the United States of America*, 107(11):4961–4966, 2010.

[54] Luca Maragliano and Eric Vanden-Eijnden. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical Physics Letters*, 426(1):168 – 175, 2006.

[55] Luca Mollica, Isabelle Theret, Mathias Antoine, Francoise Perron-Sierra, Yves Charton, Jean-Marie Fourquez, Michel Wierzbicki, Jean A. Boutin, Gilles Ferry, Sergio Decherchi, Giovanni Bottegoni, Pierre Ducrot, and Andrea Cavalli. Molecular dynamics simulations and kinetic measurements to estimate and predict protein? ligand residence times luca mollica? *Journal of Medicinal Chemistry*, 59(15):7167–7176, 2016.

[56] Luca Mollica, Sergio Decherchi, Syeda Rehana Zia, Roberto Gaspari, Andrea Cavalli, and Walter Rocchia. Kinetics of protein-ligand unbinding via smoothed potential molecular dynamics simulations. *Scientific Reports*, 5(11539), June 2015.

[57] Anna Maria Capelli and Gabriele Costantino. Unbinding pathways of vegfr2 inhibitors revealed by steered molecular dynamics. *Journal of Chemical Information and Modeling*, 54(11):3124–3136, 2014. PMID: 25299731.

[58] Cosma Shalizi. Advanced data analysis from an elementary point of view, 2013.

[59] Ulrich H.E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1):140 – 150, 1997.

[60] Pratyush Tiwary and Michele Parrinello. From metadynamics to dynamics. *Phys. Rev. Lett.*, 111(23):230602, dec 2013.

[61] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(5):826–843, 2011.

[62] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187 – 199, 1977.

[63] Daniel M. Zuckerman and Lillian T. Chong. Weighted ensemble simulation: Review of methodology, applications, and software. *Annual Review of Biophysics*, 46(1):43–57, 2017. PMID: 28301772.

[64] Lawrence R Pratt. A statistical method for identifying transition states in high dimensional problems. *The Journal of chemical physics*, 85(9):5045–5048, 1986.

[65] PeteráG Bolhuis et al. Sampling ensembles of deterministic transition pathways. *Faraday Discussions*, 110:421–436, 1998.

[66] Titus S. van Erp, Daniele Moroni, and Peter G. Bolhuis. A novel path sampling method for the calculation of rate constants. *The Journal of Chemical Physics*, 118(17):7762–7774, 2003.

[67] Rosalind J. Allen, Patrick B. Warren, and Pieter Rein ten Wolde. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.*, 94(1):018104, jan 2005.

[68] Frédéric Cérou and Arnaud Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007.

[69] G A Huber and S Kim. Weighted-ensemble brownian dynamics simulations for protein association reactions. *Biophysical Journal*, 70(1):97–110, jan 1996.

[70] David Aristoff. Analysis and optimization of weighted ensemble sampling. *ESAIM: M2AN*, 2017.

[71] J L Adelman and M Grabe. Simulating rare events using a weighted ensemble-based string method. *J Chem Phys*, 138(4):044105–044105, jan 2013.

[72] Atipat Rojnuckarin, Dennis R. Livesay, and Shankar Subramaniam. Bimolecular reaction simulation using weighted ensemble brownian dynamics and the university of houston brownian dynamics program. *Biophysical Journal*, 79(2):686 – 693, 2000.

[73] Bin W. Zhang, David Jasnow, and Daniel M. Zuckerman. Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin. *Proceedings of the National Academy of Sciences*, 104(46):18043–18048, 2007.

[74] Ali S. Saglam and Lillian T. Chong. Highly efficient computation of the basal kon using direct simulation of protein–protein association with flexible molecular models. *The Journal of Physical Chemistry B*, 120(1):117–122, 2016. PMID: 26673903.

[75] Alex Dickson, Anthony M. Mustoe, Loic Salmon, and Charles L. Brooks III. Efficient in silico exploration of rna interhelical conformations using euler angles and wexplore. *Nucleic Acids Research*, 42(19), 2014.

[76] Badi' Abdul-Wahid, Haoyun Feng, Dinesh Rajan, Ronan Costaouec, Eric Darve, Douglas Thain, and Jesús A. Izaguirre. Awe-wq: Fast-forwarding molecular dynamics using the accelerated weighted ensemble. *Journal of Chemical Information and Modeling*, 54(10):3033–3043, 2014. PMID: 25207854.

[77] Joshua L. Adelman and Michael Grabe. Simulating current–voltage relationships for a narrow ion channel using the weighted ensemble method. *Journal of Chemical Theory and Computation*, 11(4):1907–1918, 2015.

[78] Matthew C. Zwier, Adam J. Pratt, Joshua L. Adelman, Joseph W. Kaus, Daniel M. Zuckerman, and Lillian T. Chong. Efficient atomistic simulation of pathways and calculation of rate constants for a protein–peptide binding process: Application to the mdm2 protein and an intrinsically disordered p53 peptide. *The Journal of Physical Chemistry Letters*, 7(17):3440–3445, 2016. PMID: 27532687.

[79] Haoyun Feng, Ronan Costaouec, Eric Darve, and Jesús A. Izaguirre. A comparison of weighted ensemble and markov state model methodologies. *The Journal of Chemical Physics*, 142(21):214113–, jun 2015.

[80] Tom Dixon, Samuel D. Lotz, and Alex Dickson. Predicting ligand binding affinity using on- and off-rates for the sampl6 sampling challenge. *Journal of Computer-Aided Molecular Design*, 32(10):1001–1012, oct 2018.

[81] Rosalind J Allen, Daan Frenkel, and Pieter Rein ten Wolde. Simulating rare events in equilibrium or nonequilibrium stochastic systems. *The Journal of chemical physics*, 124(2):024102, 2006.

[82] Marco J Morelli, Pieter Rein ten Wolde, and Rosalind J Allen. Dna looping provides stability and robustness to the bacteriophage λ switch. *Proceedings of the National Academy of Sciences*, 106(20):8101–8106, 2009.

[83] Rory M. Donovan, Jose-Juan Tapia, Devin P. Sullivan, James R. Faeder, Robert F. Murphy, Markus Dittrich, and Daniel M. Zuckerman. Unbiased rare event sampling in spatial stochastic systems biology models using a weighted ensemble of trajectories. *PLOS Computational Biology*, 12(2):1–25, 02 2016.

[84] Rory M Donovan, Andrew J Sedgewick, James R Faeder, and Daniel M Zuckerman. Efficient stochastic simulation of chemical kinetics networks using a weighted ensemble of trajectories. *The Journal of chemical physics*, 139(11):09B642_1, 2013.

[85] J Tse Margaret, Brian K Chu, Mahua Roy, and Elizabeth L Read. Dna-binding kinetics determines the mechanism of noise-induced switching in gene networks. *Biophysical journal*, 109(8):1746–1757, 2015.

[86] Bernie J Daigle Jr, Min K Roh, Dan T Gillespie, and Linda R Petzold. Automated estimation of rare event probabilities in biochemical systems. *The Journal of chemical physics*, 134(4):01B628, 2011.

[87] Min K Roh, Bernie J Daigle Jr, Dan T Gillespie, and Linda R Petzold. State-dependent doubly weighted stochastic simulation algorithm for automatic characterization of stochastic biochemical rare events. *The Journal of chemical physics*, 135(23):234108, 2011.

[88] J Tse Margaret, Brian K Chu, Cameron P Gallivan, and Elizabeth L Read. Rare-event sampling of epigenetic landscapes and phenotype transitions. *PLoS computational biology*, 14(8):e1006336, 2018.

[89] Manuel Villén-Altamirano and José Villén-Altamirano. Restart: a straightforward method for fast simulation of rare events. In *Proceedings of Winter Simulation Conference*, pages 282–289. IEEE, 1994.

[90] Jérôme Morio and Mathieu Balesdent. *Estimation of rare event probabilities in complex aerospace and other systems: a practical approach.* Woodhead Publishing, 2015.

[91] Bin W. Zhang, David Jasnow, and Daniel M. Zuckerman. The “weighted ensemble” path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *The Journal of Chemical Physics*, 132(5), 2010.

[92] A Dickson and C L Brooks. Wexplore: hierarchical exploration of high-dimensional spaces using the weighted ensemble algorithm. *Journal of Physical Chemistry B*, 118(13):3532–3542, apr 2014.

[93] Nazanin Donyapour, Nicole M Roussey, and Alex Dickson. Revo: Resampling of ensembles by variation optimization. *The Journal of Chemical Physics*, 150(24):244112, 2019.

[94] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13, 2009.

[95] Riccardo Capelli, Paolo Carloni, and Michele Parrinello. Exhaustive search of ligand binding pathways via volume-based metadynamics. *The journal of physical chemistry letters*, 10(12):3495–3499, 2019.

[96] Matthew Skala. Measuring the difficulty of distance-based indexing. In Mariano Consens and Gonzalo Navarro, editors, *String Processing and Information Retrieval*, pages 103–114, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[97] Hanan Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[98] Yanay Ofran and Burkhard Rost. Analysing six types of protein–protein interfaces. *Journal of Molecular Biology*, 325(2):377 – 387, 2003.

[99] Wei Gao, Mohammad Reza Farahani, Muhammad Imran, and M. R. Rajesh Kanna. Distance-based topological polynomials and indices of friendship graphs. *SpringerPlus*, 5(1), sep 2016.

[100] Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM computing surveys (CSUR)*, 33(3):273–321, 2001.

[101] Rui Mao, Willard L. Miranker, and Daniel P. Miranker. Pivot selection: Dimension reduction for distance-based indexing. *Journal of Discrete Algorithms*, 13:32 – 46, 2012. Best Papers from the 3rd International Conference on Similarity Search and Applications (SISAP 2010).

[102] Samuel D Lotz and Alex Dickson. Unbiased molecular dynamics of 11 min timescale drug unbinding reveals transition state stabilizing interactions. *Journal of the American Chemical Society*, 140(2):618–628, January 2018.

[103] Alex Dickson. Mapping the ligand binding landscape. *Biophysical Journal*, 115(9):1707 – 1719, 2018.

[104] Jeremy Copperman and Daniel Zuckerman. Accelerated estimation of long-timescale kinetics by combining weighted ensemble simulation with markov model "microstates" using non-markovian theory, 2019. cite arxiv:1903.04673.

[105] Alex Dickson, Aryeh Warmflash, and Aaron R. Dinner. Separating forward and backward pathways in nonequilibrium umbrella sampling. *Journal of Chemical Physics*, 131(154104), 2009.

[106] Eric Vanden-Eijnden and Maddalena Venturoli. Exact rate calculations by trajectory parallelization and tilting. *Journal of Chemical Physics*, 131(044120), 2009.

[107] Terrell L Hill. *Free Energy Transduction and Biochemical Cycle Kinetics*. Springer, 1989.

[108] Samuel D. Lotz and Alex Dickson. Wepy: A flexible software framework for simulating rare events with weighted ensemble resampling. *ACS Omega*, 5(49):31608–31623, 2020.

[109] Numba Contributors. Numba. https://github.com/numba/numba, 2020. [Online; accessed 2020-03-17].

[110] Dask Contributors. dask. https://github.com/dask/dask, 2020. Accessed online 2020-03-17.

[111] P Eastman, M S Friedrichs, J D Chodera, R J Radmer, C M Bruns, J P Ku, K A Beauchamp, T J Lane, L P Wang, D Shukla, T Tye, M Houston, T Stich, C Klein, M R Shirts, and V S Pande. Openmm 4: A reusable, extensible, hardware independent library for high performance molecular simulation. *Journal of Chemical Theory and Computation*, 9(1):461–469, jan 2013.

[112] Mike Folk, Albert Cheng, and Kim Yates. Hdf5: A file format and i/o library for high performance computing applications. In *Proceedings of Supercomputing*, volume 99, pages 5–33, 1999.

[113] Matthew C. Zwier, Joshua L. Adelman, Joseph W. Kaus, Adam J. Pratt, Kim F. Wong, Nicholas B. Rego, Ernesto Suárez, Steven Lettieri, David W. Wang, Michael Grabe, Daniel M. Zuckerman, and Lillian T. Chong. Westpa: An interoperable, highly scalable software package for weighted ensemble simulation and analysis. *Journal of Chemical Theory and Computation*, 11(2):800–809, 2015.

[114] Ernesto Suárez, Steven Lettieri, Matthew C. Zwier, Carsen A. Stringer, Sundar Raman Subramanian, Lillian T. Chong, and Daniel M. Zuckerman. Simultaneous computation of dynamical and equilibrium information using a weighted ensemble of trajectories. *Journal of Chemical Theory and Computation*, 10(7):2658–2667, 2014. PMID: 25246856.

[115] Ronan Costaouec, Haoyun Feng, Jesús Izaguirre, and Eric Darve. Analysis of the accelerated weighted ensemble methodology. *Discrete and Continuous Dynamical Systems*, pages 171–181, 2013.

[116] Douglas L. Theobald. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Crystallographica Section A Foundations of Crystallography*, 61(4):478–480, jun 2005.

[117] Scipy Contributors. scipy. https://github.com/scipy/scipy, 2020.

[118] Robert T. McGibbon, Kyle A. Beauchamp, Matthew P. Harrigan, Christoph Klein, Jason M. Swails, Carlos X. Hernández, Christian R. Schwantes, Lee-Ping Wang, Thomas J. Lane, and Vijay S. Pande. Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. *Biophysical Journal*, 109(8):1528 – 1532, 2015.

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

[120] Ahmet Bakan, Lidio M. Meireles, and Ivet Bahar. Prody: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 2011.

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

[122] W A Baase, L Liu, D E Tronrud, and B W Matthews. Lessons from the lysozyme of phage t4. *Protein Sci*, 19(4):631–641, apr 2010.

[123] Yong Wang, Elena Papaleo, and Kresten Lindorff-Larsen. Mapping transiently formed and sparsely populated conformations on a complex energy landscape. *Elife*, 5:e17505, 2016.

[124] Ariane Nunes-Alves, Daniel M. Zuckerman, and Guilherme Menegon Arantes. Escape of a small molecule from inside t4 lysozyme by multiple pathways. *Biophysical Journal*, 114(5):1058 – 1066, 2018.

[125] Jamie M Schiffer, Victoria A Feher, Robert D Malmstrom, Roxana Sida, and Rommie E Amaro. Capturing invisible motions in the transition from ground to rare excited states of t4 lysozyme l99a. *Biophysical journal*, 111(8):1631–1640, 2016.

[126] Yinglong Miao, Victoria A Feher, and J Andrew McCammon. Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *Journal of chemical theory and computation*, 11(8):3584–3595, 2015.

[127] Tom Dixon, Arzu Uyar, Shelagh Ferguson-Miller, and Alex Dickson. Membrane-mediated ligand unbinding of the pk-11195 ligand from the translocator protein (tspo). *bioRxiv*, 2020.

[128] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs: fast, flexible, and free. *Journal of computational chemistry*, 26(16):1701–1718, 2005.

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

[130] David A Pearlman, David A Case, James W Caldwell, Wilson S Ross, Thomas E Cheatham III, Steve DeBolt, David Ferguson, George Seibel, and Peter Kollman. Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications*, 91(1-3):1–41, 1995.

[131] Kevin J Bowers, David E Chow, Huafeng Xu, Ron O Dror, Michael P Eastwood, Brent A Gregersen, John L Klepeis, Istvan Kolossvary, Mark A Moraes, and Federico D

Sacerdoti. Scalable algorithms for molecular dynamics simulations on commodity clusters. In *SC'06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing*, pages 43–43. IEEE, 2006.

[132] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[133] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, and Cory Hargus. The atomic simulation environment – a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017.

[134] Tiejun Cheng, Qingliang Li, Zhigang Zhou, Yanli Wang, and Stephen H. Bryant. Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS Journal*, 14(1):133–141, 2012.

[135] A Lavecchia and C Di Giovanni. Virtual screening strategies in drug discovery: A critical review. *Current Medicinal Chemistry*, 20:2839–2860, 2013.

[136] Ragul Gowthaman, Sergey Lyskov, and John Karanicolas. Darc 2.0: Improved docking and virtual screening at protein interaction sites. *PLoS One*, 10(7):e0131612, 2015.

[137] John D Chodera, David L Mobley, Michael R Shirts, Richard W Dixon, Kim Branson, and Vijay S Pande. Alchemical free energy methods for drug discovery: progress and challenges. *Current Opinion in Structural Biology*, 21(2):150 – 160, 2011.

[138] Joseph W Kaus, Edward Harder, Teng Lin, Robert Abel, J Andrew Mccammon, and Lingle Wang. How to deal with multiple binding poses in alchemical relative protein – ligand binding free energy calculations. *Journal of Chemical Theory and Computation*, 11:2670–2679, 2015.

[139] Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K Dahlgren, Jeremy Greenwood, Donna L Romero, Craig Masse, Jennifer L Knight, Thomas Steinbrecher, Thijs Beuming, Wolfgang Damm, Ed Harder, Woody Sherman, Mark Brewer, Ron Wester, Mark Murcko, Leah Frye, Ramy Farid, Teng Lin, David L Mobley, William L Jorgensen, Bruce J Berne, Richard A Friesner, and Robert Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137:2695–2703, 2015.

[140] Raffaele Curcio, Amedeo Caflisch, and Emanuele Paci. Change of the unbinding mechanism upon a mutation: A molecular dynamics study of an antibody–hapten complex. *Protein Science*, 14(10):2499–2514, 2005.

[141] A. Aird, J. Wrachtrup, K. Schulten, and C. Tietz. Possible pathway for ubiquinone shuttling in rhodospirillum rubrum revealed by molecular dynamics simulation. *Biophysical Journal*, 92(1):23 – 33, 2007.

[142] Francesco Colizzi, Remo Perozzo, Leonardo Scapozza, Maurizio Recanatini, and Andrea Cavalli. Single-molecule pulling simulations can discern active from inactive enzyme inhibitors. *Journal of the American Chemical Society*, 132(21):7361–7371, 2010. PMID: 20462212.

[143] Laura J Kingsley and Markus A Lill. Including ligand-induced protein flexibility into protein tunnel prediction. *Journal of Computational Chemistry*, 35(24):1748–56, 2014.

[144] F Pietrucci, F Marinelli, P Carloni, and A Laio. Substrate binding mechanism of hiv-1 protease from explicit-solvent atomistic simulations. *Journal of the American Chemical Society*, 131(33):11811–11818, 2009.

[145] Vittorio Limongelli, Massimiliano Bonomi, Luciana Marinelli, Francesco Luigi Gervasio, Andrea Cavalli, Ettore Novellino, and Michele Parrinello. Molecular basis of cyclooxygenase enzymes (coxs) selective inhibition. *Proceedings of the National Academy of Sciences*, 107(12):5411–5416, 2010.

[146] Matthew W. Harding, Andrzej Galat, David E. Uehling, and Stuart L. Schreiber. A receptor for the immuno-suppressant fk506 is a cis-trans peptidyl-prolyl isomerase. *Nature*, 341:758–760, October 1989.

[147] John J. Siekierka, Shirley H. Y. Hung, Martin Poe, C. Shirley Lin, and Nolan H. Sigal. A cytosolic binding protein for the immunosuppressant fk506 has peptidyl-prolyl isomerase activity but is distinct from cyclophilin. *Nature*, 341(6244):755–757, oct 1989.

[148] Gregory D. Van Duyne, Robert F. Standaert, P.Andrew Karplus, Stuart L. Schreiber, and Jon Clardy. Atomic structures of the human immunophilin fkbp-12 complexes with fk506 and rapamycin. *Journal of Molecular Biology*, 229(1):105 – 124, 1993.

[149] J P Griffith, J L Kim, E E Kim, M D Sintchak, J A Thomson, M J Fitzgibbon, M A Fleming, P R Caron, K Hsiao, and M A Navia. X-ray structure of calcineurin inhibited by the immunophilin-immunosuppressant fkbp12-fk506 complex. *Cell*, 82(3):507–522, aug 1995.

[150] Charles R. Kissinger, Hans E. Parge, Daniel R. Knighton, Cristina T. Lewis, Laura A. Pelletier, Anna Tempczyk, Vincent J. Kalish, Kathleen D. Tucker, Richard E. Showalter, Ellen W. Moomaw, Louis N. Gastinel, Noriyuki Habuka, Xinghai Chen, Fausto Maldonado, John E. Barker, Russell Bacquet, and J. Ernest Villafranca. Crystal structures of human calcineurin and the human fkbp12-fk506-calcineurin complex. *Nature*, 378(6557):641–644, dec 1995.

[151] Jungwon Choi, Jie Chen, Stuart L. Schreiber, and Jon Clardy. Structure of the fkbp12-rapamycin complex interacting with binding domain of human frap. *Science*, 273(5272):239–242, July 1996. DOI: 10.1126/science.273.5272.239.

[152] Andrzej Galat. Functional diversity and pharmacological profiles of the fkbps and their complexes with small natural ligands. *Cellular and Molecular Life Sciences*, 70:3243–3275, 2013.

[153] Peter Burkhard, Paul Taylor, and Malcolm D Walkinshaw. X-ray structures of small ligand-fkbp complexes provide an estimate for hydrophobic interaction energies1. *Journal of Molecular Biology*, 295(4):953 – 962, 2000.

[154] K. Vanommeslaeghe and A. D. MacKerell. Automation of the charmm general force field (cgenff) i: Bond perception and atom typing. *Journal of Chemical Information and Modeling*, 52(12):3144–3154, 2012.

[155] K. Vanommeslaeghe, E. Prabhu Raman, and A. D. MacKerell. Automation of the charmm general force field (cgenff) ii: Assignment of bonded parameters and partial atomic charges. *Journal of Chemical Information and Modeling*, 52(12):3155–3168, 2012.

[156] Alex Dickson, Mark Maienschein-Cline, Allison Tovo-Dwyer, Jeff R. Hammond, and Aaron R. Dinner. Flow-dependent unfolding and refolding of an rna by nonequilibrium umbrella sampling. *Journal of Chemical Theory and Computation*, 7(9):2710–2720, 2011. PMID: 26605464.

[157] Kyle A. Beauchamp, Gregory R. Bowman, Thomas J. Lane, Lutz Maibaum, Imran S. Haque, and Vijay S. Pande. Msmbuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. *Journal of Chemical Theory and Computation*, 7(10):3412–3419, 2011. PMID: 22125474.

[158] Martin Senne, Benjamin Trendelkamp-Schroer, Antonia S.J.S. Mey, Christof Schütte, and Frank Noé. Emma: A software package for markov model building and analysis. *Journal of Chemical Theory and Computation*, 8(7):2223–2238, 2012. PMID: 26588955.

[159] Sebastian Salentin, Sven Schreiber, V. Joachim Haupt, Melissa F. Adasme, and Michael Schroeder. Plip: fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, 43(W1):W443–W447, 2015.

[160] Roderick E Hubbard and Muhammad Kamran Haider. Hydrogen bonds in proteins: Role and strength. In *Encyclopedia of Life Science*. John Wiley & Sons, Ltd., Chichester, 2010.

[161] Justin P. Gallivan and Dennis A. Dougherty. Cation-pi interactions in structural biology. *Proceedings of the National Academy of Science*, 96(17):9459–9464, August 1999.

[162] Lianqing Zheng, Mengen Chen, and Wei Yang. Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. *Proceedings of the National Academy of Sciences*, 105(51):20227–20232, 2008.

[163] Xiongwu Wu and Bernard R. Brooks. Self-guided langevin dynamics simulation method. *Chemical Physics Letters*, 381(3):512 – 518, 2003.

[164] Ting Wang and Yong Duan. Ligand entry and exit pathways in the beta2-adrenergic receptor. *Journal of Molecular Biology*, 392(4):1102–15, 2009.

[165] Peter Carlsson, Sofia Burendahl, and Lennart Nilsson. Unbinding of retinoic acid from the retinoic acid receptor by random expulsion molecular dynamics. *Biophysical Journal*, 91(9):3151–61, 2006.

[166] Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D. Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of Chemical Physics*, 134(17), 2011.

[167] Francesco Rao and Amedeo Caflisch. The protein folding network. *Journal of Molecular Biology*, 342(1):299 – 306, 2004.

[168] Alex Dickson and Charles L. Brooks. Native states of fast-folding proteins are kinetic traps. *Journal of the American Chemical Society*, 135(12):4729–4734, 2013. PMID: 23458553.

[169] Albert C. Pan, David W. Borhani, Ron O. Dror, and David E. Shaw. Molecular determinants of drug–receptor binding kinetics. *Drug Discovery Today*, 18(13–14):667 – 673, 2013.

[170] Ning Yin, Jianfeng Pei, and Luhua Lai. A comprehensive analysis of the influence of drug binding kinetics on drug action at molecular and systems levels. *Mol. BioSyst.*, 9(6):1381–1389, 2013.

[171] John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*, 25:135 – 144, 2014.

[172] Shaoyong Lu, Shuai Li, and Jian Zhang. Harnessing allostery : A novel approach to drug discovery. *Medicinal Research Reviews*, 34(6):1242–1285, 2014.

[173] Ran Dai, Todd W. Geders, Feng Liu, Sae Woong Park, Dirk Schnappinger, Courtney C. Aldrich, and Barry C. Finzel. Fragment-based exploration of binding site flexibility in mycobacterium tuberculosis bioa. *Journal of Medicinal Chemistry*, 58(13):5208–5217, 2015.

[174] Elena N. Laricheva, Garrett B. Goh, Alex Dickson, and Charles L. Brooks. ph-dependent transient conformational states control optical properties in cyan fluorescent protein. *Journal of the American Chemical Society*, 137(8):2892–2900, 2015. PMID: 25647152.

[175] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[176] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$

and $\chi 2$ dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012. PMID: 23341755.

[177] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, 2010.

[178] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016.

[179] Samuel Lotz. Multi-atom structure/selection toolkit with interaction capabilities (mastic) v0.2.1-beta release. 2017.

[180] Yannick Hold-Geoffroy, Olivier Gagnon, and Marc Parizeau. Once you scoop, no need to fork. In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, page 60. ACM, 2014.

[181] Weinan E. and Eric Vanden-Eijnden. Towards a theory of transition paths. *Journal of Statistical Physics*, 123(3):503–523, 2006.

[182] Philipp Metzner, Christof Schütte, and Eric Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Modeling and Simulation*, 7(3):1192–1219, 2009.

[183] Hang Chen, Ying Zhang, Liang Li, and Ju-Guang Han. Probing ligand-binding modes and binding mechanisms of benzoxazole-based amide inhibitors with soluble epoxide hydrolase by molecular docking and molecular dynamics simulation. *Journal of Physical Chemistry B*, 116(34):10219–10233, 2012. PMID: 22857012.

[184] Kin Sing Stephen Lee, Niel M. Henriksen, Connie J. Ng, Jun Yang, Weitao Jia, Christophe Morisseau, Armann Andaya, Michael K. Gilson, and Bruce D. Hammock. Probing the orientation of inhibitor and epoxy-eicosatrienoic acid binding in the active site of soluble epoxide hydrolase. *Archives of Biochemistry and Biophysics*, 613:1 – 11, 2017.

[185] Li Xing, Joseph J. McDonald, Steve A. Kolodziej, Ravi G. Kurumbail, Jennifer M. Williams, Chad J. Warren, Janet M. O'Neal, Jill E. Skepner, and Steven L. Roberds. Discovery of potent inhibitors of soluble epoxide hydrolase by combinatorial library design and structure-based virtual screening. *Journal of Medicinal Chemistry*, 54(5):1211–1222, 2011. PMID: 21302953.

[186] Loïc Salmon, Logan S. Ahlstrom, Scott Horowitz, Alex Dickson, Charles L. Brooks, and James C. A. Bardwell. Capturing a dynamic chaperone–substrate interaction using nmr-informed molecular modeling. *Journal of the American Chemical Society*, 138(31):9826–9839, 2016. PMID: 27415450.

[187] Nina Singhal, Christopher D. Snow, and Vijay S. Pande. Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *Journal of Chemical Physics*, 121(1):415–425, 2004.

[188] Lauren A. Spagnuolo, Sandra Eltschkner, Weixuan Yu, Fereidoon Daryaee, Shabnam Davoodi, Susan E. Knudson, Eleanor K. H. Allen, Jonathan Merino, Annica Pschibul, Ben Moree, Neil Thivalapill, James J. Truglio, Joshua Salafsky, Richard A. Slayden, Caroline Kisker, and Peter J. Tonge. Evaluating the contribution of transition-state destabilization to changes in the residence time of triazole-based inha inhibitors. *Journal of the American Chemical Society*, 139(9):3417–3429, 2017. PMID: 28151657.

[189] Andrew Chang, Johannes Schiebel, Weixuan Yu, Gopal R. Bommineni, Pan Pan, Michael V. Baxter, Avinash Khanna, Christoph A. Sotriffer, Caroline Kisker, and Peter J. Tonge. Rational optimization of drug-target residence time: Insights from inhibitor binding to the staphylococcus aureus fabi enzyme–product complex. *Biochemistry*, 52(24):4217–4228, 2013. PMID: 23697754.

[190] James S. Butler and Stewart N. Loh. Kinetic partitioning during folding of the p53 dna binding domain. *Journal of Molecular Biology*, 350(5):906 – 918, 2005.

[191] Pratyush Tiwary and B. J. Berne. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proceedings of the National academy of Sciences of the United States of America*, 113(11):2839–2844, 2016.

[192] Kin Sing Stephen Lee, Christophe Morisseau, Jun Yang, Peng Wang, Sung Hee Hwang, and Bruce D. Hammock. Förster resonance energy transfer competitive displacement assay for human soluble epoxide hydrolase. *Analytical Biochemistry*, 434(2):259 – 268, 2013.

[193] Peter Csizmadia. Marvinsketch and marvinview: molecule applets for the world wide web. 1999.

[194] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):1–14, 2011.

[195] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013. PMID: 23379370.

[196] Guido van Rossum and Fred L. Drake. *The Python Language Reference Manual.* Network Theory Ltd., 2011.