UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Science
School of Chemistry and Chemical Engineering
Essex Group

Development and Application of Grand Canonical Nonequilibrium Candidate Monte Carlo for in silico Prediction of Fragment Binding Sites, Modes, and Affinities

by

William George Poole
MChem
ORCiD: 0009-0003-2441-8794

A thesis for the degree of
Doctor of Philosophy

June 2025

University of Southampton

Abstract

Faculty of Engineering and Physical Science
School of Chemistry and Chemical Engineering

Doctor of Philosophy

Development and Application of Grand Canonical Nonequilibrium Candidate Monte Carlo for in silico Prediction of Fragment Binding Sites, Modes, and Affinities

by William George Poole

Structure and fragment-based drug design are increasingly popular approaches to drug discovery. Computational tools have become integral to these campaigns and provide a route to library design, virtual screening, property prediction, identifying putative binding sites, elucidating binding geometries, and predicting accurate binding affinities. This thesis discusses various molecular simulation methods and assesses their applicability to these drug discovery regimes.

Molecular dynamics-based simulations are a useful tool in computer-aided drug design but are often limited by sampling issues related to the simulation timescales obtainable. Here, we develop, implement, validate, and test the application of grand canonical nonequilibrium candidate Monte Carlo (GCNCMC) to accurately predict the binding sites, modes, and affinities of fragment-like molecules. To this end, we develop the Python module, *grandlig*. GCNCMC simulations can accurately predict the location of small molecules in protein-ligand systems by attempting the insertion and deletion of molecules to, or from, a region of interest; each proposed move is subject to a rigorous acceptance test based on the thermodynamic properties of the system.

We first set the scene and highlight the limitations of basic MD simulations by applying a variety of methods to the ERK2 protein. The theory and development of ligand-based GCNCMC is then presented with a rigorous validation of the method. The subsequent chapters then present various ways in which GCNCMC can be used to enhance the drug discovery pipeline by applying the method to two protein-ligand systems, T4L99A and MUP1. We demonstrate the ability of fragment-based GCNCMC to rapidly and reliably find experimental fragment binding sites, show that the method can accurately sample multiple fragment binding modes without any prior knowledge of their existence, and finally demonstrate the method's ability as a free energy estimator.

We present two novel applications of GCNCMC; the integration of GCNCMC into mixed solvent MD, a popular method for binding site identification, and as a fragment screening tool. In both cases, we observe promising results and outline steps for the future which could make this method a powerful tool in the computational-aided drug design arsenal.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: Poole, W.; Samways, M.; Branduardi, D.; Taylor, R.; Verdonk, M.; Essex, J. Accelerating Fragment Based Drug Discovery Using Grand Canonical Nonequilibrium Candidate Monte Carlo. ChemRxiv September 30, 2024. https://doi.org/10.26434/chemrxiv-2024-q9l5z

Signed:..... Date:.....

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Jonathan Essex for his support throughout all of these years. Jon has provided me with everything I needed to succeed in this PhD, and his advice both academically and pastorally has always been so useful. Finally, his guidance has helped me in my own personal and professional development more than he knows.

I would like to thank my industrial collaborators at Astex Pharmaceuticals, particularly Davide, Pavel, and Marcel, for both funding and for the extremely useful discussions throughout. Being able to discuss scientific problems with those who have a lot of experience and an industry perspective is invaluable. I would also like to thank them for the fun and useful experience I gained while on placement at Astex.

In a similar vein, I would like to thank those at UCB Pharma. I was lucky enough to engage in a second work placement at UCB, and I would like to thank them for the opportunity to do something a bit different. In that respect, I would like to particularly thank Jon Shearer for looking after me during this time. I also thank Marley and Rich, who have provided support and useful discussions throughout my PhD and placement. I would like to thank Marley even more for his support and teaching throughout the past 6 years from when I first joined the group. He has never shied away from a panicked late evening meltdown about GCMC.

I will always look back on my time in the Essex group very fondly. This is in part due to the amazing group members, many of whom have become mates for life, especially my housemates, Jack and Ollie. I feel very lucky to work with such a great, social, bunch of people with whom I can have a scientific discussion one minute to the pub the next. I would like to thank Victoria, my project student, for all of her hard and amazing work which has helped to really propel part of this study. Finally, a special mention to Cameron - you're a massive legend and I can't wait to drag you to more Fred gigs.

To my amazing partner, Anna, I would like to thank you for your unwavering support and for putting up with my incessant moaning and ailments. Meeting you in this group was one of the best things I could have wished for.

To all my friends at home, I want to thank you all for being my best friends for the best part of 16 years. Having such a solid friend group is rare and I am very lucky to have you. Finally, to my family, Trish, Simon, Katie and Lucy (yes, even Lucy..), I would like to thank you all for your support throughout, even when I have been busy and distant. Dr Katie next??

Definitions and Abbreviations

Common Mathematical Terms, Symbols, and Constants

$[X]$	Concentration of Species X
β	Thermodynamic Beta ($1/k_BT$)
$\Delta G^{\ominus}$	Standard State Binding Free Energy (Gibbs)
δt	Time Step
ΔX	Change in X. Often used for change in Free Energies e.g. ΔG
Λ	Thermodynamic Wavelength
λ	Alchemical Coupling Parameter
Λ_p	NCMC Protocol
$\langle A \rangle$	Ensemble Average of Property A
μ	Chemical Potential
μ'	Chemical Potential (Excess)
μ^{id}	Chemical Potential (Ideal)
Ω	Grand Potential
$\pi(x)$	Equilibrium Probability of State x (Monte Carlo)
ρ	Number Density
$\rho(\mathbf{r}^N, \mathbf{p}^N)$	Equilibrium Probability of a given Microstate
τ	GCNCMC Switching Time
$\mathbf{p}^N$	Momenta Vector for N Particles
$\mathbf{r}$	Atomic Position Vector
$\mathbf{r}^N$	Position Vector for N Particles

$\mathbf{s}^N$	Scaled Position Vector for N Particles
$\mathbf{v}$	Atomic Velocity Vector
Ξ	Grand Canonical Partition Function
$A(\mathbf{r}^N, \mathbf{p}^N)$	Value of A in a given Microstate
B	Adams Value
B_{50}	Adams Value at 50% Average Occupancy, also the dimensionless Free Energy of Transfer.
B_{eq}	Adams Value at Equilibrium
c_x	Concentration of Species x
E	Total Energy
F	Helmholtz Free Energy
F'	Helmholtz Free Energy (Excess)
F^{id}	Helmholtz Free Energy (Ideal)
G	Gibbs Free Energy
h	Planck's Constant
K_b	Binding Constant
K_D	Dissociation Constant
k_B	Boltzmann Constant
N	Number of Particles
$N(B)$	Number of Molecules at Adams Value, B
N_A	Avogadro's Number
n_{pert}	Number of Perturbations in an NCMC Move
n_{prop}	Number of MD Steps Betweeen Each Perturbation in an NCMC Move
P	Pressure
Q_{NVT}	Canonical Partition Function
S	Entropy
T	Temperature

U	Potential Energy
V	Volume
$V(c)$	Average Volume per Molecule at Concentration, c
$V^{\ominus}$	Standard State Volume
V_{GCMC}	Volume of GCMC Region
W	Work
W_p	NCMC Protocol Work
Z_{NPT}	Isothemal-isobaric Partition Function
$B_{eq}(c)^{\ominus}$	Adams Value at Equilibrium with a Non-standard State
$B_{eq}^{\ominus}$	Adams Value at Equilibrium with the Standard State
HAC	Heavy Atom Count
LE	Ligand Efficiency

Common Acronyms and Abbreviations

βCD	β-cyclodextrin
μVT	Grand Canonical Ensemble
ABFE	Absolute Binding Free Energy
AR	Androgen Receptor
BAR	Bennett Acceptance Ratio
BLUES	Binding modes of Ligands Using Enhanced Sampling
ERα	Estrogen Receptor α
ERK2	Extracellular signal-regulated kinase 2
FBDD	Fragment-Based Drug Design
FEP	Free Energy Perturbation
GAR	GCMC Acceptance Ratio
GCI	Grand Canonical Integration
GCMC	Grand Canonical Monte Carlo
GCNCMC/MD	Combined GCNCMC Sampling with Regular MD

GPB	Glycogen Phosphorylase B
HAC	Heavy Atom Count
HEWL	Hen Egg White Lysozyme
HMR	Hydrogen Mass Repartitioning
HSP90	Heat Shock Protein 90
LE	Ligand Efficiency
LJ	Lennard-Jones
MBAR	Multistate Bennett Acceptance Ratio
MC	Monte Carlo
MD	Molecular Dynamics
MSCS	Multiple Solvent Crystal Structures
MSMD	Mixed Solvent Molecular Dynamics
MUP1	Major Urinary Protein 1
NCMC	Grand Canonical Nonequilibrium Candidate Monte Carlo
NCMC	Nonequilibrium Candidate Monte Carlo
NPT	Isothermal–isobaric Ensemble
NVT	Canonical Ensemble
p53	Tumor Protein 53 Y220C
PDK1	Phosphoinositide-Dependent Kinase-1
PME	Particle Mesh Ewald
PPARγ	Peroxisome Proliferator-Activated Receptor γ
PTP1B	Protein Tyrosine Phosphatase 1B
RBFE	Relative Binding Free Energy
SBDD	Structure-Based Drug Design
T4L99A	L99A Mutant of T4 lysozyme
T4L99A/M102Q	L99A/M102Q Mutant of T4 lysozyme
TI	Thermodynamic Integration

Chapter 1

Introduction

1.1 Fundamentals of Small Molecule Drug Discovery

Small molecule drug design is a major field of medicinal chemistry which focuses on the development of synthetic low molecular weight compounds that can modulate biological processes by interacting with a specific molecular target, typically a protein. The goal is to identify compounds that bind selectively and potently to a given target, thereby altering its function in a way that treats or prevents disease. The process typically begins with identifying a disease-relevant target, commonly enzymes, receptors, or ion channels, and validating its role in the disease of interest. Targets may have 'druggable' sites, usually a pocket or groove where a small molecule can bind with sufficient affinity and specificity to modulate the target's activity. However, many drug targets lack apparent binding sites but may have 'cryptic' sites which only become visible upon ligand binding.[2–4]

Ligands are the small molecules designed or discovered to bind to these targets. They aim to mimic, enhance or diminish the behaviour of natural substances within the body. The nature of the interaction between ligand and target is critical and is governed by many different physicochemical properties, such as hydrogen bonding, hydrophobic interactions, and electrostatics. Drug design involves optimizing these interactions to improve binding affinity, selectivity, and pharmacokinetic properties while minimizing toxicity. There are two main strategies in small molecule drug design: structure-based and ligand-based. Structure-based design leverages the three-dimensional structure of the target to rationally design molecules, often using computational methods like molecular docking. Ligand-based design, in contrast, relies on knowledge of other molecules that bind to the target and uses their properties to guide the design of new compounds.[2,4]

As of April 2025, there are approximately 2,990 small molecule drugs approved by the United States Food and Drug Administration (FDA), making up 65% of all approved

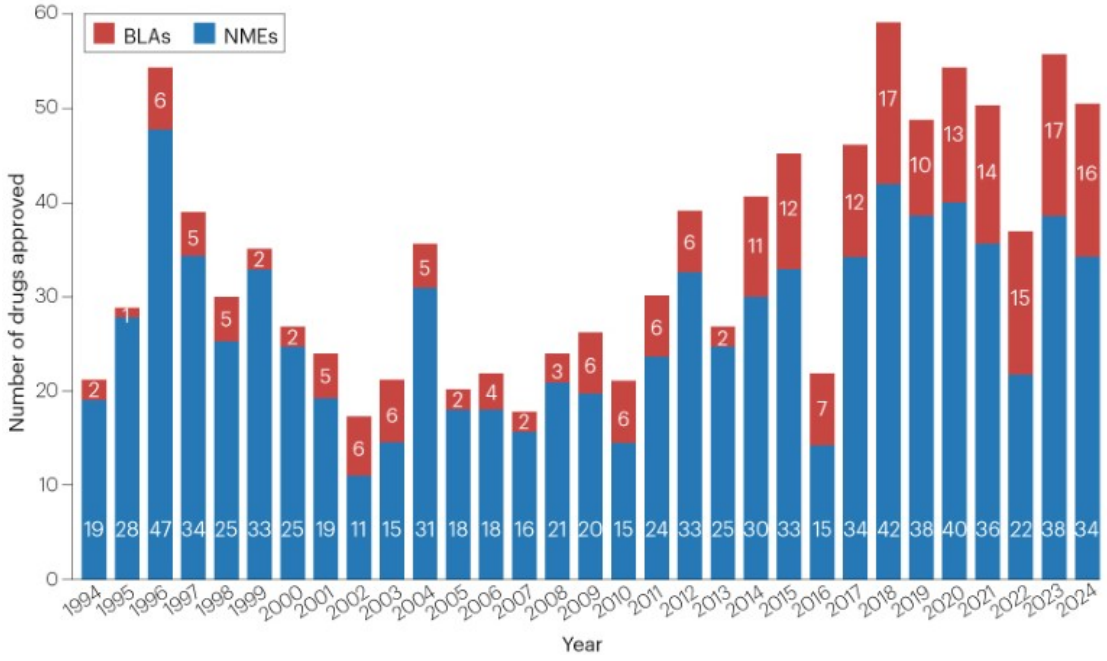


FIGURE 1.1: Numbers of new molecular entities (NMEs), or small molecules, approved by the FDA's Center for Drug Evaluation and Research between 1994 and 2024.

drugs.[5] Figure 1.1 shows the number of FDA drug approvals between 1994-2024, highlighting the continued dominance of small molecules as a therapeutic strategy.[6]

1.2 Structure-Based Drug Design

Structure-based drug design (SBDD)[7,8] is a powerful form of rational drug discovery. Fundamental to SBDD is using the 3D structure of a target protein to rationally and iteratively design new molecular entities using various tools and one's intuition. Rather than blindly screening large libraries of molecules against a target, SBDD is a more methodical approach. Often, smaller screens are performed, either experimentally or virtually, to identify initial hits and novel binding sites. Then, by looking at the resulting structures of these screens, medicinal chemists can identify and exploit useful interactions to develop lead-like molecules. Sequencing efforts such as the human genome project,[9] the advent of high throughput methods of structure determination, and public databases such as the Protein Data Bank,[10] have laid the foundations for SBDD.

SBDD generally follows the iterative Design, Make, Test, Analyze (DMTA) cycle.[7] Once a 3D structure of the target is obtained through either experiment or computation, the next step is to identify a binding pocket. Binding site identification can be achieved by experimentally screening small fragment molecules (such as MiniFrags[11]), with computational methods or even with extensive human knowledge

of the target. This is where the DMTA cycle begins. Using knowledge of the binding site and molecules from an initial screen, medicinal chemists and modellers can begin to design new molecules. They may use virtual or experimental screens to aid this process. Once new designs are developed, they can be synthesised and tested for potency using either functional assays or biophysical techniques. In many cases, a new updated structure is obtained as the starting point for further optimization. This process continues until a lead compound is developed or the project is cancelled owing to financial constraints or lack of progress.

While many lead compounds are developed using SBDD, some examples of successful drugs developed using a primarily SBDD approach include the HIV protease inhibitors amprenavir and nelfinavir,[12] zanamivir which targets neuraminidase,[13] tomudex designed against thymidylate synthase[14] and imitinab mesylate which inhibits Abl tyrosine kinase.[15] Crucially, SBDD relies heavily on *in silico* methods at almost every stage, many of which are described in a following section.

1.3 Fragment-Based Drug Design and MiniFrags

Drug design is one of the largest industries in the chemical sciences, and chemists are continually trying to find ways to optimise the process. There are many bottlenecks in current drug design processes, one of which is the identification of lead compounds.[16–24] Traditionally, lead compounds are found by high throughput screening (HTS) methods that screen large libraries of drug-like molecules to find compounds that cause some change in a functional or biochemical assay. Leads are then optimized to increase the binding affinity while maintaining or enhancing the drug-like properties of the compound. Drugs are then thoroughly tested before being subjected to clinical trials, where the failure rate for the drug design process is highest, with lack of efficacy often cited as the reason. Specifically, analysis of clinical trials between 2010 and 2017 revealed that 80% of clinical candidates failed owing to poor clinical efficacy and safety concerns.[22,23,25–30]

Many pharmaceutical companies have access to large commercial and confidential libraries of chemical compounds that can be screened against target-specific assays. A hit is then loosely defined as a compound that gives a desired outcome. In a biochemical assay for example, a hit may reduce downstream processes which can be measured. Traditionally, these libraries contain drug-like compounds that are found to loosely obey Lipinski's rule of 5 (Table 1.1).[31] However, it has been shown that these libraries cannot possibly account for all the molecules available in the chemical space, which could amount to 10^{60} molecules with ≤ 30 non-hydrogen atoms.[32,33] Generally, the hit rate (number of hits versus number of molecules screened) for screening large molecules is low, and optimisation can be cumbersome given the already complex

TABLE 1.1: Summary of the "guidelines" used in the drug discovery process. "Rule of Five"[31] for drugs and the "Rule of Three"[37] for fragments.

	"Rule of Five"	**"Rule of Three"**
Molecular weight	< 500	< 300
ClogP	≤ 5	≤ 3
Hydrogen bond donors	≤ 5	≤ 3
Hydrogen bond acceptors	≤ 10	≤ 3

structure of the drug-like hit. As the size of the molecules is reduced, the number of possible structures decreases exponentially, and therefore, it becomes more efficient to screen smaller libraries (10^3) of smaller-sized molecules.[32,34] This is the underpinning philosophy of Fragment-based drug design (FBDD).

FBDD optimises hit identification by screening libraries of smaller molecules that tend to generate more, but less potent, hits compared to larger compounds and is a complementary approach to SBDD.[32,35,36] Although less potent, these fragment hits often provide a more useful starting point for optimization as they are simple compounds with greater options for modification. In 2003, Congreve *et al.* proposed that fragments should obey the "rule of three"[37] which is an adaptation of Lipinski's "rule of five".[31] It states that fragments should have a molecular weight of < 300, logP ≤ 3, the number of hydrogen bond donors and acceptors should each be ≤ 3 and finally the number of rotatable bonds should be ≤ 3 (Table 1.1). Although not a strict rule, many fragment libraries follow these guidelines.

Other than a valid binding event, other metrics for a 'good' hit in FBDD include Ligand Efficiency (LE),[38] Group Efficiency (GE) and the Ligand-Lipophilicity Efficiency (LLE). These metrics are a relatively new development and can be used to compare hits and guide the optimisation process. Ligand efficiency is defined as the total free energy of binding of a ligand for a specific target, averaged for each heavy atom in the fragment.[39,40] Ligand efficiency provides an alternative means to compare hits based upon their ratio of binding affinity to the number of atoms, whereas traditionally, hits would be selected on their binding affinity alone, which becomes biased towards larger compounds.[40] Analysis of drug-like molecules that obey Lipinski's rules gives the lower limit for ligand efficiency as 0.3; this serves as a guideline as to which fragment hits may be worth pursuing. Fragment hits which have LEs higher than this value, make for a good candidate as the LE is likely to decrease during optimisation, and thus any screened fragments with a LE lower than 0.3 may be automatically discarded.[39–41] The equation for the ligand efficiency is shown below where HAC is the heavy atom count and the binding free energy $\Delta G^{\ominus}$ is measured in units of kcal mol^{-1}.

$$LE = -\frac{\Delta G^{\ominus}}{HAC} \tag{1.1}$$

Group efficiency[42] (GE) is a metric derived from the LE and is used to estimate the effect that adding a new functional group to a compound will have on the molecule's overall binding free energy ($\Delta G^{\ominus}$). This gives medicinal chemists a quick way of comparing the viability of different possible modifications or additions to the fragment. Unsurprisingly, it transpires that only adding groups with similar or higher GE compared to the LE of the fragment will maintain or improve the potency of the compound.[39–41] Group efficiency is defined as:

$$\begin{aligned} GE &= -\frac{\Delta\Delta G}{\Delta HAC} \\ \Delta\Delta G &= \Delta G(B) - \Delta G(A) \\ \Delta HAC &= HAC(B) - HAC(A), \end{aligned} \tag{1.2}$$

where the binding affinity gained by the new compound, B, owing to the introduction of more heavy atoms (ΔHAC) to fragment A, can be expressed as the difference in the binding free energies between the optimised compound B and fragment A (ΔΔG).[40]

The final metric, Ligand-Lipophilicity Efficiency (LLE),[43] is a useful measure of the ADMET properties (Absorption, Distribution, Metabolism, Excretion, and Toxicity). Generally, the more lipophilic the ligand, the more favourable the desolvation of the ligand is upon binding (thus increasing the binding affinity). However, if the lipophilicity becomes too high, it can result in a loss of specificity, leading to an increased risk of unwanted, potentially toxic side effects.[39,43] LLE is therefore a measure used to optimise the affinity while ensuring the compound does not become toxic.[39,40] The LLE is given by Equation 1.3, where IC_{50} is the half maximal inhibitory concentration and $\log P$ is the logarithm of the partition coefficient of a molecule between water and an organic solvent and serves as a proxy for lipophilicity.

$$LLE = pIC_{50} - c \log P \tag{1.3}$$

A high LLE suggests that the compound gains a lot of its binding affinity through direct interactions rather than favourable desolvation, thus meaning the compound likely has a high specificity. It has been suggested that drug candidates maintain an LLE in the range of 5-7 or higher.[39,40,43]

As of February 2024, seven FDA-approved drugs have resulted from fragment-based drug discovery campaigns with another 59 molecules in clinical trials.[44] Between 2015 and 2022, 180 fragment-to-lead studies have been published, and approximately 7% of

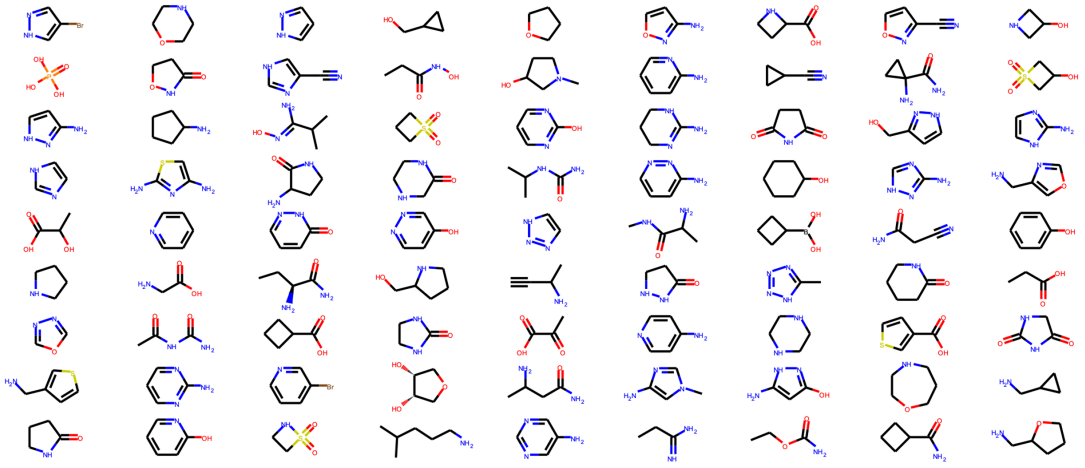


FIGURE 1.2: The 81 compounds in the MiniFrag library.

all clinical candidates published in the Journal of Medicinal Chemistry between 2018 and 2021 originated from fragment screening. This trend highlights the increasing importance of fragments in the drug discovery process.[21,45,46]

The MiniFrags[11] library, designed by Astex Pharmaceuticals, contains only the smallest fragments with a heavy atom count of 5 to 7. The library is comprised of 81 carefully selected, chemically diverse and highly soluble ligands which cover a large range of the chemical space (Figure 1.2). In line with the principles of FBDD, the advantages of a small library include faster screening efforts and improved hit rates owing to the simpler molecular structure of MiniFrags. Table 1.2 compares the average properties of the MiniFrag library to a standard fragment library used by Astex.

TABLE 1.2: Average properties of the MiniFrags library compared to an in-house Astex X-ray Fragment library. The standard deviations are in parentheses.

Average Properties									
Library	N	HAC	MW	clogP	PSA	NDon	NAcc	Nrot	Fsp^3
MiniFrags	81	6.4	94	-0.4	43	1.9	1.5	0.4	0.5
		(0.8)	(16)	(1.0)	(16)	(1.3)	(0.8)	(0.6)	(0.4)
X-ray set	440	10.6	150	0.6	50	1.7	2.0	0.8	0.3
		(1.6)	(25)	(1.1)	(17)	(1.2)	(1.0)	(0.9)	(0.3)

As well as designing this library, Astex have also proposed an experimental procedure to screen the MiniFrags. They present an X-ray crystallography method that soaks the protein using 1 M aqueous solutions of fragments. The authors note that the multiple solvent crystal structures[47] (MSCS) methodology uses a similar protocol but requires much higher concentrations of organic solvents (1-20 M) to the point where very few crystal systems can tolerate the high organic loads employed during soaking, often leading to denaturation. Additionally, the non-aqueous conditions make the results

less physiologically relevant. Owing to the properties of the MiniFrags, the authors find that the harsh conditions employed by the MSCS method are not required for MiniFrag binding, therefore making it a more widely applicable approach. Lastly, in a more recent study, an analogous electrophilic MiniFrag library has been proposed as a means of mapping binding sites for potential covalent inhibitors.[48]

A central question which remains unanswered is how can computational methods be utilised to complement this experimental regime. Such methods are not limited by the same practical considerations such as fragment aggregation, solubility and protein denaturation, but rather they come with their own set of limitations usually associated with the simulation timescales accessible. In this report, we explore the use of some older and newer methods in the context of very small molecule binding.

1.4 Experimental Methods in Structure and Fragment-Based Drug Design

As already mentioned, SBDD and FBDD tend to work hand in hand with each other and, as such, share many of the same experimental methods. Fragments, due to their small size, tend to be soluble in organic medium and have a low binding affinity to their targets with dissociation constants (K_D) typically in the millimolar range. These binding events are hard to detect and many specialized techniques have been developed to aid this task. Many of these methods involve the use of biophysical techniques and a brief overview of some of the more prevalent methods is discussed below. These methods are loosely separated into methods of determining 3D structure and methods used to calculate how tightly a molecule binds to its target. All experimental methods have advantages and disadvantages and in practice, any one of them could be used depending on the specific project and desired outcomes.

1.4.1 Structure Determination

Fundamental to SBDD are 3D structures of the biological target of interest. Depending on what stage of the cycle the campaign is in, the structure may or may not be solved in the presence of a fragment, a hit, or a lead. By far the most popular method of structure determination is X-ray crystallography. Despite having a relatively low throughput, it yields both hit detection and a starting point for optimisation by giving a complete picture of ligand binding in terms of its atomic coordinates.[7,32,49] In these experiments, the structure of the crystal is modelled based on the electron density derived from the diffraction pattern produced when the crystal is hit with a beam of incident X-rays. Crucially, X-ray crystallography requires a large amount of protein (milligrams) and a stable crystal. Producing the latter is often the most time

consuming as conditions need to be optimized in a system dependent manner. That said, improvements in robotics, automation and cell lines are making it increasingly easier to produce high quality crystals. In terms of fragment screening, or producing structures of target-ligand complexes, there are two routes. The first, crystal soaking, such as that used in the MiniFrags study, transfers the pure protein crystal to a solution containing the molecules of interest allowing the molecules to diffuse through the crystal solvent channels and bind to the protein. The second is co-crystallisation where the ligand is added to the protein solution before any crystal formation begins. Often, the latter is more difficult as different ligands may alter the conditions at which crystals form.[36,50–53]

Other limitations can arise in the assignment of electron density. Proteins, or bound ligands, with a high degree of disorder or high mobility, can result in weak or scrambled electron density which cannot be resolved. At best, an average of the different states may be assigned but this can sometimes lead to confusion. An example would be a highly mobile side chain which flips readily between conformations, in which case there will be electron density present for both conformers and the final result may be a combination of both.[54,55] Further, resolving the electron density of small fragments can be problematic especially if they are mobile. Take pyridine for example, it is usually difficult to assign the position of the nitrogen atom using the electron density alone as the molecule will appear to a crystallographer as simply a small region of electron density shaped like a 6-membered ring. Atoms which are isoelectronic, such as water and sodium ions, cannot be distinguished on the basis of electron density alone. Atoms with similar properties such as nitrogen and oxygen are often hard to assign. Finally, it is often forgotten that a crystal structure deposited onto the PDB[56] is ultimately one crystallographer's subjective interpretation of the measured electron density and although automation tools help significantly in maintaining consistency, similar structures can be assigned different atomic coordinates by different practitioners.[21,32,51,57]

Cryogenic electron microscopy, or cryo-EM, in contrast to X-ray crystallography, fires high energy electrons at a sample which has been rapidly frozen to cryogenic temperatures.[58,59] The scattered electrons produce thousands of 2D projections of individual particles which are then computationally reconstructed to build a 3D image of the sample. The method has long been tainted by poor throughput and low resolution structures, but with more recent advancements in data collection and analysis, cryo-EM can now produce structures at near atomic resolution and thus offers an alternative but complementary method to X-ray crystallography.[60–62] The most beneficial aspect to cryo-EM is that there is no need for crystallization and therefore it is much more applicable to large, dynamic and flexible structures such as membrane proteins, multimer complexes and intrinsically disordered proteins. Cryo-EM is now feasible in the context of FBDD and will continue to improve.[51,63]

Nuclear magnetic resonance (NMR), is another tool widely utilized in FBDD and SBDD.[36,49–51,64–66] In the context of structure prediction, NMR has a unique advantage over the above methods in that it better captures the dynamic nature of proteins in solution, giving the researcher even more information on the behaviour of the protein and may give access to conformations missed in static structure methods. That said, NMR's applicability in structure determination is limited to smaller proteins ($<$ 40 kDa) as the spectra become increasingly complicated to assign for larger structures.

When it comes to fragment screening, NMR is advantageous as it provides a quick and convenient way of hit detection. Protein-observed NMR, such as chemical-shift perturbation NMR,[67] requires isotopically labelling the protein with ^{15}N or ^{13}C and comparing spectra of the protein with and without the presence of a ligand. Upon ligand binding, the local environment is changed and the chemical shifts of surrounding amino acids are affected, and as such, this method not only detects hits but also gives an idea of their location. The method is highly sensitive and can detect even the weakest of interactions. Gradually increasing the concentration of ligand while monitoring the change in chemical shifts of specific protein residues provides a means of calculating binding affinity.[68] Protein-observed NMR can be time consuming as it requires a full mapping of the chemical shifts, and methods of isotopically labelling the protein.

A simpler method, saturation-transfer difference (STD) NMR,[69,70] is a ligand-observed technique which measures the change in the 1H NMR spectrum of the ligand with and without the presence of the protein. In STD, radiofrequency pulses are selectively applied to saturate the NMR signals of the protein without affecting the unbound ligand in the solution. As the ligand binds there is a small transfer of saturation, and consequently, a decrease in the intensity of the ligand spectrum indicates a valid binding event. Again, this technique is very sensitive and can detect even the weakest binders. Affinity prediction is possible in principle with STD, but in practice difficult; as such, STD is regarded as more of a qualitative method.[50]

1.4.2 Affinity Measurement

Key to the SBDD and FBDD pipelines is the ability to measure how strongly a molecule binds to its target. Here, we discuss some more biophysical techniques which give affinity measurements to varying degrees of accuracy. Affinity prediction is extremely important in identifying fragments with high potential (e.g. high ligand efficiency), in the optimisation of a hit to a lead, in developing a structure-activity relationship (SAR) and in enhancing hit validation.

Fragments, in general, often do not elicit responses in functional assays (an assay specific to the target where a desired agonism or antagonism is measured) the same

way as larger lead-like molecules.[36,50] It is for this reason that affinity prediction using biophysical techniques is preferred but it should be noted that affinity does not always translate to activity or inhibition and this relationship should be investigated on a per target basis.

Isothermal Titration Calorimetry (ITC)[71] is a highly sensitive technique that measures the heat changes associated with molecular interactions and is often regarded as the gold standard in affinity measurement. In an ITC experiment, a ligand solution is slowly injected into a sample solution of protein, and as the ligand binds to the protein, heat is either released or absorbed which is measured by highly sensitive sensors. This heat change can be measured as a function of ligand concentration and gives direct measurements of the binding affinity, the reaction enthalpy and the entropy change. Unfortunately, fragment interactions are sometimes too weak to be detected by ITC which limits its application in high throughput screening. Additionally, it requires a large amount of protein and can be time consuming. That said, it is very useful in the hit-to-lead optimisation stage of the drug design process.[36]

Surface plasmon resonance (SPR)[72,73] is a more sensitive affinity prediction method which is more applicable to fragment binding. In SPR, the target is immobilised on a sensor chip coated with a thin layer of gold or silver and the fragment solution is flowed over the surface of the chip. Upon ligand binding the refractive index of the surface changes and therefore reflects light differently. Like ITC, the change in refractive index can be measured as a function of fragment concentration to predict binding affinity. Interestingly, the time taken between adding the ligand solution and the change in refractive index indicates the on rates, and vice versa the off rates, giving valuable kinetic information as well. SPR is high throughput and requires less protein than ITC making it a cost and time effective method of screening fragments.[36,50]

Finally, thermal shift assay (TSA),[74] or differential scanning fluorimetry (DSF), are reliable and simple techniques which measure the temperature at which a protein is denatured. Fundamentally, the stability of a protein correlates to its melting temperature and is related to environmental factors such as solution composition, amino acid mutations, pH and crucially the binding of a ligand. In DSF, a fluorescent dye is introduced to the protein solution which shows a weak fluorescence in polar environments and larger signals in apolar environments. The protein solution is then heated to the point at which the protein becomes denatured, thus exposing its apolar core and producing a more intense fluorescence signal, meaning the melting temperature of the protein can be determined. As mentioned, the binding of a ligand can either stabilise or destabilise the protein leading to a change in the protein melting point and as such a change in melting point is indicative of ligand binding. TSA's are very high throughput and therefore serve as a very useful primary screening step in FBDD, though it has been noted that TSA often has a large number of false positives and it is difficult to relate melting temperature to binding affinity.[36]

1.5 Computational Methods in Structure and Fragment-Based Drug Design

In recent years, *in silico* methods, which involve the use of computer simulations and computational techniques, have seen a significant rise due to advancements in computational power, more sophisticated algorithms, machine learning, and the growing availability of large biological datasets, making them increasingly integral to research and drug discovery processes. Notably, computational tools can be used in library design, virtual screening, property prediction, ligand characterisation and in hit to lead optimization.

In silico methods are not drop-in replacements for experimental methods but are powerful complements as they offer speed, scalability, and cost-efficiency. Various computational methods are described in the following sections and are loosely split into different categories. Like experimental methods, there is never a 'one size fits all' solution and drug discovery campaigns are likely to make use of many different techniques at differing stages of the process. Finally, computational method development is an ever evolving field with new ideas, concepts and technologies emerging rapidly. The following are some examples of ongoing work in this field but is by no means an exhaustive list.[20,24,75]

1.5.1 Structure Prediction

As mentioned, experimental methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy can provide highly accurate structural data but are often resource intensive, time consuming, and sometimes inherently limited by the protein systems being studied. Some proteins, such as membrane-bound or highly flexible proteins, can be particularly challenging to crystallize or analyze using these methods. In contrast, computational approaches can predict protein structures with fewer physical constraints, making them invaluable for rapidly generating models for proteins that are difficult to study experimentally.

One common method of *in silico* structure prediction is homology modelling,[76] which leverages the fact that protein structures are more conserved than sequences. In this approach, the sequence of a target protein is aligned to a known structure (template) that shares a significant degree of sequence similarity. By using the template's 3D structure as a framework, homology modelling algorithms predict the folding of the target protein. This method can be quite accurate when the template has a high sequence identity with the target protein. Two tools for homology modelling are MODELLER[77] and SWISS-MODEL,[78] both of which generate models by adapting

known structures to the query sequence. The accuracy of homology modelling is very dependent on the template and its similarity to the query sequence.

More recently, deep learning methods, such as RoseTTAFold[79] and AlphaFold 3,[80] have revolutionised *in silico* structure prediction.[81] AlphaFold and RoseTTAFold use a neural network architecture trained on massive datasets of known protein structures. These methods can learn millions of parameters by integrating multiple sources of information culminating in the ability to model the long-range dependencies in the protein sequence. From a sequence, AlphaFold predicts pairwise residue distances and angles, effectively building a model of the protein spatial conformation, where it is then fed to an iterative cycle which further refines the structure to give a low energy conformation. In general, AlphaFold has outperformed all other methods, highlighted by its performance in the blind Critical Assessment of Structure Prediction competition.[82] These methods are more versatile than homology modelling and do not require the use of a template, however, there are limitations associated with the quality of the training data.

Firstly, an obvious challenge is that predicting the structure of a protein with few close relatives is difficult as there are not enough sequences in the training set against which to compare. Further, protein systems which are under-represented in training sets, such as membrane-bound proteins can be hard to predict. Conversely, protein conformations which are over-represented in the training set can bias results. For example, it has been shown that the majority of the PDB contains ligand bound structures and as such AlphaFold generally predicts the structure of proteins more akin to their ligand bound conformation.[83–85]

1.5.2 Hotspots and Binding Site Identification

A key step in the SBDD pipeline is the identification and characterisation of ligand binding pockets within a target protein. As already mentioned, the MiniFrags and other experimental methods can be used to find regions where ligands bind. Computationally, a few methods exist to do the same and can loosely be split into static structure methods and molecular dynamics (MD) based methods. The former usually takes a static 3D structure of the protein and exhaustively samples and maps putative binding sites by placing probe/fragment molecules over the structure and then scoring the resulting poses using an energy function. These methods include GRID,[86] Multiple Copy Simultaneous Search[87] and FTMap,[88] all of which are computationally efficient and have had continued success. That said, the lack of protein flexibility can limit the application of these methods to protein systems with pre-formed binding sites, though in principle, one could feed these algorithms with an ensemble of different conformations generated by simulation or machine learning techniques. More details on the FTMap algorithm are found in a following chapter.

MD-based methods provide a way to incorporate protein dynamics and flexibility and, as such, are more applicable to a range of problems. One flavour of simulation, analogous to crystal soaking, is called mixed-solvent molecular dynamics (MSMD) whereby a protein is solvated in a solution of water and small organic cosolvent molecules. Several short simulations are then run with the intention that these probe molecules, which are no larger than MiniFrags, bind favourably in druggable binding pockets.[89] MDMix,[90] SILCS,[91] and MixMD[92] are three well-known implementations of MSMD and differ slightly in their respective design philosophies. For example, MixMD uses a relatively low concentration of fully miscible probes to prevent aggregation while SILCS uses a high concentration and implements a repulsive potential between probes to maintain a well-mixed solution. The resulting data of MSMD simulations can be used to identify hotspots and favourable interactions which can be exploited when designing lead compounds. Furthermore, the resulting interaction patterns can be used to develop pharmacophore models that guide docking studies by prioritising ligands, and conformations, with the same pattern as the binding site.

MSMD has been shown to outperform static structure methods[93] but can often fail when the binding event, even for these small probes, happens over a time scale longer than what can be reasonably simulated. This is particularly a problem for binding sites which are occluded from the solvent, or for cryptic pockets which require large conformational changes to occur as the ligand binds. SILCS have presented a grand canonical Monte Carlo-like approach to facilitate binding in occluded binding pockets.[94] This method oscillates the chemical potential of the probes to drive insertions/deletions from the binding pockets but it should be noted that this method does not follow detailed balance. In other MDMix and MixMD studies, the MSMD protocols have been combined with accelerated MD, which is a method that adds a bias potential to the system's potential energy encouraging the sampling of conformational transitions and allowing the identification of transient or cryptic binding pockets which would otherwise be difficult to observe within classical simulation timescales.[95,96] MSMD methods are explored in further detail in Chapter 7.

Sampling Water Interfaces through Scaled Hamiltonians[97,98] (SWISH) is another MSMD-based method aimed at identifying cryptic pockets, or pockets which require a large conformation change to be revealed, which uses a Hamiltonian Replica Exchange[99] protocol to scale the interactions between water molecules and apolar protein atoms. The implication is that these 'strengthened' molecules may pry open cryptic pockets so that a small fragment can bind, triggering the full opening of the pocket. SWISH has had ongoing development and success in identifying cryptic pockets in a wide range of proteins.[100]

1.5.3 Virtual Screening and Hit Identification

Once a putative binding site is found, the next stage is to find hits to start optimising. While it is possible to use the results of an MSMD simulation as a starting point, it is often useful to start with a more substantial molecule, although one may argue that a bound MiniFrag may suffice. Instead, it is often better to use MSMD results to guide virtual screening efforts.

Virtual screening generally involves using computational methods to evaluate a large library of compounds for their potential to bind in the selected binding site and aims to accelerate the drug design process by narrowing down potentially tens of thousands of molecules to a more manageable number. Molecular docking algorithms are well suited for this task as they are computationally efficient and somewhat reliable.[101,102] In general, docking algorithms will place a molecule in the specified binding site, sample different possible configurations of both the ligand and surrounding side chains and score each configuration according to some scoring function. The molecules, or scaffolds, with the best score can then be taken forward for optimisation. Examples of empirical docking algorithms, or algorithms with their own scoring function, include DOCK,[103] GOLD[104] and AutoDock,[105] and examples of algorithms which use a molecular mechanics-based scoring function include Glide[106] and ROSETTALigand.[107] In the context of FBDD, it has been shown that docking algorithms with empirical scoring functions are usually trained on drug-like molecules and can sometimes fail to rank small fragments, or even differentiate binding modes of the same fragment, correctly.[102,108]

At this stage of the cycle, molecular simulations are somewhat unfeasible owing to the sheer number of potential hits that need to be evaluated. That said, simulations are well-placed to enhance virtual screening and docking. As docking often uses a mostly rigid representation of the protein, it can sometimes be useful to seed the docking algorithms with different conformations from MD simulations. Conversely, it could be useful to take docked structures and run an MD simulation to assess the stability of the predicted binding poses. Binding modes of Ligands Using Enhanced Sampling (BLUES),[109,110] for example, is a nonequilibrium candidate Monte Carlo-based[111] method designed to sample different ligand binding modes within a binding site. BLUES can therefore be used to generate an ensemble of stable configurations and help resolve multiple binding modes in X-ray or docked structures giving a more accurate picture of the binding. While some degree of knowledge of the binding site is required, BLUES has successfully identified multiple binding modes of fragments in T4-Lysozyme and Soluble Epoxide Hydrolase.[109,110] A similar method, Adaptive Alchemical Sequential Monte Carlo uses importance resampling to explore ligand conformational degrees of freedom.[112]

Finally, the ranking of docked poses can be improved by performing free energy calculations to rank said poses based on their binding affinity. However, these calculations are expensive compared to the empirical scoring functions, but when performed correctly, they give a theoretically rigorous and accurate means of ranking and optimising ligands.

1.5.4 Affinity Prediction

While molecular dynamics simulations can fully incorporate protein dynamics in principle, binding events often occur over longer timescales than can be simulated in a reasonable time frame, and to achieve converged results, these binding and unbinding events would need to be simulated multiple times.[113–116] Therefore, one cannot simply run an MD simulation, and record the on and off events to calculate a ligand's affinity. Instead, more complex methods are used.

Molecular Mechanics Generalized Born Surface Area (MM-GBSA),[117] is a relatively cheap method of calculating free energies and is applicable to a relatively large pool of bound/docked molecules.[118–120] These calculations combine molecular mechanics energies with solvation effects, approximated through the Generalized Born model and the solvent-accessible surface area. The MM component considers the potential energy of the system including bonded and non-bonded interactions within the molecular complex, described in Chapter 2. The GB model approximates electrostatic interactions between solutes and the solvent, while the SA term accounts for non-polar solvation effects, which are related to the surface area of the molecule interacting with the solvent.[117] The MM-PBSA method uses the Poisson–Boltzmann (PB) equation to calculate the solvation energies and, in general, is more accurate than MM-GBSA and are traditionally performed in conjunction with MD simulations.[117,118]

While having high throughput, the inherent approximations of the GB model can lead to inaccurate predictions in systems with strong electrostatic interactions.[119] The implicit solvent model is also a simplification, as it does not fully capture dynamic water interactions that are often critical to binding. MM-GBSA methods are sensitive to the conformational state of the system which means the best results are obtained when performing the calculation many times on different snapshots of an MD trajectory, lowering its throughput. In the context of early-stage drug discovery, MM-GBSA calculations may be performed on docked structures to give a complementary ranking to the docking scores.[118,120]

An alternative approach, alchemical free energy calculations, or free energy perturbation (FEP),[121–123] may be used once a binding mode is known. When performed correctly, these are a highly accurate and reliable way to rank a series of ligands *in silico*.[124,125] Relative binding free energy (RBFE) calculations provide means

of predicting the free energy difference between two closely related ligands by perturbing one or two functional groups into another. In the context of fragment-to-lead optimization, RBFEs provide a very convenient way of understanding how a potential modification will affect the binding affinity of a hit.

Absolute binding free energy (ABFE) calculations, on the other hand, calculate the total binding affinity of a molecule to a target and are generally more useful when ranking ligands with different scaffolds; a retrospective study by Alibay *et al.* shows their applicability in FBDD.[126] In practice, most fragment-based drug discovery campaigns can benefit in different ways from both free energy approaches.[123] For both methods there is a requirement for high-quality structural data from either experiment or computation, prior knowledge of the binding modes, and, in the case of ABFEs, a series of user-defined restraints to maintain the complex, as the ligand is decoupled.[127,128]

While many different codes now exist to make the setup and execution of free energy simulations simpler, the various time scale limitations of MD-based methods remain.[129–131] Incorrect sampling of the bound and unbound state can lead to inaccuracies, for example, decoupling the ligand from a *holo* state can lead to overestimating the binding free energy as the ligand is being removed from its most stable pose and the simulations are unlikely to account for any protein reorganization during the actual binding event.[112,132,133] Ligands with more than one binding mode are also challenging. Experimental data includes an ensemble of all possible binding poses and therefore to compare simulated results to experimental data reliably, each unique binding mode needs to be simulated and the free energy of binding for each pose needs to be combined via a Boltzmann average.[109,128] This requirement can make free energy calculations prohibitively expensive, particularly for fragments, which may have many different binding modes.[109] General forcefield errors, convergence problems, and sensitivities to the initial protein/ligand structure also plague the method with many recent studies aimed at improving these issues.[112,134,135]

Traditionally, FEP calculations are only performed on small subsets of ligands owing to time and compute cost constraints. However, this is changing with continuous improvements in computing, more sophisticated automated setups and the advent of active learning protocols.[136–138] Active learning in the context of free energy calculations and virtual screening opens the method up to larger pools of ligands, a by-product of which is greater diversity. In general, a few FEP calculations are performed on a random set of ligands, the active learning algorithms then suggest a new set of ligands to test based on a complex network of chemical properties. The results from each round of calculations are fed back into the training set for further model improvement. As the model iterates, it continuously refines its predictions, homing in on promising ligand candidates while reducing the need for exhaustive calculations. This is especially useful in drug discovery, where active learning helps

identify top-binding ligands more efficiently, reducing the time and computational expense required to screen large molecular libraries.[136–139]

A recent review by Gorantla *et al.*[139] investigated the use of different active learning protocols applied to four datasets. For the largest set, TYK2, it was shown that performing only 360 explicit free energy calculations was enough to predict the affinities of 10000 compounds with an R-squared of over 0.8 compared with experiment while recalling over 50% of the 200 top ranking ligands.

1.5.5 Grand Canonical Monte Carlo

Grand Canonical Monte Carlo (GCMC) simulations have been routinely used to simulate the grand canonical (μVT) ensemble allowing the number of molecules in the system to fluctuate while keeping the overall chemical potential of the system constant.[140–143] In recent years, GCMC has been used to sample buried water molecules in protein-ligand binding regions, to predict favourable water sites, and calculate the free energies of water networks.[135,141,144–147] A particularly interesting use case of water-based GCMC is its ability to enhance free energy calculations by enhancing the sampling of water molecules while an alchemical change is applied to the ligand.[135,148] This facilitates the rearrangement of water networks or the displacement of tightly bound/occluded water molecules which may otherwise fail to be displaced under the timescales of a typical MD simulation.

In practice, this sampling is achieved by trailing random, instantaneous, insertions and deletions of water molecules into/from a region of interest. The move is subjected to a Monte Carlo acceptance test which accounts for the equilibrium properties of the system.[135,147,149,150] However, the acceptance of such moves, particularly in condensed phase systems, can become prohibitively low as each insertion has a high probability of a steric clash with other particles, while the deletion of stable molecules can leave the system destabilized.[135,147,149,150]

In a more recent study, the application of nonequilibrium candidate Monte Carlo (NCMC) to GCMC has been investigated. NCMC itself is an enhanced sampling method developed to improve the acceptance of low-probability Monte Carlo moves via a nonequilibrium switching process.[111,150] In other words, NCMC breaks up a Monte Carlo move into a series of small perturbations with relaxation steps interspersed to allow the surrounding environment to respond to each perturbation. In the context of GCMC, the addition of NCMC means that the insertion or deletion of molecules can occur gradually over a series of alchemical states allowing the molecule to bind with an induced fit mechanism.[149,150] The acceptance rates of this combined "Grand Canonical nonequilibrium candidate Monte Carlo" (GCNCMC) method were found to be significantly improved compared to instantaneous GCMC while also

giving greater insight into the dynamics of water binding, and indirectly improving protein and ligand sampling.[150]

1.6 Objectives

MiniFrags have clear benefits compared to other fragment-based methods, however, there is yet to be any direct computational method that aids this drug discovery regime. This project aims to use experimental MiniFrag, and MiniFrag-like, binding data to develop simulation approaches that can accurately and quickly model the binding of these small fragments. The development of such *in silico* screening methods can be used to not only guide future experiments but can also be used for cases where an experimental approach may not feasible.

Of particular interest is the use of GCNCMC applied to fragment binding.[143,147,149–151] This methodology involves slowly inserting or deleting molecules into a region of interest and then accepting/rejecting the move based on the work done. We aim to develop and expand on the previously published theory and code that performs this sampling with water molecules.[147] The applicability of ligand-based GCNCMC/MD will be evaluated by assessing how effectively and efficiently the method can reproduce experimentally published binding sites in various protein systems. As an extension, the method will also be tested on how accurately it can predict binding free energies and compared on its relative merits to other free energy methods. The method is first tested on a simple host-guest cyclodextrin system and then further tested on simple protein systems such as T4-Lysozyme. The method is developed to use the OpenMM[152] MD engine owing to its versatility and customizability. A Python module, *grandlig*, is developed and made available to the wider community (`https://github.com/essex-lab/grand-lig/`).

Throughout this thesis, we also make use of other popular *in silico* methods such as mixed solvent MD and absolute binding free energy calculations. We primarily use these to validate the GCNCMC implementation but also to demonstrate their applicability in small fragment simulations.

Chapter 3 involves preliminary MD simulations of the MiniFrag-ERK2 system which highlights the need for more sophisticated enhanced sampling methods. Chapters 4, 5, and 6 outline the development and application of GCNCMC to small molecule binding. The similarities and differences between this implementation and water sampling are highlighted. These chapters are ordered in terms of increasing difficulty and relevance to drug design. First, in Chapter 4, we validate the method by reproducing a simple ensemble property, concentration, and studying the method's behaviour in a host-guest system. In Chapter 5, the method's ability to calculate free energies is demonstrated by first calculating hydration free energies for a subset of the

FreeSolv[153] database and then binding affinities for a pool of guest ligands binding to the same host system. Key results in Chapter 5 include the derivation of two related free energy estimators that use and manipulate the nonequilibrium works derived from GCNCMC simulations. Lastly, in Chapter 6, the method is applied to two protein systems, T4L99A and MUP1, where its use in computational SBDD is highlighted by first predicting the binding sites and then calculating affinities for a range of ligands.

Finally, Chapters 7, 8, and 9 highlight different use cases of ligand-GCNCMC which warrant further investigation. Specifically, Chapter 7 combines the method with mixed solvent MD simulations to create a GCNCMC-enhanced mixed solvent MD protocol. Chapter 8 demonstrates how the method can be used in the context of fragment screening and finally, Chapter 9 involves a long discussion on various avenues for further development with some initial data.

Chapter 2

Theory and Methods

2.1 Molecular Dynamics

Molecular Dynamics (MD) is a computational method used to simulate the interactions of molecules through time. The trajectory (containing sets of atomic positions through time) of a simulation can be acquired by successive integration of Newton's second law of motion.

$$\mathbf{F} = m\mathbf{a} \tag{2.1}$$

where the acceleration (**a**) of the molecule can be written as the second derivative of the molecular coordinates with respect to time:

$$\mathbf{F} = m\frac{d^2\mathbf{r}}{dt^2} \tag{2.2}$$

It can also be said that the force (**F**) acting on a particle is equal to the negative gradient of the potential energy, U, with respect to the atomic coordinates (**r**), and as such, Equation 2.2 can be rewritten as the following:

$$\mathbf{F} = -\frac{dU}{d\mathbf{r}} = -\nabla U = m\frac{d^2\mathbf{r}}{dt^2} \tag{2.3}$$

$$-\frac{1}{m}\nabla U = \frac{d^2\mathbf{r}}{dt^2} \tag{2.4}$$

Equation 2.4 shows that a change in a particle's position is directly related to the potential energy of the system and, as such, if the potential energy can be calculated, molecular dynamics simulations can extrapolate the state of the system at a future time based on its current state.[154]

The potential energy between two atoms is dependent on their separation and is constantly changing in a dynamic system. The continuous nature of the dynamic

TABLE 2.1: Suggested time steps for systems with varying degrees of freedom.[154]

System	Motions present	Time step
Flexible molecules + flexible bonds	Vibration, Torsion, Rotation, Translation	0.5 - 1 fs
Flexible molecules with "frozen" bonds	Torsion, Rotation, Translation	2 fs
Flexible molecules with "frozen" bonds and hydrogen mass = 3-4 amu	Torsion, Rotation, Translation	4 fs
Rigid molecules	Rotation and Translation	5 fs
Single atoms	Translation only	10 fs

potentials requires that Newton's equations of motion be integrated at very short intervals. At each step, the forces on the atoms are calculated and combined with the current positions and velocities of the atoms in order to generate new positions and velocities a short time ahead.[154]

If the time step is too large, instabilities in the system can arise since large jumps in time based only on the current potential energy can lead to violations of momentum and energy conservation. If the time step is too small, more computational resources would be required to simulate the same amount of time as if an appropriate time step was used.

As a general rule of thumb, to obtain the most accurate results, the time step should be chosen so that it is at least one order of magnitude smaller than the fastest motion in the system.[154] This tends to be very small for some systems, given that they may contain high frequency motions such as bond vibrations (10^{15} s^{-1}). A simple way to reduce the impact these high frequency motions have on the time step is to 'freeze out' the vibrations by constraining 'spectator' bonds to their equilibrium values. For simulations of biomolecules with bonds to hydrogen frozen, a 2 femtosecond time step is often used.

Another optimisation one may employ to increase the accessible timescale of a simulation is a method called hydrogen mass repartitioning (HMR).[155] In HMR, the mass of hydrogen atoms is increased by subtracting mass from their corresponding heavy atom. This has the effect of making the hydrogens heavier, and therefore slower, while maintaining the overall mass of the system. In turn, the fastest motions of the hydrogen are slowed, allowing for an even larger timestep. Table 2.1 details some suggested time steps for different systems.

2.1.1 Forcefields

In modern techniques, the potential energy of a system is usually described by a forcefield which encompasses the molecular mechanics of the system. The underlying

concepts of these descriptions are fairly simple and based upon a soft ball and springs representation of molecules. In MD calculations the Born-Oppenheimer approximation is used so that there is no explicit representation of electrons which saves computing power and thus time. Both covalent and non-covalent bonds are considered, however, covalent bonds cannot be formed or broken in a basic simulation.

For a typical forcefield, the potential energy of a molecule is described in terms of bond lengths, bond angles, dihedral angles, electrostatics, and Lennard-Jones interactions (Figure 2.1). The former three are 'bonded' terms and only occur between connected atoms of the same molecule (intramolecular), while the latter two are 'nonbonded' terms and describe the long-range interactions between particles and are both inter and intramolecular.

Both the bond length and angle potential energy terms can be approximated using a harmonic oscillator. This highlights the energetic penalty associated with deviating from the equilibrium bond length and angle.

$$U = \sum_{bonds} k_b (r - r_0)^2 \tag{2.5}$$

$$U = \sum_{angles} k_\theta (\theta - \theta_0)^2 \tag{2.6}$$

where k_b and k_θ are the spring constants for the bond length and angle respectively, and r_0/θ_0 is the equilibrium bond length/angle.

The torsion angle, unlike the bond length and angle, is referred to as a "soft" degree of freedom. This means that there is not a single optimal value, but instead multiple

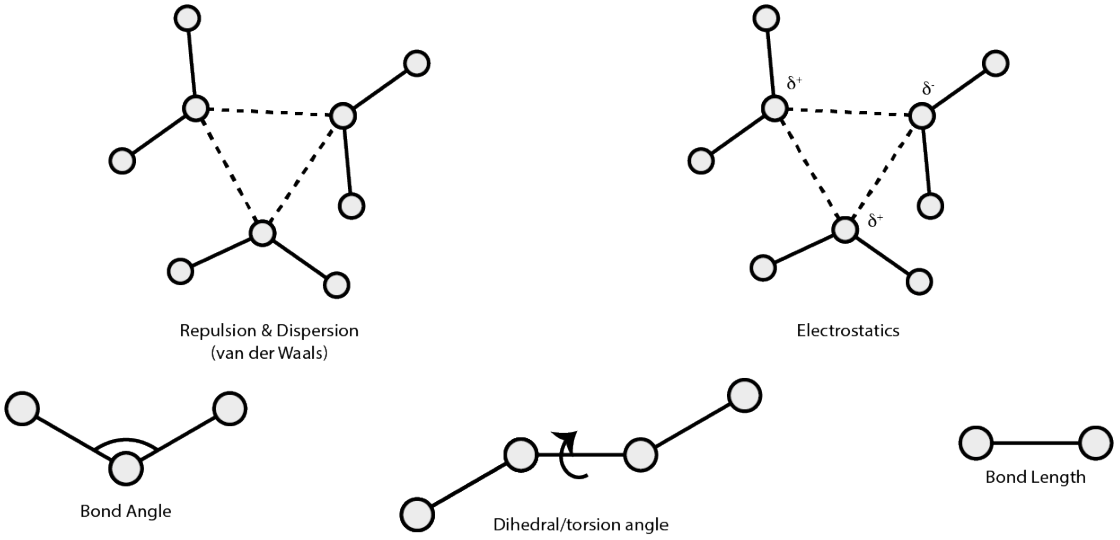


FIGURE 2.1: Schematic representation of the key contributors to a molecular mechanics forcefield.

values which arise in a regular sequence. One classic example is the multiple conformations adopted by ethane, where there are 3 staggered conformations corresponding to energy minima and three eclipsed conformations corresponding to energy maxima. In most forcefields, the torsion potentials are expressed as a set of cosine functions.

$$U = \sum_{dihedrals} A[1 + \cos(n\phi - \phi_0)] \tag{2.7}$$

where ϕ is the torsion angle, n is the multiplicity and defines the number of minima as the bond is rotated 360 degrees, ϕ_0 is the phase factor which determines where the torsion angle passes through its minimum value, and finally A is the barrier height.

The first non-bonded term models the electrostatic interactions between atoms and is modelled using Coulomb's law. This law assumes a partial atomic charge model where more electronegative elements will attract less electronegative elements.[154] The partial charges on atoms i and j are denoted by q_i and q_j respectively and the distance between the two atoms is given by r_{ij}. ε_0 is the vacuum permittivity.

$$U = \sum_{pairs(i,j)} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.8}$$

Lastly, repulsion and dispersion interactions between two atoms are modelled using a Lennard-Jones potential. Despite being small, the van der Waals forces are still significant enough to be included in the forcefield. Both σ and ε are adjustable parameters, where σ is the collision diameter (the separation of two particles where the energy is zero) and ϵ is the well depth parameter which defines the 'stickiness' of the potential. The deeper the well the stronger the interaction is between two particles. The Lennard-Jones potential is shown mathematically in Equation 2.9 and graphically in Figure 2.2.

$$U = 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] \tag{2.9}$$

The total potential energy of a system is given by the sum of all the above terms:

$$U_{tot} = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} A[1 + \cos(n\phi - \phi_0)] + \sum_{pairs(i,j)} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.10}$$

Equation 2.10 represents a Class 1 molecular mechanics forcefield, which is the most widely used for atomistic molecular simulation. Currently, there is much research in

the field focused on improving this class of forcefield and the design of new classes including polarizable and machine learning forcefields.[156–161]

A forcefield that uses the Class 1 representation for the potential energy requires parameters for the various constants including k, A, ϵ_{ij} and σ_{ij}. Generally, the parameters used to define a forcefield are derived from experimental data or are tuned such that they can accurately reproduce different physical properties determined experimentally. Some forcefields may use the same functional form but have different values for the parameters. However, it is strictly incorrect to mix and match parameters from different forcefields since the individual terms can all be related. There is some relaxation of this rule with the bond and angle terms which tend to be sufficiently independent of the other terms.[154]

An important feature of forcefields is the transferability of parameters. This means that the same parameters can be used for a whole series of molecules so that new parameters do not need to be defined for every new molecule. For example, it is beneficial to have one forcefield which can define the parameters for all n-alkanes; without a transferable forcefield, it would be necessary to define parameters for methane, ethane, propane and so on. A useful feature is the use of "atom types" which group similar atoms based on their properties such as hybridisation state, the local environment and even neighbouring atoms, and assign values to the necessary constants based on these properties. Clear use of atom types can be shown in terms of the equilibrium bond angle parameter, θ_0, where it is clear the reference angle will differ between sp^3-hybridised centres and sp^2 centres.

The Lennard-Jones parameters, σ and ϵ, are defined for each atom type, however, to define the interactions between two different atom types we must define a combining

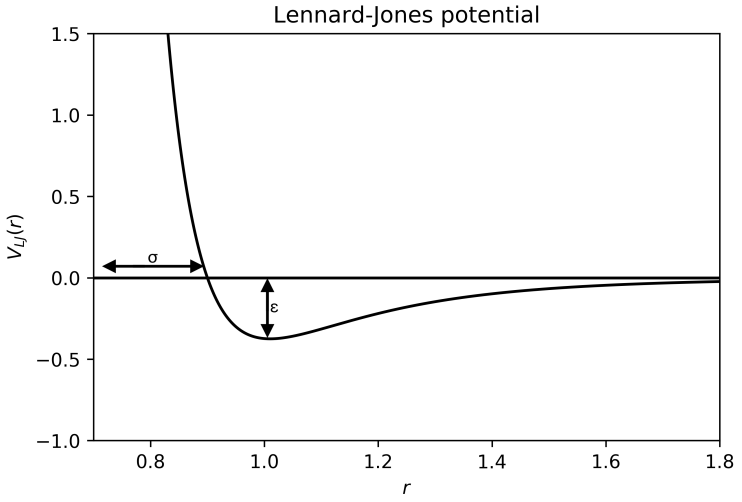


FIGURE 2.2: The Lennard-Jones potential.

rule. A few exist but by far the most popular is the Lorentz-Berthelot combining rules:[162]

$$\sigma_{ij} = \frac{1}{2}(\sigma_{ii} + \sigma_{jj}) \tag{2.11}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{2.12}$$

2.1.2 Integration

As mentioned, the forces acting on a particle are dependent on its position and the positions of many other particles around it, so, for a moving system, the forces are continuously changing. This gives rise to a "many-body" problem where Newton's laws of motion cannot be integrated analytically (for systems with more than two interacting particles), instead, they must be solved using numerical methods based upon the finite differences method.[154,163]

In these so-called finite difference techniques, the integration is broken down into separate smaller integrations. Given the total force on a particle at a given time, t, one can determine the acceleration of the particle and, by combining this with the velocity and position of the particle at time, t, the new positions and velocities can be calculated at a time, $t + \delta t$, where δt is the time step. This process is continuously repeated for the length of the simulation. Many finite difference integration algorithms assume that positions, velocities and acceleration can be approximated using a Taylor series series expansion such that:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 \tag{2.13}$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \mathbf{a}(t)\delta t + \frac{1}{2}\mathbf{b}(t)\delta t^2 \tag{2.14}$$

$$\mathbf{a}(t + \delta t) = \mathbf{a}(t) + \mathbf{b}(t)\delta t + \frac{1}{2}\mathbf{c}(t)\delta t^2 \tag{2.15}$$

where $\mathbf{r}$ is an atoms position, $\mathbf{v}$ is the velocity (the 1st derivative of the positions with respect to time), $\mathbf{a}$ is the acceleration (the second derivative) and $\mathbf{b}$ and $\mathbf{c}$ are further derivatives.

There are many different integration methods, one of the most simple is the Verlet algorithm.[164] The Verlet algorithm uses positions and accelerations at time t, and the positions from time $t - \delta t$ (the previous step), to calculate the new positions at time,

$t + \delta t$. Using the Taylor series expansions for the positions:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 \tag{2.16}$$

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 \tag{2.17}$$

and adding together gives:

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \mathbf{a}(t)\delta t^2 \tag{2.18}$$

In the Verlet algorithm there is no explicit calculation for the velocities and therefore they must be calculated separately. A simple way to do this is to divide the difference in positions by $2\delta t$.

$$\mathbf{v}(t) = \frac{\mathbf{r}(t + \delta t) - \mathbf{r}(t - \delta t)}{2\delta t} \tag{2.19}$$

The Verlet algorithm is relatively simple and requires little computer memory to implement. However, the addition of the $\mathbf{a}(t)\delta t^2$ term to the difference of the two larger terms, $2\mathbf{r}(t)$ and $\mathbf{r}(t - \delta t)$ can lead to a loss of precision. More importantly, the Verlet equations lack a term for the explicit velocity, making it difficult to calculate accurate velocities for the particles at time $t + \delta t$, meaning the temperature of the system cannot be reliably represented or controlled. Finally, at $t = 0$, there cannot possibly be a set of positions for time, $t - \delta t$. This is characteristic of what is known as a "self-starting algorithm" - the positions at $t - \delta t$ must be generated by other means, which can lead to further inaccuracies.

In 1982, Swope *et al*[165] developed a new algorithm, velocity Verlet, a variation on the Verlet method outlined above. The velocity Verlet algorithm provides both the atomic positions and velocities at the same time:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\delta t^2\mathbf{a}(t) \tag{2.20}$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t[\mathbf{a}(t) + \mathbf{a}(t + \delta t)] \tag{2.21}$$

In order to calculate the new velocities, two sets of acceleration data are required, and the procedure can be split into three individual equations:

$$\mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t\mathbf{a}(t) \tag{2.22}$$

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \mathbf{v}(t + \frac{1}{2}\delta t)\delta t \tag{2.23}$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t + \frac{1}{2}\delta t) + \frac{1}{2}\delta t[\mathbf{a}(t + \delta t)] \tag{2.24}$$

The velocity Verlet method makes use of half steps to calculate the final velocity. First, the half step velocity is calculated using the acceleration and the velocity from time, t.

Secondly, the atomic positions at time $t + \delta t$ are calculated using the current positions and the half step velocity, then finally, using the forces calculated at the new position the half step velocity is 'upgraded' to a full step.[154] Other popular integrators that take a similar form include the leap-frog[166] algorithm and Beeman's algorithm.[167] Newer integrators based on Langevin dynamics[168] are becoming more popular as they can be used to control system temperature *in situ* and have been shown to more accurately sample the intended ensemble with low error owing to integrator precision. See Section 2.1.3.2 for a full definition.

2.1.3 Practical Considerations for MD Simulations

While the above integrators are sufficient at sampling the microcanonical (NVE) ensemble, where the number of particles, volume and energy are held held constant, modifications and additions are needed in order to sample the canonical ensemble (NVT) and the isothermal–isobaric ensemble (NPT). The temperature and pressure need to be regulated by adjusting particle velocity and/or system volume respectively. Further details on these ensembles are found in the following Section (2.2).

2.1.3.1 Temperature Regulation

For systems performed under constant temperature (NVT, NPT and μVT), a thermostat is required to control the temperature of the system by modifying the velocities of the particles. The average kinetic energy of the system is directly related to the system temperature (Eq. 2.25) showing that appropriately altering the particles' velocities is a viable route to regulating the system temperature, T.

$$\left\langle \sum_{i=1}^{N} \frac{m_i \mathbf{v}_i^2}{2} \right\rangle = \frac{3}{2} N k_B T \tag{2.25}$$

where k_B is the Boltzmann constant.

Two such methods of temperature regulation using the theory outlined in Equation 2.25 are "Velocity Rescale"[169] and the "Berendsen Thermostat".[170] The former, and most simple, rescales all of the velocities in the system at each integration timestep by a factor of λ, given by $\lambda = \sqrt{T_{\text{target}}/T(t)}$, where T_{target} is the desired temperature and $T(t)$ is the observed temperature at time, t. Although conceptually simple, the velocity rescale algorithm is very rigid and does not allow for temperature fluctuations. The Berendsen thermostat is slightly more complex but does allow for some degree of fluctuation in the temperatures. The system is considered to be weakly

coupled to an external heat source at a temperature, T_{target}, which gives/takes thermal energy to or from the system when appropriate. The scaling factor therefore becomes:

$$\lambda = \sqrt{1 + \frac{\delta t}{\tau}\left(\frac{T_{\text{target}}}{T(t)} - 1\right)} \tag{2.26}$$

where τ controls how often the system temperature converges to the external bath temperature. However, like the velocity rescale algorithm, the Berendsen thermostat has been shown to incorrectly sample the canonical ensemble and it has been recommended that all use of the Berendsen thermostat be discontinued.[171]

Methods which better sample the canonical ensemble are more complex and include the stochastic "Anderson thermostat"[172] and the deterministic "Nosé-Hoover thermostat".[173,174] The Anderson thermostat builds on the concept of an external heat bath, however, in this case, the heat bath randomly emits 'thermal particles' into the system which subsequently collide with the atoms. In practice, this is implemented by selecting a random particle and reassigning its velocity to a random selection from the Maxwell-Boltzmann distribution. Each collision is simulated in the microcanonical ensemble (NVE) in order to transport the system from its original state to the updated state without disrupting the system's energy distribution. This, however, leads to a non-smooth trajectory which is one drawback of this method. Another consideration is the collision frequency, too low and the system will not correctly sample the canonical distribution and too high can lead to little temperature fluctuation as well as becoming computationally expensive.

The Nosé-Hoover thermostat is a deterministic approach that adopts an extended system method where the heat bath is considered an integral part of the system. The thermostat introduces an extra degree of freedom, s, and has a potential energy of $V = (f+1)k_B T_{\text{target}} \ln s$, where f is the number of degrees of freedom in the system and T_{target} is the desired temperature. The bath also has kinetic energy given by $\frac{Q}{2}(\frac{ds}{dt})^2$ where Q can be considered as the 'fictitious' mass of the new degree of freedom and has units of energy $\times$ time^2. The magnitude of Q is what determines the level of coupling between the system and the heat bath and as such influences the temperature fluctuations. It is suggested that Q should be proportional to fk_BT, where the proportionality constant can be obtained by performing a series of test simulations, however, this can become tiresome.

2.1.3.2 Langevin Dynamics

Another method to regulate temperature, and the method used in these studies, is to use Langevin dynamics where the temperature control is implemented into the integrator itself.[168,175,176] Langevin dynamics considers the heat bath as an implicit

solvent that directly interacts with the system. In Langevin dynamics, the force acting on a particle, i, now has a stochastic component such that Equation 2.4 becomes:

$$\mathbf{F_i} = -\nabla U_i - \gamma \mathbf{v_i} + \sqrt{2\gamma k_B T}\mathcal{N} \tag{2.27}$$

where γ is a friction coefficient and $\mathcal{N}$ represents a random number selected from a normal distribution. The friction coefficient determines the viscosity of the implicit solvent and controls the balance between deterministic and stochastic. A coefficient of zero will yield fully deterministic dynamics with poor temperature control and when the coefficient is infinitely high, the system will follow Brownian dynamics.[177]

Various Langevin integrators have been proposed. One of which is the popular BAOAB integrator (also known as VRORV):

$$\mathbf{v}(t + \frac{1}{2}\delta t) = \mathbf{v}(t) - \frac{1}{2m}\delta t \nabla U(\mathbf{r}(t)) \tag{2.28}$$

$$\mathbf{r}(t + \frac{1}{2}\delta t) = \mathbf{r}(t) + \frac{1}{2}\delta t \mathbf{v}(t + \frac{1}{2}\delta t) \tag{2.29}$$

$$\mathbf{v}^*(t + \frac{1}{2}\delta t) = \frac{1}{m}e^{-\gamma\delta t}\mathbf{v}(t + \frac{1}{2}\delta t) + (\frac{k_B T}{m}(1 - e^{-2\gamma\delta t}))^{\frac{1}{2}}\mathcal{N}(t) \tag{2.30}$$

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t + \frac{1}{2}\delta t) + \frac{1}{2}\delta t \mathbf{v}^*(t + \frac{1}{2}\delta t) \tag{2.31}$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}^*(t + \frac{1}{2}\delta t) - \frac{1}{2m}\delta t \nabla U(\mathbf{r}(t + \delta t)) \tag{2.32}$$

Here, steps 1 and 5 (the B steps) are deterministic updates to the velocity (V), steps 2 and 4 (A steps) are deterministic position updates (R) and finally the third step (O) is a stochastic update to the system velocity ($\mathbf{v}^*$). Other Langevin integrators exist but differ in the order in which these steps are evaluated meaning that the behaviour of each integrator is different. It has been shown that the BAOAB integrator samples the canonical ensemble particularly well with a high level of accuracy.[178]

2.1.3.3 Pressure Regulation

Simulations performed at constant pressure (NPT) require the use of a barostat to maintain the pressure by modifying the system volume by a scaling factor of λ. Incidentally, as the system volume is scaled the particle coordinates must also be scaled by a factor of $\lambda^{\frac{1}{3}}$ to represent the expansion or compression of the simulation box. Similar to temperature regulation, there are barostats which are analogous to the rescale, Berenden and Anderson thermostats.[170,172] However, these methods, like for temperature, do not correctly sample fluctuations in the system pressure and struggle to correctly sample the NPT ensemble.

The Monte Carlo barostat[179,180] uses a Monte Carlo algorithm that relies on repeated random sampling to generate a numerical result. This type of sampling can be used to maintain system pressure by proposing random changes in the system volume at regular intervals. These changes are then accepted or rejected based on acceptance criteria. For a system of N particles in a box with a volume, V, and scaled coordinates, $\mathbf{s}^N$, a change in the volume, ΔV, is suggested to get a new volume, $V_{\text{new}} = V + \Delta V$. Assuming that ΔV is selected from a uniform distribution centred on zero, the acceptance ratio is as follows:

$$\begin{aligned}
\frac{A\left(V_{\text{new}} \mid V\right)}{A\left(V \mid V_{\text{new}}\right)} &= \frac{\pi_{\text{NPT}}\left(\mathbf{s}^N; V_{\text{new}}\right)}{\pi_{\text{NPT}}\left(\mathbf{s}^N; V\right)} \\
&= \frac{Z_{\text{NPT}}^{-1}\beta P \Lambda^{-3N}(N!)^{-1} V_{\text{new}}^N e^{-\beta P V_{\text{new}}} e^{-\beta U\left(\mathbf{s}^N; V_{\text{new}}\right)} \mathrm{d}\mathbf{s}^N \, \mathrm{d}V}{Z_{NPT}^{-1}\beta P \Lambda^{-3N}(N!)^{-1} V^N e^{-\beta P V} e^{-\beta U(\mathbf{s}^N; V)} \mathrm{d}\mathbf{s}^N \, \mathrm{d}V} \\
&= \left(\frac{V_{\text{new}}}{V}\right)^N e^{-\beta P\left(V_{\text{new}} - V\right)} e^{-\beta \Delta U} \\
&= \left(\frac{V_{\text{new}}}{V}\right)^N e^{-\beta P \Delta V} e^{-\beta \Delta U}
\end{aligned} \tag{2.33}$$

where $A(y|x)$ is the acceptance probability of a move to y from x, π_{NPT} is the probability of the NPT microstate, Z_{NPT} is the isothermal-isobaric partition function (Sec. 2.2), Λ is the thermal wavelength of a particle, $U\left(\mathbf{s}^N; V\right)$ is the potential energy of a particle at the specified volume and finally ΔU is the difference in potential energy caused by the change in volume.[181] A more in-depth theoretical outline and examples of Monte Carlo sampling are detailed in Section 2.5.

2.1.3.4 Periodic Boundary Conditions

To make our simulations as realistic as possible, rather than just simulating a single biomolecule in an isolated box, we use periodic boundary conditions (PBC). PBCs serve to reduce finite size effects, also known as boundary effects, where essentially anything outside of the simulation box appears as a vacuum making particles at the edge of the box behave strangely.

In practice, PBCs use a single simulation box (unit cell) surrounded by an infinite number of identical cells (periodic images), giving the illusion of a continuous, and more realistic, solution phase system (Fig. 2.3). Each atom in the unit cell still interacts with its nearest neighbours but this now includes atoms which may appear on the opposite side of the unit cell since they are close by in the next image. Lastly, if an atom translates out of the unit cell, it will simply reappear at the opposite side as that atoms image translates in the same way.

Simulations which employ PBCs also use the 'minimum image convention' to calculate the non-bonded interactions. This is where the interaction between two particles is taken to be between their closest images regardless of where the particle is in the unit cell.

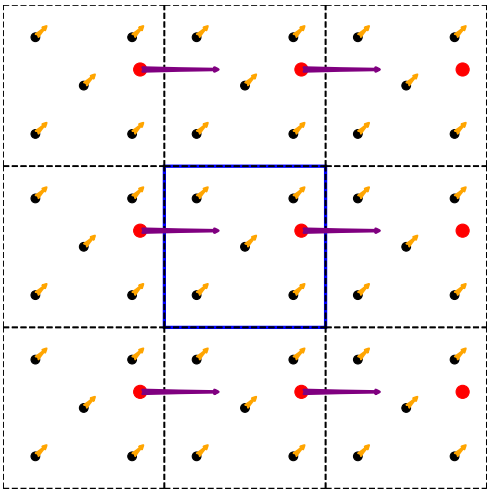


FIGURE 2.3: Periodic Boundary Conditions. The simulation unit cell is shown in the middle with its surrounding periodic images. When an atom leaves the unit cell, it reappears from the other edge as if it has been replaced by its periodic image.

2.1.3.5 Handling of Long-Range Nonbonded Interactions

In traditional molecular simulations, the forces defined by the forcefield can be categorised into short and long range interactions. The former are interactions between atoms which are close to each other and encompass all three of the bonded forces (bonds, angles, and dihedrals). The latter refers to interactions which happen at both short and long distances and encompass the two nonbonded forces, the electrostatic and Lennard Jones interactions. As an example, the charge on one atom can influence the force on another atom even at long distances, and therefore the nonbonded interactions must be calculated for every pair of atoms in the system causing the cost of calculating these potentials to scale as $\mathcal{O}(N^2)$. This calculation results in the majority of the computing cost in molecular simulations. As the distance between two particles increases, the potential energy becomes weaker. To reduce computing time, explicit calculations of these interactions are only performed for pairs of atoms within a certain distance cutoff, r_c. Of course, calculating the distances between atom pairs to evaluate whether they are within the distance cutoff also scales as $\mathcal{O}(N^2)$. Instead, simulation packages use 'neighbour lists' where for each atom there is a list of other atoms within the cutoff and only the interactions between these atoms are evaluated. The neighbour list must be updated regularly to not miss interactions. Additionally, simply truncating the interactions at a certain distance can lead to infinitely high forces and the derivative of the potential energy (the force) is no

longer smooth. To overcome this, a switching function is used whereby the interactions at distances between where the switching function is employed, r_s, and the cutoff, are scaled smoothly to zero (Fig. 2.4).

$$S = 1 - 6x^5 + 15x^4 - 10x^3 \tag{2.34}$$

where $x = (r - r_s/r_c - r_s)$.

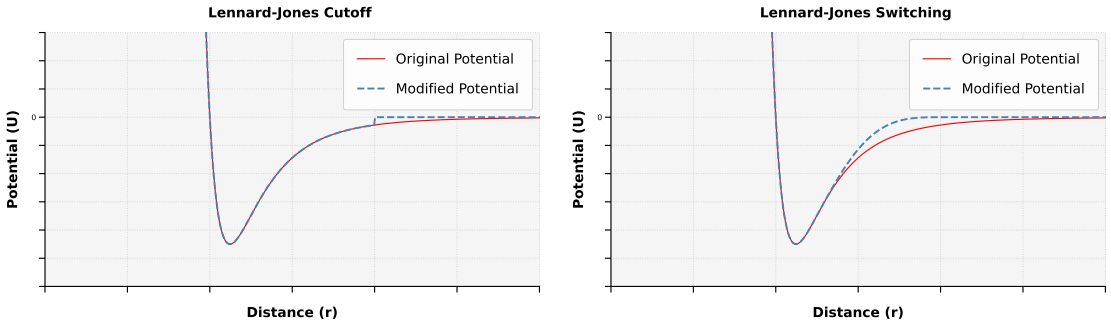


FIGURE 2.4: Left: The Lennard-Jones potential with a cutoff. At the cutoff distance, the potential becomes non-continuous leading to large forces. Right: The Lennard-Jones potential using a switching function where the potential is scaled smoothly to zero between the switching distance and the cutoff.

For the Lennard-Jones interactions, the overall potential energy of the interactions not considered beyond the cutoff can be well approximated using a dispersion correction:

$$U^{\text{vdw}}_{r>r_c} = \frac{8\pi N^2}{V}\left(\frac{\langle \epsilon_{ij}\sigma^{12}_{ij}\rangle}{9r_c^9} - \frac{\langle \epsilon_{ij}\sigma^{6}_{ij}\rangle}{3r_c^3}\right) \tag{2.35}$$

where N is the number of particles in the system, V is the system volume, r_c is the cutoff distance, ϵ_{ij} and σ_{ij} is the LJ parameters between atoms i and j and $\langle \cdots \rangle$ represents an average over all pairs of atoms in the system. This equation holds for all potentials which decay at a rate of r^{-3} or greater and therefore is only suitable for the LJ interactions which decay at a rate of r^{-6}. It should be noted, however, that the cutoff must be selected appropriately to allow for the direct summation of all the closest and most relevant interactions. Owing to the minimum image convention, which enforces that each particle interact only with the closest periodic image of another particle, when periodic boundary conditions are employed, the cutoff must not exceed more than half the length of the box to ensure an atom interacts with only one image of any other particle.

For the electrostatic interactions, which decay at r^{-1}, another solution must be employed to measure the effect of interactions beyond the cut-off. Two popular approaches are 'Reaction Field (RF)[182]' and 'Particle Mesh Ewald (PME)'.[183] Reaction field assumes that everything beyond the cutoff is a solvent with a uniform dielectric constant such that the potential energy term for the electrostatics becomes:

$$
\begin{aligned}
U^{\text{ele}} &= \frac{q_1 q_2}{4\pi\epsilon_0}\left(\frac{1}{r} + k_{rf} r^2 - c_{rf}\right) \\
k_{rf} &= \left(\frac{1}{r_c^3}\right)\left(\frac{\epsilon_{\text{solvent}} - 1}{2\epsilon_{\text{solvent}} + 1}\right) \\
c_{rf} &= \left(\frac{1}{r_c}\right)\left(\frac{3\epsilon_{\text{solvent}}}{2\epsilon_{\text{solvent}} + 1}\right)
\end{aligned}
\tag{2.36}
$$

As Equation 2.36 is only evaluated for all pairs of atoms within the cutoff, the scaling is more efficient.

The more popular particle mesh Ewald method is based on Ewald summation[154,181,184,185] and is generally regarded as more accurate than RF. To improve the convergence of the electrostatic calculations between atoms, Ewald summation splits the direct summation into two series which converge more rapidly using the following relationship:

$$
\frac{1}{r} = \frac{f(r)}{r} + \frac{1 - f(r)}{r} \tag{2.37}
$$

where $f(r)$ is chosen appropriately to handle the large variations in $1/r$ at short distances and the slow convergence at long r. In the Ewald method, the point charges of atoms are surrounded, or 'screened', by a neutralising Gaussian charge distribution of equal, but opposite, charge. At large distances, r, these functions rapidly decrease to zero. Eventually, to correct for the added Gaussian distributions, a second set of Gaussian charges are added to exactly cancel out the first set and are evaluated in Fourier space.

For Ewald summation, the calculation can be split into three terms, the direct space sum, the reciprocal space sum, and the self-energy term. The direct space describes interactions between all of the point charges and the screening Gaussian functions, the reciprocal space describes the interactions between each point charge and the second Gaussian set, and the self-energy term is included to correct for the interaction between the point charge and the Gaussian charge on a single atom.

$$
\begin{aligned}
U^{\text{ele}} &= U^{\text{ele}}_{dir} + U^{\text{ele}}_{rec} + U^{\text{ele}}_{self} \\
U^{\text{ele}}_{dir} &= \frac{1}{2}\sum_{i,j} q_i q_j \frac{\operatorname{erfc}(\alpha r_{ij})}{r_{ij}} \\
U^{\text{ele}}_{rec} &= \frac{1}{2\pi V}\sum_{i,j} q_i q_j \sum_{\mathbf{k}\neq 0} \frac{\exp(-(\pi\mathbf{k}/\alpha)^2 + 2\pi i\mathbf{k}\cdot(\mathbf{r}_i - \mathbf{r}_j))}{\mathbf{k}^2} \\
U^{\text{ele}}_{self} &= -\frac{\alpha}{\sqrt{\pi}}\sum_i q_i^2,
\end{aligned}
\tag{2.38}
$$

where

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-x^2} dx. \tag{2.39}$$

α defines the width of the Gaussians, V is the simulation volume, $\mathbf{k}$ are vectors in the Fourier space, $\mathbf{r}_i$ is the position vector of atom i. Crucially, the direct summation now converges much more rapidly and is ignored beyond the cutoff. The reciprocal calculations are performed in the Fourier space and are even quicker. As a result Ewald summation scales $\mathcal{O}(N^{3/2})$.

Finally, particle mesh Ewald builds on the Ewald summation but rather than calculating the reciprocal space sum directly, it first distributes the particle charges onto a rectangular lattice where the Fourier vectors can be evaluated using the Fast-Fourier Transform method further improving the scaling to $\mathcal{O}(N \ln N)$.

While traditionally PME is preferred to RF, a recent study by Ge *et al.*[186] compares the two methods in the context of relative free energy calculations showing very little difference in the results with RF showing greater efficiency. The conclusions suggest that RF be considered more often.

2.2 Statistical Mechanics

In the previous section, it was shown that particles within a system can be simulated using classical mechanics, in that knowledge of the past and present atomic positions and velocities allows us to calculate the future evolution of the system. However, such systems are simulated at the nano- and microscopic scale and do not correspond to the real, macroscopic world. The use of classical statistical mechanics and ensemble theory is what allows us to link microscopic configurations (microstates) to macroscopic observables such as those measured in real experiments.

An ensemble can be thought of as many closed microstates in thermal contact with one another and as such can exchange energy. The 'type' of the ensemble is then defined by the properties which remain constant between all microstates. For example, the canonical ensemble (NVT) fixes the number of particles in the microstate (N), the volume of the state (V) and the temperature (T).

Of particular interest to us, is sampling the equilibrium probability of a particular ensemble. This in turn allows us to correctly calculate an ensemble average of any hypothetical property, A, that can be accurately compared to macroscopic observables (Eq.2.40):

$$\langle A \rangle = \iint A(\mathbf{r}^N, \mathbf{p}^N) \rho(\mathbf{r}^N, \mathbf{p}^N) d\mathbf{r}^N d\mathbf{p}^N \tag{2.40}$$

where $\langle A \rangle$ represents the ensemble average of property A, $A(\mathbf{r}^N, \mathbf{p}^N)$ is the value of A for a microstate with positions $\mathbf{r}^N$ and momenta $\mathbf{p}^N$. $\rho(\mathbf{r}^N, \mathbf{p}^N)$ is the equilibrium probability density of that microstate, given by a Boltzmann distribution. An overview of two common ensembles and one less common ensemble is given in the following sections.[154,181,187]

2.2.1 Canonical Ensemble

The canonical (NVT) ensemble is one of the most simple and fixes the number of particles (N), volume (V) and temperature (T). An NVT system can be thought of as being in thermal contact with an external heat source with which the system can exchange energy in the form of heat to maintain a constant temperature. The free energy of an NVT system is referred to as the Helmholtz energy (F) and is given by:[154,181,187]

$$F = -k_B T \ln Q_{NVT} \tag{2.41}$$

where k_B is the Boltzmann constant and Q_{NVT} is the canonical partition function representing the microstates in the ensemble and is given as a sum of all energies of each microstate, E_i:

$$Q_{NVT} = \sum_i e^{-\beta E_i} \tag{2.42}$$

Within our classical based simulations, the total energy is a function of both atomic position (potential) and momenta (kinetic) which are both continuous functions. As such, the canonical partition function can be treated as an integral over all microstates with respect to atomic positions and momenta:

$$Q_{NVT} = \frac{1}{h^{3N} N!} \iint e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} d\mathbf{r}^N d\mathbf{p}^N, \tag{2.43}$$

where h is Planck's constant. When dealing with identical particles, the $(N!)^{-1}$ term is required to prevent over counting microstates with the same configurations. It follows that the potential energy and kinetic energy are separable such that the total energy is given by:

$$E(\mathbf{r}^N, \mathbf{p}^N) = U(\mathbf{r}^N) + \sum_{i=1}^{N} \frac{|\mathbf{p}_i|^2}{2m}, \tag{2.44}$$

where $U(\mathbf{r}^N$ is the potential energy function usually described by a forcefield and m is the mass of particle i. It follows that the integrals of both terms can also be calculated separately and the exponentiated kinetic term yields a Gaussian function which can be

integrated analytically:

$$\int_{-\infty}^{+\infty} \exp\left\{-\sum_{i=1}^{N} \frac{|\mathbf{p_i}|^2}{2mk_BT}\right\} d\mathbf{p}^N = (2\pi mk_BT)^{3N/2}. \tag{2.45}$$

The canonical partition function can now be simplified to yield only an integral of the potential energy with respect to the atomic positions.

$$Q_{NVT} = \frac{1}{\Lambda^{3N}N!} \int e^{-\beta U(\mathbf{r}^N)} d\mathbf{r}^N, \tag{2.46}$$

where Λ is the thermal wavelength of a particle, depending on its mass, m, and the temperature, T.

$$\Lambda = \left(\frac{h^2}{2\pi mk_BT}\right)^{\frac{1}{2}} \tag{2.47}$$

It is sometimes useful, and will become obvious later, to rewrite the partition function in terms of scaled coordinates, $\mathbf{s}^N$, where the particle positions are scaled to fit between 0 and 1:

$$Q_{NVT} = \frac{V^N}{\Lambda^{3N}N!} \int_0^1 e^{-\beta U(\mathbf{s}^N;V)} d\mathbf{s}^N \tag{2.48}$$

Having made this rearrangement, the partition function can now be seen as a product of ideal and excess components where the ideal partition function, Q^{id}_{NVT}, is equivalent to the partition function of an ideal gas, and the excess component, Q'_{NVT}, accounts for the contribution of intermolecular interactions.

$$Q^{id}_{NVT} = \frac{V^N}{\Lambda^{3N}N!} \tag{2.49}$$

$$Q'_{NVT} = \int_0^1 e^{-\beta U(\mathbf{s}^N;V)} d\mathbf{s}^N \tag{2.50}$$

It also follows that the Helmholtz free energy can also be split into its ideal and excess components such that:

$$\begin{aligned} F &= -k_BT \ln Q^{id}_{NVT} - k_BT \ln Q'_{NVT} \\ &= F^{id} + F' \end{aligned} \tag{2.51}$$

To calculate an ensemble average of a property, $\langle A \rangle$, we must calculate the equilibrium probability density of observing a given microstate:[154,181,187]

$$\rho_{NVT}(\mathbf{r}^N, \mathbf{p}^N) = Q^{-1}_{NVT} \frac{1}{h^{3N}N!} e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} \tag{2.52}$$

As we are only interested in a particular configuration of particles that define an instantaneous microstate, we can separate and integrate out the momentum

contribution such that the probability density is only dependent on the potential energy:

$$\begin{aligned}\rho_{NVT}(\mathbf{r}^N) &= Q_{NVT}^{-1}\frac{1}{h^{3N}N!}e^{-\beta E(\mathbf{r}^N)}\int_{-\infty}^{+\infty}\exp\left\{-\sum_{i=1}^{N}\frac{|\mathbf{p_i}|^2}{2mk_BT}\right\}d\mathbf{p}^N \\ &= Q_{NVT}^{-1}\frac{1}{\Lambda^{3N}N!}e^{-\beta U(\mathbf{r}^N)} \\ &= \frac{e^{-\beta U(\mathbf{r}^N)}}{\int e^{-\beta U(\mathbf{r}^N)}d\mathbf{r}^N}\end{aligned} \tag{2.53}$$

The ensemble average of an observable, A, in the canonical ensemble, can then be calculated as:

$$\langle A\rangle_{NVT} = \frac{\int A(\mathbf{r}^N)e^{-\beta U(\mathbf{r}^N)}d\mathbf{r}^N}{\int e^{-\beta U(\mathbf{r}^N)}d\mathbf{r}^N} \tag{2.54}$$

Finally, it is useful to define the following partial derivatives which relate the Helmholtz free energy to different thermodynamic properties of the system:

$$\left(\frac{\partial F}{\partial N}\right)_{V,T} = \mu \tag{2.55}$$

$$\left(\frac{\partial F}{\partial V}\right)_{N,T} = -P \tag{2.56}$$

$$\left(\frac{\partial F}{\partial T}\right)_{N,V} = -S \tag{2.57}$$

where μ is the chemical potential, P is the pressure, and S is the entropy of the ensemble. Subscripts indicate parameters which are held constant.

2.2.2 Isothermal-Isobaric Ensemble

The isothemal-isobaric (NPT) ensemble is conceptually similar to the canonical ensemble except it allows for simulations to be carried out at constant pressure (P) which is often more akin to the physiological conditions at which experiments are performed making this ensemble more relevant for the biological systems we simulate. To maintain pressure, the system can be thought of as being in contact with an ideal gas of a fixed pressure such that a hypothetical piston can increase or decrease the volume of the system in order to maintain the pressure of the ideal gas. In this ensemble, the probability of a microstate becomes dependent on both the system volume and pressure, and therefore the partition function includes an integration over

all possible volumes as well as atomic positions and momenta.[154,181,187]

$$Z_{NPT} = \frac{\beta P}{h^{3N} N!} \iiint e^{-\beta PV} e^{\beta E(\mathbf{r}^N, \mathbf{p}^N)} d\mathbf{r}^N d\mathbf{p}^N dV. \tag{2.58}$$

As with the canonical ensemble, the partition function can be simplified by scaling the coordinates and integrating out the momenta:

$$Z_{NPT} = \frac{\beta P}{\Lambda^{3N} N!} \int V^N e^{-\beta PV} dV \int e^{-\beta U(\mathbf{s}^N; V)} d\mathbf{s}^N \tag{2.59}$$

In the isothermal-isobaric ensemble, the Gibbs free energy of a system can be calculated as:

$$G = -k_B T \ln Z_{NPT} \tag{2.60}$$

The probability density of a given microstate with positions, $\mathbf{r}^N$, in a defined volume, V, is given by:

$$\rho_{NPT}(\mathbf{r}^N, V) = Z_{NPT}^{-1} \frac{\beta P}{\Lambda^{3N} N!} e^{-\beta PV} e^{-\beta U(\mathbf{r}^N)} \tag{2.61}$$

and with scaled coordinates:

$$\rho_{NPT}(\mathbf{r}^N, V) = Z_{NPT}^{-1} \frac{\beta P V^N}{\Lambda^{3N} N!} e^{-\beta PV} e^{-\beta U(\mathbf{s}^N; V)} \tag{2.62}$$

The ensemble average of an observable in the isothermal-isobaric ensemble is given by:

$$\langle A \rangle_{NPT} = Z_{NPT}^{-1} \frac{\beta P}{\Lambda^{3N} N!} \iint A(\mathbf{s}^N; V) V^N e^{-\beta PV} e^{-\beta U(\mathbf{s}^N; V)} dV d\mathbf{s}^N \tag{2.63}$$

Finally, the following partial derivatives hold for the NPT ensemble:

$$\left(\frac{\partial G}{\partial N}\right)_{P,T} = \mu \tag{2.64}$$

$$\left(\frac{\partial G}{\partial P}\right)_{N,T} = V \tag{2.65}$$

$$\left(\frac{\partial G}{\partial T}\right)_{N,P} = -S \tag{2.66}$$

2.2.3 Grand Canonical Ensemble

The grand canonical (μVT) ensemble is conceptually the most interesting and important to this work as this is the ensemble in which most of our simulations are performed. The μVT ensemble is distinctive with respect to the NVT and NPT ensembles in that the number of particles is not held fixed. Instead, the chemical potential, μ, is maintained. This ensemble can be considered as an open system in contact with an ideal gas with which the system can exchange particles (Figure 2.5).

Given its importance in this work, a full derivation of the grand canonical ensemble is given below.[149,181,187]

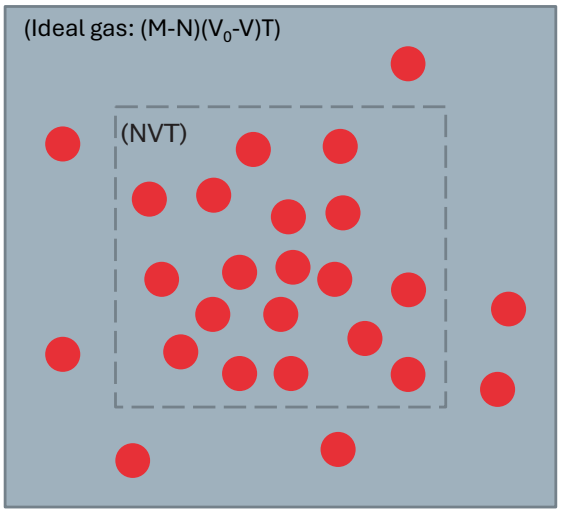


FIGURE 2.5: Graphical depiction of the grand canonical ensemble. Particles are free to move between the ideal gas and the central NVT system. At equilibrium, the chemical potential of the coupled systems is equal.

The grand canonical ensemble is similar to the NVT ensemble in that both volume and temperature are kept constant. Accordingly, both the system and the ideal gas reservoir can be considered as two individual canonical ensembles. The total system (ideal gas + system) contains M particles in a volume, V_{tot}, while the simulated system contains N particles in a volume, V_{sys}. This means the ideal gas reservoir has G particles in a volume V_{gas} where $G = M - N$ and $V_{gas} = V_{tot} - V_{sys}$. If we first consider the complete system where there is no exchange of particles between the ideal gas and the simulated volume, we can write the canonical partition function of the combined system:

$$Q_{MVT} = \frac{1}{h^{3M} M!} \iint e^{-\beta E(\mathbf{r}^M, \mathbf{p}^M)} d\mathbf{r}^M d\mathbf{p}^M \tag{2.67}$$

Since the two systems do not interact nor exchange particles, the energy, or Hamiltonian, of the system can be split such that:

$$E(\mathbf{r}^M, \mathbf{p}^M) = E(\mathbf{r}^N, \mathbf{p}^N) + E(\mathbf{r}^G, \mathbf{p}^G) \tag{2.68}$$

It follows that the partition function can also be split such that:

$$Q_{MVT} = \frac{1}{h^{3M} M!} \iint e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} d\mathbf{r}^N d\mathbf{p}^N \iint e^{-\beta E(\mathbf{r}^G, \mathbf{p}^G)} d\mathbf{r}^M d\mathbf{p}^G \tag{2.69}$$

The individual partition functions for the separated systems are given by:

$$Q_{NV_{sys}T} = \frac{1}{h^{3N}N!}\iint e^{-\beta E(\mathbf{r}^N,\mathbf{p}^N)}d\mathbf{r}^N d\mathbf{p}^N \tag{2.70}$$

$$Q_{GV_{gas}T} = \frac{1}{h^{3G}G!}\iint e^{-\beta E(\mathbf{r}^G,\mathbf{p}^G)}d\mathbf{r}^G d\mathbf{p}^G \tag{2.71}$$

and can be substituted into Equation 2.69:

$$\begin{aligned} Q_{MVT} &= \frac{1}{h^{3M}M!}Q_{NV_{sys}T}(h^{3N}N!)Q_{GV_{gas}T}(h^{3G}G!) \\ &= \frac{N!G!}{M!}Q_{NV_{sys}T}Q_{GV_{gas}T} \end{aligned} \tag{2.72}$$

From this result, we must now consider all the possible ways in which the M number of particles can be arranged into the groups N and $M-N$. Currently, as written, for a non-exchanging system the right hand side of Equation 2.72 represents only one possible arrangement of the particles between the two systems. The partition function now becomes a sum over all possible configurations:

$$Q_{MVT} = \sum_{N=0}^{M} g(N,G)\frac{N!G!}{M!}Q_{NV_{sys}T}Q_{GV_{gas}T} \tag{2.73}$$

where we have introduced a degeneracy factor to correctly count the number of ways the particles can be distributed:

$$g(N,G) = \frac{M!}{N!G!} \tag{2.74}$$

Substituting this into the above gives:

$$Q_{MVT} = \sum_{N=0}^{M} Q_{NV_{sys}T}Q_{GV_{gas}T} \tag{2.75}$$

The partition function of the ideal gas is related to its Helmholtz free energy ($F_{GV_{gas}T} = -k_BT\ln Q_{GV_{gas}T}$) and given that the total number of gas particles, G, is much greater than the number of interacting particles, N, and $V_{gas} \gg V_{sys}$, we can approximate the Helmholtz free energy of the ideal gas by performing a first order Taylor expansion around the Helmholtz free energy of the total system.

$$\begin{aligned} F_{GV_{gas}T} &\approx F_{MV_{tot}T} - N\frac{\partial F_{MV_{tot}T}}{\partial N} - V_{sys}\frac{\partial F_{MV_{tot}T}}{\partial V_{tot}} \\ &\approx F_{MVT} - \mu N + PV_{sys} \end{aligned} \tag{2.76}$$

The Helmholtz free energy of the ideal gas reservoir can then replace its corresponding partition function in Equation 2.75 and with other rearrangements we find that:

$$\begin{aligned} Q_{MVT} &= \sum_{N=0}^{M} e^{-\beta(F_{MVT}-\mu N+PV_{sys})} Q_{NV_{sys}T} \\ &= \sum_{N=0}^{M} e^{-\beta F_{MVT}} e^{\beta\mu N} e^{-\beta PV_{sys}} Q_{NV_{sys}T} \\ &= e^{-\beta PV_{sys}} Q_{MVT} \sum_{N=0}^{M} e^{\beta\mu N} Q_{NV_{sys}T} \end{aligned} \tag{2.77}$$

where the result of Equation 2.77 contains Q_{MVT} on both sides and can be cancelled out to yield:

$$e^{\beta PV_{sys}} = \sum_{N=0}^{M} e^{\beta\mu N} Q_{NV_{sys}T} \tag{2.78}$$

where there is now no explicit reference to the ideal gas reservoir except for the summation. However, by assuming M is infinitely large, we can consider these equations only in terms of the overall grand canonical system by replacing V_{sys} with V_{tot}:

$$e^{\beta PV_{tot}} = \sum_{N=0}^{\infty} e^{\beta\mu N} Q_{NV_{tot}T} \tag{2.79}$$

Analogous to the Helmholtz free energy and the Gibbs free energy, the "grand potential" can be written in terms of the grand canonical partition function, $\Xi_{\mu VT}$:

$$\Omega = -k_B T \ln \Xi_{\mu VT} \tag{2.80}$$

As before, it is useful to define the following partial derivatives of the grand potential:

$$\left(\frac{\partial\Omega}{\partial\mu}\right)_{V,T} = -\langle N\rangle_{V,T} \qquad \left(\frac{\partial\Omega}{\partial V}\right)_{\mu,T} = -P \qquad \left(\frac{\partial\Omega}{\partial T}\right)_{\mu,V} = -S \tag{2.81}$$

As such, the grand potential can then be written as:

$$\begin{aligned} \Omega &= V\frac{\partial\Omega}{\partial V} \\ &= -PV \end{aligned} \tag{2.82}$$

which combined with Equation 2.80 gives the following:

$$\begin{aligned} k_B T \ln \Xi_{\mu VT} &= PV \\ \Xi_{\mu VT} &= e^{\beta PV} \end{aligned} \tag{2.83}$$

and substituting this into Equation 2.79 gives us the expression for the grand canonical partition function:

$$\Xi_{\mu VT} = \sum_{N=0}^{\infty} e^{\beta\mu N} Q_{NV_{tot}T} \tag{2.84}$$

It follows that the probability density of a microstate in the grand canonical ensemble with N particles is given by (note that as previously, the momenta is integrated out):

$$\rho_{\mu VT}(\mathbf{r}^N) = \Xi_{\mu VT}^{-1} \frac{e^{\beta\mu N}}{\Lambda^{3N} N!} e^{-\beta U(\mathbf{r}^N)} \tag{2.85}$$

and as such the ensemble average for an observable, A, is given by:

$$\langle A \rangle_{\mu VT} = \Xi_{\mu VT}^{-1} \sum_{N=0}^{\infty} \frac{e^{\beta\mu N}}{\Lambda^{3N} N!} \int A(\mathbf{r}^N) e^{-\beta U(\mathbf{r}^N)} d\mathbf{r}^N \tag{2.86}$$

2.3 Chemical Potential

When defining the grand canonical ensemble, we refer to the chemical potential, μ, which is kept constant in the ensemble. The chemical potential of a species is the energy which is either absorbed or released due to a change in the particle number of said species. In more abstract terms, it defines the flow of particles where, in general, particles move to areas with a lower chemical potential to reduce the overall free energy.[181,187] To relate the chemical potential to the grand canonical ensemble we must first define a Legendre transformation of the Helmholtz free energy such that:

$$\begin{aligned} \Omega &= F - N \left(\frac{\partial F}{\partial N} \right)_{V,T} \\ &= F - N\mu \end{aligned} \tag{2.87}$$

where, using the product rule, the derivative can be defined as:

$$d\Omega = dF - Nd\mu - \mu dN. \tag{2.88}$$

Using the partial derivatives defined above, we can also say a small change in the grand potential can be expressed as:

$$\begin{aligned} d\Omega &= \left(\frac{\partial \Omega}{\partial \mu} \right)_{V,T} d\mu + \left(\frac{\partial \Omega}{\partial V} \right)_{\mu,T} dV + \left(\frac{\partial \Omega}{\partial T} \right)_{\mu,V} dT \\ &= -Nd\mu - PdV - SdT. \end{aligned} \tag{2.89}$$

Substituting Equation 2.89 into Equation 2.88 we get a value for dF:

$$\begin{aligned} dF &= d\Omega + Nd\mu + \mu dN \\ dF &= -PdV - SdT - \mu dN. \end{aligned} \tag{2.90}$$

If we fix the temperature and the volume, we can relate the chemical potential to the Helmholtz free energy as a partial derivative with respect to particle number, in turn giving an indication as to whether the addition or removal of a particle is favourable or not:

$$\mu = \left(\frac{\partial F}{\partial N}\right)_{V,T} \tag{2.91}$$

It was shown in Equation 2.51 that the Helmholtz energy can be written in terms of its ideal and excess contributions and, as such, so can the chemical potential:

$$\begin{aligned} \mu &= \frac{\partial}{\partial N}(F^{id} + F') \\ &= \frac{\partial F^{id}}{\partial N} + \frac{\partial F'}{\partial N} \\ &= \mu^{id} + \mu' \end{aligned} \tag{2.92}$$

where μ^{id} is the ideal chemical potential, and μ' is the excess chemical potential that arises due to interactions between particles. It is worth reminding that a particle in the ideal gas does not interact with other particles and as such will only have an ideal component.

2.3.1 Ideal Chemical Potential

As shown in Equation 2.49, the ideal canonical partition function, Q^{id}, can be calculated analytically given the volume and the number of particles in the system.[181,187] As such the ideal contribution to the Helmholtz free energy can be calculated analytically:

$$\begin{aligned} F^{id}(N) &= -k_B T \ln Q^{id}_{NVT} \\ &= -k_B T \ln\left(\frac{V^N}{\Lambda^{3N} N!}\right). \end{aligned} \tag{2.93}$$

In an ideal gas, the value of N is infinitely large meaning we can use Stirling's approximation to remove the factorials in the above equation:

$$F^{id}(N) \approx -k_B T\left(N \ln\left(\frac{V}{\Lambda^3}\right) - N \ln N + N\right) \tag{2.94}$$

the derivative of which, leads us to an equation for the ideal chemical potential:

$$\mu^{id} = \frac{\partial F^{id}}{\partial N} = -k_B T \ln\left(\frac{V}{N\Lambda^3}\right) \tag{2.95}$$

which can finally be rewritten in terms of the number density, ρ_{ideal}, for convenience:

$$\mu^{id} = k_B T \ln(\rho_{ideal}\Lambda^3), \tag{2.96}$$

where

$$\rho_{ideal} = \frac{N}{V}. \tag{2.97}$$

2.3.2 Excess Chemical Potential

The excess chemical potential can be calculated from the excess contribution to the canonical partition function (Eq. 2.50) via the excess Helmholtz free energy:

$$\begin{aligned} F'(N) &= -k_B T \ln Q'_{NVT} \\ &\approx -k_B T \ln\left\{\int_0^1 e^{-\beta U(\mathbf{s}^N;V)} d\mathbf{s}^N\right\}, \end{aligned} \tag{2.98}$$

however, the configurational integral above is impossible to solve analytically for all but the simplest of systems. For the large and complex systems that we are interested in, the excess free energy (and thus chemical potential) must be calculated numerically. The Widom particle insertion method is one approach which is conceptually simple and forms the basis for more complex calculations. Simply, the method involves assuming that the excess chemical potential of a system with a large number of particles, N, can be approximated by rewriting the partial derivative in Equation 2.92 as a finite difference derivative:[188]

$$\mu' \approx \frac{\Delta F'}{\Delta N} \tag{2.99}$$

from this, we can take the smallest possible value for ΔN to be equal to 1, meaning that the excess chemical potential can be calculated as the excess free energy of adding one additional particle to the system:

$$\begin{aligned} \mu' = F'(N+1) - F'(N) &= -k_B T \ln \frac{Q'_{(N+1)VT}}{Q'_{NVT}} \\ &= -k_B T \ln\left\{\frac{\int e^{-\beta U(\mathbf{s}^{N+1};V)} d\mathbf{s}^{N+1}}{\int e^{-\beta U(\mathbf{s}^N;V)} d\mathbf{s}^N}\right\} \end{aligned} \tag{2.100}$$

This equation can be simplified further by noting that the potential energy of the $N+1$ system is equal to the sum of the energy of the N system and the additional

energy of the excess particle interacting with the N system:

$$\begin{aligned} \Delta U_{N+1} &= U(\mathbf{s}^{N+1}; V) - U(\mathbf{s}^{N}; V) \\ U(\mathbf{s}^{N+1}; V) &= U(\mathbf{s}^{N}; V) + \Delta U_{N+1} \end{aligned} \tag{2.101}$$

where ΔU_{N+1} is the potential energy of the $(N+1)^{th}$ particle with all the other particles in the system. This means that Equation 2.100 can be simplified:

$$\mu' = -k_B T \ln \left\{ \int \langle e^{-\beta \Delta U_{N+1}} \rangle_N d\mathbf{s}_{N+1} \right\} \tag{2.102}$$

Importantly, the integral is now only carried out over the positions of just the $(N+1)^{th}$ particle. Note that the chemical potential is now also shown as an ensemble average, $\langle \cdots \rangle_N$, over the system containing N particles, such that it can now be sampled in a computer simulation.

In practice, the Widom particle insertion method solves the integral using a brute force Monte Carlo scheme whereby many configurations of the N-particle system are generated according to its equilibrium probability, then, at regular intervals, the $(N+1)$ particle is randomly placed in the system and Equation 2.102 is evaluated by measuring the change in potential energy, ΔU_{N+1}. The resulting values are averaged to calculate the excess chemical potential.[181]

While the theory of the Widom method is useful, practically the method is very inefficient in dense biomolecular systems. This is because many of the random insertions will result in a large potential energy owing to steric clashes. It is therefore recommended that excess chemical potential is calculated using alchemical methods (Section 2.4) which give much better phase-space overlap.

2.4 Methods of Calculating Binding Free Energy

In these studies, understanding the binding of small molecules to a receptor is crucial. A few computational methods to calculate binding affinity exist with alchemical free energy calculations, also known as Free Energy Perturbation (FEP), being one of the most widely adopted. Alchemical calculations involve intermediate non-physical 'alchemical' states in which the interactions of parts of a system (typically a small molecule) with their environment are modified. These so-called alchemical states bridge two different physical states. In this section, we will first discuss the statistical thermodynamics of ligand binding followed by a deeper understanding of how binding affinity is calculated in simulation using alchemical methods.

2.4.1 Statistical Thermodynamics of Protein-Ligand Binding

Throughout this section, the notation P will represent a protein/receptor while L refers to a ligand. The reversible reaction of interest is the binding of a ligand to a receptor which is given by:

$$P + L \leftrightharpoons PL. \tag{2.103}$$

These binding events are an equilibrium process with an equilibrium constant, K_b, which is also known as the binding constant or dissociation constant for the reverse process. It follows that the condition for equilibrium in a solution is given by a sum of the species' chemical potentials:

$$\mu_{sol,P} + \mu_{sol,L} = \mu_{sol,PL}, \tag{2.104}$$

where $\mu_{sol,i}$ is the chemical potential of species i in solution and is given by:

$$\mu_{sol,i} = \mu^o_{sol,i} + k_B T \ln \frac{\gamma_i C_i}{C^o}, \tag{2.105}$$

where $\mu^o_{sol,i}$ and C_i are the standard chemical potential and concentration of species i respectively, k_B is the Boltzmann constant, T is temperature and γ_i is the activity coefficient of i. C^o is equal to 1 M and is included to cancel the units of C_i. The standard chemical potential defines the chemical potential of a species when in a hypothetical non-interacting standard state at a concentration of C^o.

It can then be said that the standard free energy of binding of L and P to give PL is equal to the difference of the chemical potentials and as such rearrangement of Equation 2.104 and substitution of Equation 2.105 leads to the following expression:

$$\begin{aligned} \Delta G^o_{PL} &\equiv \mu_{sol,PL} - \mu_{sol,P} - \mu_{sol,L} \\ &= -k_B T \ln \left(\frac{\gamma_{PL}}{\gamma_P \gamma_L} \frac{C^o C_{PL}}{C_P C_L} \right)_{eq} \equiv -k_B T \ln K^o_b \end{aligned} \tag{2.106}$$

At low concentrations, like those used in simulation, to a good approximation $\gamma_i = 1$ therefore Equation 2.106 simplifies to:

$$\Delta G^o_{PL} = -k_B T \ln \left(C^o \frac{C_{PL}}{C_P C_L} \right), \tag{2.107}$$

giving an expression for the standard binding constant as:

$$K^o_b = C^o \frac{C_{PL}}{C_P C_L}. \tag{2.108}$$

The concentrations of the bound and unbound states are related to the probability of finding the system in said state and therefore the standard free energy can be written

as:

$$\Delta G^o_{PL} = -k_B T \ln \frac{P(PL)}{P(P+L)} \tag{2.109}$$

The probability of finding configuration or state, $\mathbf{q}$, in a set of configurations, Γ, is given by the Boltzmann probability density function:

$$P(\mathbf{q}) = \frac{\exp(\beta U(\mathbf{q}))}{\int_\Gamma \exp(\beta U(\mathbf{q}))\, d\mathbf{q}}, \tag{2.110}$$

where $\beta = (k_B T)^{-1}$, $U(\mathbf{q})$ is the potential energy of configuration $\mathbf{q}$ and the integration is over all accessible configurations. The bound and unbound configurations can then be considered a subset of all possible configurations ($\Gamma_{bound}, \Gamma_{unbound} \subset \Gamma$), such that to find the probability of just the desired configuration, Γ_{bound} and $\Gamma_{unbound}$, from all possible states we can integrate further over only these states to get:

$$P(PL) = \int_{\Gamma_{bound}} P(\mathbf{q})\, d\mathbf{q} = \frac{\int_{\Gamma_{bound}} \exp(\beta U(\mathbf{q}))\, d\mathbf{q}}{\int_\Gamma \exp(\beta U(\mathbf{q}))\, d\mathbf{q}} \tag{2.111}$$

$$P(P+L) = \int_{\Gamma_{unbound}} P(\mathbf{q})\, d\mathbf{q} = \frac{\int_{\Gamma_{unbound}} \exp(\beta U(\mathbf{q}))\, d\mathbf{q}}{\int_\Gamma \exp(\beta U(\mathbf{q}))\, d\mathbf{q}} \tag{2.112}$$

Finally, we can rewrite the probability ratio of bound and unbound states as:

$$\frac{P(PL)}{P(P+L)} = \frac{\int_{\Gamma_{bound}} \exp(\beta U(\mathbf{q}))\, d\mathbf{q}}{\int_{\Gamma_{unbound}} \exp(\beta U(\mathbf{q}))\, d\mathbf{q}} = \frac{Z(PL)}{Z(P+L)} \tag{2.113}$$

where we have defined the configurational integrals, $Z(X)$, for the bound and unbound state. A more rigorous derivation can be found in the works of Gilson *et al.*[189,190] and further defined by Mobley *et al.*[128,191] however this goes beyond the scope of this project. The key result is shown in Equation 2.114.

$$\Delta G^o_{PL} = -k_B T \ln \left[\frac{C^o}{8\pi^2} \frac{\sigma_P \sigma_L}{\sigma_{PL}} \frac{Z_{PL} Z_0}{Z_P Z_L} \right] + P^o \Delta V_{PL}, \tag{2.114}$$

where Z_0 is the configurational integral for the solvent of N atoms with no solute present, σ_i is the symmetry number for species i and lastly $P^o \Delta V_{PL}$ represents the change in equilibrium volume upon ligand binding. This final equation shows that simply running a simulation and measuring the probability of finding the system in a bound, or unbound, configuration could in principle give an estimate for the binding affinity. However, in practice simulating multiple binding and unbinding events is almost impossible. The computational cost of these calculations far exceeds what would be deemed feasible in a live drug discovery setting. For example, the rate of dissociation for typical drug molecules can reach into seconds while simulation time is often limited to microseconds for atomistic systems.[113,114,192–194]

2.4.2 Absolute Alchemical Free Energy Calculations

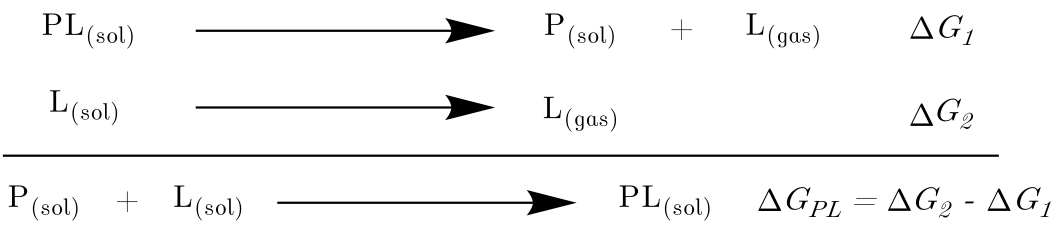


FIGURE 2.6: Basic overview of the double decoupling method. The binding free energy of a ligand to a protein is equal to the difference in the free energy of decoupling the ligand from the receptor and solvent.

Rather than simulating the actual binding event, we can exploit the fact that free energy is a state function and is therefore independent of the route taken. The double-decoupling (Figure 2.6) method yields predictions that do not require direct simulation of binding and unbinding events.[189] The double decoupling method shows that the free energy of binding can be calculated as the difference in free energy between the free energy change of decoupling the ligand from its environment in complex and in solvent. It is now useful to define a controlling parameter, λ, which controls the potential energy function ($U(\mathbf{q}, \lambda)$) such that at $\lambda = 1$ at state A and 0 at state B. In the context of double decoupling, state A can be thought of as the left hand side of the equations (PL_{sol} or L_{sol}) in Figure 2.6 and state B, as the right hand side ($P_{sol} + L_{gas}$ or L_{gas}). Following a similar logic as above, the free energy of these individual reactions can be estimated by:

$$\Delta G_{env} = -k_B T \ln \frac{Z(\lambda_B)}{Z(\lambda_A)} = -k_B T \ln \frac{\int_{\Gamma_{env}} \exp(\beta U(\mathbf{q}, \lambda_B))\, d\mathbf{q}}{\int_{\Gamma_{env}} \exp(\beta U(\mathbf{q}, \lambda_A))\, d\mathbf{q}} \tag{2.115}$$

where *env* refers to where the decoupling is occurring, in this case, complex or solvent. Generally, it is not feasible to compute these partition functions directly, instead, several estimators have been proposed and are discussed in the following section. In theory, the two free energies, ΔG_1 and ΔG_2, can be obtained by simply measuring potential energy samples of just the initial (A) and final states (B) and using the estimators to integrate between the two, however, this is often very inaccurate as the current integration methods require good phase space overlap between the two states.[191] To be precise, the sampling of state A must also reflect the Boltzmann distribution of state B which is almost impossible when simulating a bound and unbound state.

To circumvent this issue of poor overlap, alchemical intermediate states are introduced to interpolate smoothly between states A and B. A system's potential

energy, on which the configurational integrals in Equation 2.114 depend, can now be defined in terms of λ:

$$U(\lambda) = \lambda U_A + (1-\lambda)U_B. \tag{2.116}$$

It follows that the free energy change from states A to B can then be estimated as a sum of the free energy differences between each alchemical intermediate:

$$\Delta G_{env} = k_B T \sum_{k=0}^{k-1} \Delta f(\lambda_k, \lambda_{k+1}) \tag{2.117}$$

where Δf is the dimensionless free energy difference between two lambda states:

$$\Delta f(\lambda_k, \lambda_{k+1}) = f(\lambda_{k+1}) - f(\lambda_k) = -\ln \frac{Z(\lambda_{k+1})}{Z(\lambda_k)} \tag{2.118}$$

2.4.2.1 Practical Considerations

As the interactions of the ligand with its environment are switched from state A to B (on to off) we encounter a problem associated with the Lennard-Jones potential whereby at short distances, such as those encountered by mostly non-interacting molecules, the associated repulsive potential energy becomes infinitely high. This high energy is not very much like the final state of a non-interacting ligand and can severely limit the phase space overlap between neighbouring λ values. Instead, it is best practice to replace the LJ potential with a 'softcore' alternative to smooth the transition between λ values towards the non-interacting state. An example of a softcore potential is shown below.[191]

$$U\left(\vec{r}_{ij}, \lambda\right) = 4\epsilon_{ij}\lambda \left(\frac{1}{\left(\alpha(1-\lambda) + \left(r_{ij}/\sigma_{ij}\right)^6\right)^2} - \frac{1}{\alpha(1-\lambda) + \left(r_{ij}/\sigma_{ij}\right)^6} \right) \tag{2.119}$$

where r_{ij} is the distance between two particles i and j and α is a constant, with $\alpha = 0.5$ being the convention. This form of the Lennard-Jones equation recovers the exact LJ potential at $\lambda = 1$ (fully interacting) and is exactly zero when $\lambda = 0$ (fully non-interacting). Most importantly, as λ tends to zero, the $\alpha(1-\lambda)$ term lowers the high energies experienced at short r_{ij} distances. This potential is shown graphically in Figure 2.7.

Until now, there has been no mention of how the electrostatic interactions can be decoupled. While the softcore Lennard-Jones reduces the repulsive potentials at short r_{ij} distances, we can still get incredibly large electrostatic interactions which scale with decreasing atomic distance. While it is possible to apply a softcore approach to the Coloumb interactions, it is often easier to perform the transformations in sequence. This involves splitting the λ variable into electrostatic and van der Waals components,

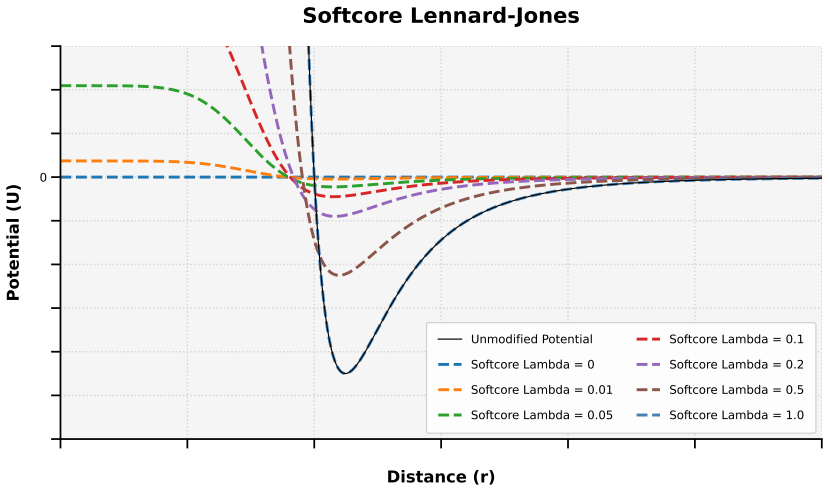


FIGURE 2.7: Softcore Lennard-Jones function (Eq. 2.119) at various values of λ. Crucially, the original potential is restored at $\lambda = 1$ and a flat potential at $\lambda = 0$.

λ_{ele} and λ_{vdW}. To avoid the high forces when decoupling a ligand from its environment, one would first turn off the electrostatics while the repulsive LJ potential maintains sensible distances and then follow up by turning off the softcore van der Waals.

A final consideration in absolute free energy calculations is the use of restraints. The complex leg of the double decoupling method results in a state where all the interactions of the ligand with the rest of the system have been turned off. When the ligand is mostly/entirely non-interacting with the environment it can easily dissolve from the binding site and begin to move around the simulation box by a random walk. This issue is magnified when simulating weak binders that may have a dissociation rate shorter than the length of the simulation. The random walk leads to a lack of phase space overlap between even the closest of lambda neighbours giving inaccurate results. Additionally, if the ligand samples regions of configurational space that are not relevant to the bound state making the final free energy estimate meaningless.

Preventing the ligand from sampling irrelevant regions of the configurational space can be done by trapping it into the relevant regions using restraints. The use of restraints then introduces a needed correction to the overall binding free energy:[191]

$$\Delta G^{\ominus}_{\text{restr}} = -k_B T \ln \left(c^{\ominus} V_L\right) - k_B T \ln \left(\frac{\xi_L}{8\pi^2}\right) \tag{2.120}$$

where V_L and ξ_L are the volume of the translational and rotational degrees of freedom of the non-interacting ligand in the system and are entirely restraint dependent. They can usually be obtained analytically or numerically by solving the relevant integrals. Note that the standard state correction ($c^{\ominus} V_L$), required for absolute free energy calculations, can also be incorporated into this correction.

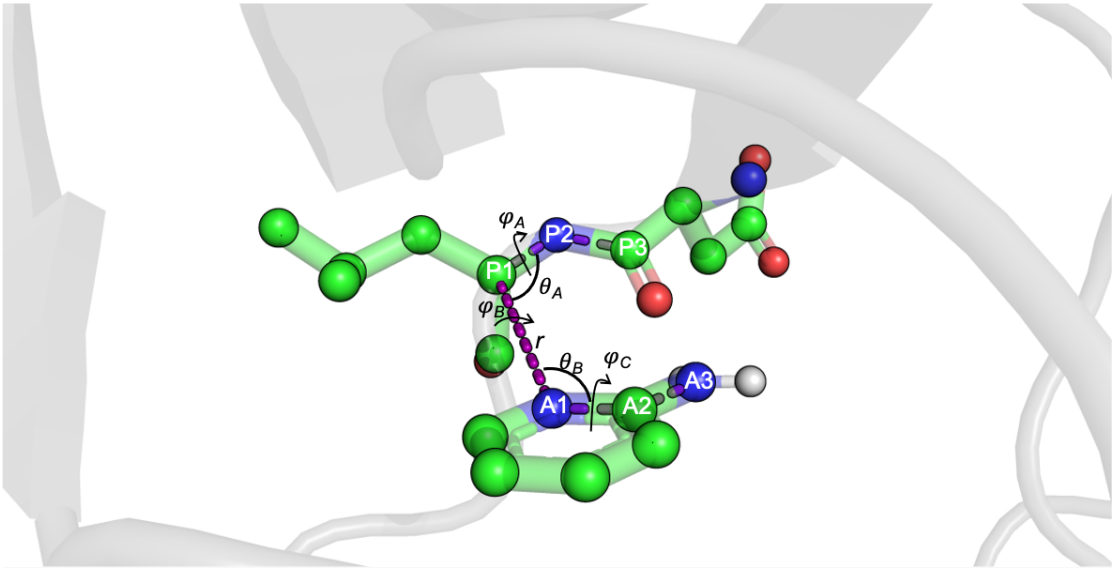


FIGURE 2.8: Schematic representation of the six degrees of freedom that define the orientation of the ligand relative to the protein. The distance r, the angles θ_A & θ_B and the dihedrals ϕ_A, ϕ_B & ϕ_C are all restrained using a separate harmonic potential for each.

Two simple restraints include a harmonic restraint and a flat bottom restraint, both of which do not limit the rotation of the ligand and therefore the rotational term in Equation 2.120 drops out. These distance restraints are useful for systems with multiple binding modes as the restraints do not restrict the sampling of these other conformations. However, for weaker binders, these restraints are often not strong enough.[127,128,191]

Boresch restraints[127] are another commonly used set of restraints which involve applying a harmonic potential to all six degrees of freedom that define the orientation of a ligand relative to its receptor. To define the restraint 3 protein atoms and 3 ligand atoms are selected to give one distance restraint, two angle restraints and three dihedral restraints. This is shown schematically in Figure 2.8.

The free energy change associated with imposing these restraints must be calculated and accounted for in the final estimate of the binding free energy. This can be done simply in simulation using a lambda scheme to gradually turn on the restraints measuring the potential energy which can then be estimated using a standard free energy estimator. The potential energy of the restraints can be written as:

$$U(\xi_i;\lambda) = \sum_{i=1}^{6} \frac{K_0\lambda}{2}(\xi_i - \xi_0)^2 \tag{2.121}$$

where ξ_i denotes the degree of freedom being restrained, ξ_0 is the reference value and K_0 is the harmonic force constant for that restraint.

Finally, the free energy of removing the restraints, which can also be thought of as the correction term mentioned above, can be calculated analytically using the expression derived by Boresch *et al.*:[127]

$$\Delta G^{\ominus}_{\text{restr, off}} = -k_B T \ln \left[\frac{8\pi^2 V_0}{r_0^2 \sin\theta_{A,0} \sin\theta_{B,0}} \frac{\left(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C}\right)^{1/2}}{(2\pi k_B T)^3} \right] \tag{2.122}$$

Bringing together the various concepts discussed, we can finally build the full thermodynamic cycle for the double decoupling method. The cycle indicates which simulations need to be performed and in which order the results must be used to calculate the absolute free energy of binding. Simulations are performed in solvent and in complex with the electrostatics being decoupled before the softcore Lennard-Jones interactions to avoid high forces.

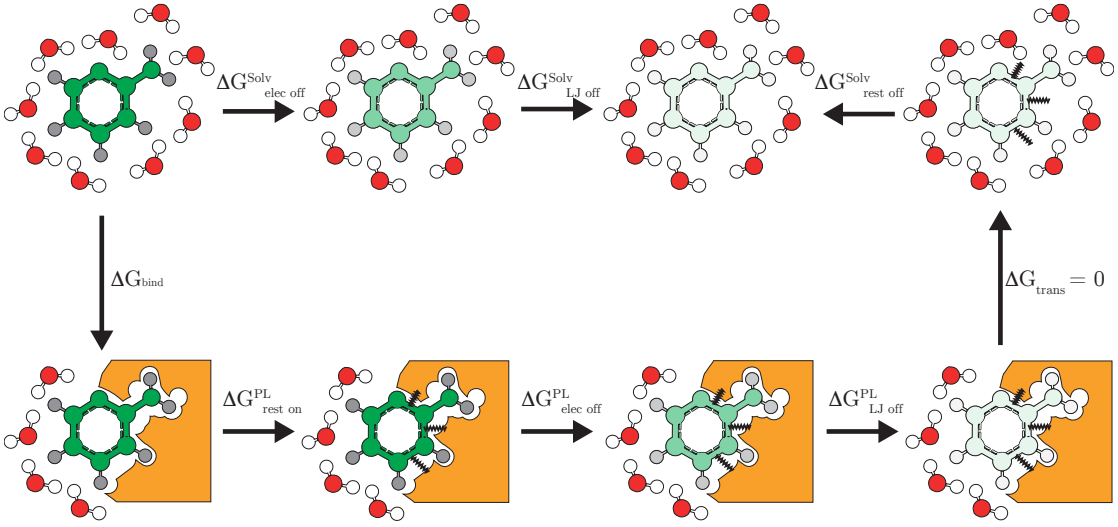


FIGURE 2.9: Thermodynamic cycle for a double decoupling absolute binding free energy calculation. $\Delta G^{Solv}_{elec,off}$, $\Delta G^{Solv}_{LJ,off}$, $\Delta G^{PL}_{elec,off}$, $\Delta G^{PL}_{LJ,off}$ are the free energies associated with turning of the electrostatics and Lennard-Jones in a solvent simulation and in complex respectively. $\Delta G^{PL}_{rest,on}$ is the free energy associated with imposing Boresch restraints on the system. Lastly, $\Delta G^{Solv}_{rest,off}$ is the free energy of removing the Boresch restraints and is calculated using Equation 2.122. The binding free energy can finally be calculated as: $\Delta G_{bind} = \Delta G^{Solv}_{elec,off} + \Delta G^{Solv}_{LJ,off} - \Delta G^{Solv}_{rest,off} - 0 - \Delta G^{PL}_{LJ,off} - \Delta G^{PL}_{elec,off} - \Delta G^{PL}_{rest,on}$

2.4.2.2 Sampling Schemes

The most common method of performing the free energy calculations described in the previous section is by simulating each lambda value in a separate simulation, or by moving through the lambda scheme in a single simulation and allowing the system to equilibrate at each change. This is known as equilibrium FEP and generally provides

good overlap between lambda states provided the spacing is appropriate. In equilibrium, FEP potential energy samples of each state are recorded and are used in the free energy estimators described below.

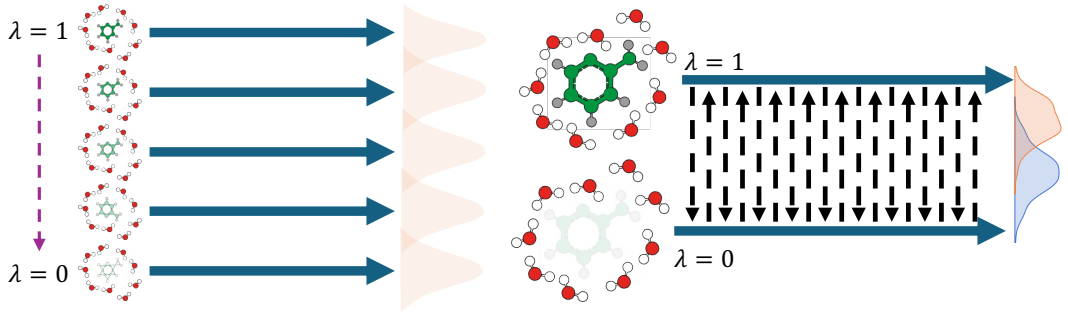


FIGURE 2.10: The two sampling schemes discussed. Left: Equilibrium FEP with multiple lambda states ran in parallel. Right: Nonequilibrium sampling scheme with rapid transitions between the two end states. Adapted from Mey *et al.*[191]

An alternative approach is the use of Nonequilibrium FEP (NEQ), where only the two physical end states need to be simulated at equilibrium with rapid nonequilibrium switches between the two.[191,195–197] In the case of ABFE calculations this involves rapidly switching the ligand from interacting to non-interacting and vice versa. The same thermodynamic cycle as in Figure 2.9 holds. Figure 2.10 graphically highlights the differences between the two sampling methods.

Practically, NEQ methods are conceptually very similar to EQ methods, however, rather than measuring the potential energy at equilibrium, the work done throughout the switch is measured. The free energy of the transition can be estimated using the Jarzynski estimator which can use either $A \rightarrow B$ or $B \rightarrow A$ works:[198]

$$\Delta F_{AB} = -k_B T \ln \langle e^{-\beta W_{AB}} \rangle_A \tag{2.123}$$

where W_{AB} is the work done in perturbing A to B. The value ΔF_{AB} is an average over all the works and will converge with more sampling. Since the Jarzynski equation only depends on the work done for a transition in one direction it is uni-directional, however, studies have shown that including work values in both directions can lead to a more accurate result. The Crooks theorem is a development to the Jarzynski equality that allows for bi-directional transitions that include both $A \rightarrow B$ and $B \rightarrow A$ works:[199]

$$\frac{P_F(W)}{P_R(-W)} = e^{\beta(W-\Delta F)} \tag{2.124}$$

where $P_F(W)$ and $P_R(-W)$ are the work distributions of the forward and reverse transitions. The free energy can then be estimated as the work value where both distributions overlap, $P_F(W) = P_R(-W)$.

A final analysis method is to use the Bennett Acceptance Ratio, described in Section 2.4.3.3, by simply replacing the potential energy term with work done for the forward and reverse transitions. It has been shown empirically that BAR outperforms the previous two estimation methods.[200,201] Further details on practical considerations and sampling schemes, particularly around the handling of multiple binding modes, are discussed in Chapter 5 Section 5.2.3.

2.4.3 Free Energy Estimators

As we know from Equation 2.114, the free energy difference between two states is related to the ratio of probabilities of those states. However, to estimate this difference we must convert the ratio of configurational partition functions into something that can be measured in simulation, such as the potential energy.

2.4.3.1 Zwanzig Relationship

Zwanzig, also known as one-sided exponential reweighting, is the most simple method for calculating the free energy between two states A and B can be calculated as follows:[202]

$$\Delta F_{AB} = -k_B T \ln \left\langle e^{-\beta(U_B - U_A)} \right\rangle \tag{2.125}$$

where U_A and U_B is the potential energy of states A and B. As mentioned before, the transition from A to B can be split into smaller steps and the total free energy can be calculated using Equation 2.117. Almost identical to the Jarzynski estimator, Zwanzig suffers from poor numerical convergence whereby if the standard deviation of the potential energy distribution is high, the result becomes biased and noisy as the tails of the distribution are exponentially weighted.[203,204]

2.4.3.2 Thermodynamic Integration (TI)

Thermodynamic integration[191,205] treats the partition function as a function of lambda where the derivative of the free energy ($F = -k_B T \ln Q$) with respect to λ can be shown as:

$$\begin{aligned}
\frac{dF}{d\lambda} &= -k_B T \frac{d}{d\lambda} \ln Q(\mathbf{q}, \lambda) \\
&= -k_B T \cdot Q(\mathbf{q}, \lambda)^{-1} \frac{d}{d\lambda} \int e^{-\beta U(\mathbf{q},\lambda)} d\mathbf{q} \\
&= -k_B T \cdot Q(\mathbf{q}, \lambda)^{-1} \cdot -\beta \int \frac{dU(\mathbf{q}, \lambda)}{d\lambda} e^{-\beta U(\mathbf{q},\lambda)} d\mathbf{q} \\
&= \left\langle \frac{dU(\mathbf{q}, \lambda)}{d\lambda} \right\rangle_\lambda
\end{aligned} \tag{2.126}$$

where $\mathbf{q}$ is the degrees of freedom in the coordinate space, Q the canonical ensemble (slight modifications are required for other ensembles) and $U(\mathbf{q}, \lambda)$ is the potential energy of the molecule with coordinates $\mathbf{q}$ at a set λ value. The final result shows that the derivative of the free energy is equal to the ensemble average of the derivative of the potential energy with respect to λ. Finally, the free energy difference between two states can then be written as an integral over the range of lambda values:

$$\Delta F_{AB} = \int_0^1 \left\langle \frac{dU(\mathbf{q}, \lambda)}{d\lambda} \right\rangle_\lambda d\lambda \tag{2.127}$$

Equation 2.127 is the final result of the Thermodynamic Integration derivation and can typically be evaluated using numerical methods that take the form:

$$\Delta F_{AB} \approx \sum_{k=1}^{K} w_k \left\langle \frac{dU(\mathbf{q}, \lambda)}{d\lambda} \right\rangle_k \tag{2.128}$$

where w_k are weights which depend on the numerical integration style of choice. For the most common, trapezium rule, even weights are selected such that $w_1 = w_k = \frac{1}{2(K-1)}$ and $w_{k\neq 1,K} = \frac{1}{K-1}$.

2.4.3.3 Bennett Acceptance Ratio

The Bennett Acceptance Ratio[206] (BAR) requires configuration information from two neighbouring states to calculate the free energy difference. It stands that the ratio of partition functions of states A and B can be written as a ratio of ensemble averages:

$$\frac{Q_A}{Q_B} = \frac{\left\langle \alpha(q) e^{-\beta U_A} \right\rangle_B}{\left\langle \alpha(q) e^{-\beta U_B} \right\rangle_A} \tag{2.129}$$

which holds for any value of $\alpha(q) > 0$. The free energy difference can then be written as:

$$\beta \Delta F_{AB} = \ln \left\langle \alpha(q) e^{-\beta U_A} \right\rangle_B - \ln \left\langle \alpha(q) e^{-\beta U_B} \right\rangle_A \tag{2.130}$$

Bennett finds that the free energy can be estimated by finding a value of $\alpha(q)$ that minimises the variance in ΔF_{AB}. The value of $\alpha(q)$ can either be guessed or found in a self-consistent manner by slowly iterating to find the minimum variance. Following the derivations by Bennett *et al.* we find that when self-consistency is attained, the free energy can be solved numerically using the following equation:

$$\sum_{A=1}^{n_A} \frac{1}{1+\exp\left(\ln\left(n_A/n_B\right)+\beta\Delta U_{AB}-\beta\Delta F_{AB}\right)} - \sum_{B=1}^{n_B} \frac{1}{1+\exp\left(\ln\left(n_B/n_A\right)-\beta\Delta U_{BA}+\beta\Delta F_{AB}\right)} = 0 \quad (2.131)$$

2.4.3.4 Multistate Bennett Acceptance Ratio (MBAR)

The Multistate Bennett Acceptance Ratio[207] is a direct extension to BAR and uses configuration data from all the lambda states to calculate the free energy of a single state. A full derivation of MBAR can be found in the works by Shirts *et al.*[207] with the final result showing that the free energy of a single state can be calculated as follows:

$$\hat{F}_A = -\beta^{-1} \ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{\exp\left[-\beta U_i\right]}{\sum_{k=1}^{K} N_k \exp\left[\beta \hat{F}_k - \beta U_k\right]} \quad (2.132)$$

where K is the number of individual states and N_K is the number of samples from each K state. Note that this result gives the free energy for a single state and therefore to calculate the free energy difference between two states you must use Equation 2.132 to calculate the free energy of both states and take the difference. MBAR has been shown to be significantly more accurate than other free energy estimators and is recommended for almost every use case.[203,207]

2.5 Monte Carlo

Monte Carlo (MC) sampling in the context of simulations is a technique used to generate samples of a system that satisfy the system's equilibrium probabilities. However, unlike molecular dynamics, MC does not generate a smooth trajectory nor any dynamic information about the system being studied. Instead, Monte Carlo generates a Markov chain of states by randomly changing the system and subjecting the "move" to an acceptance criteria that accounts for the equilibrium probability of the proposed state.[154,181] To rigorously sample the equilibrium distribution, we must impose the condition of microscopic reversibility via the detailed balance condition. This condition ensures that for any two microstates, the net flux of probability going between the two is zero, or in other words, the rate of transitions between any two states must be equal in both directions.

$$\pi(x)P(y|x)A(y|x) = \pi(y)P(x|y)A(x|y) \quad (2.133)$$

where $\pi(x)$ is the equilibrium probability of microstate x, $P(y|x)$ is the conditional probability of proposing a move to state y from x, and $A(y|x)$ is the conditional probability of accepting that move. A simple rearrangement of this equation gives a ratio of acceptance probabilities:

$$\frac{A(y|x)}{A(x|y)} = \frac{P(x|y)\pi(y)}{P(y|x)\pi(x)} \tag{2.134}$$

Unfortunately, the above relationship still does not give an explicit calculation for either of the acceptance probabilities. Instead, we can use the Metropolis-Hastings criteria to calculate the acceptance probability of a move $(y|x)$ without having to know the acceptance probability of the reverse move $(x|y)$:[208,209]

$$A(y|x) = \min\left[1, \frac{A(y|x)}{A(x|y)}\right] \tag{2.135}$$

The calculated acceptance ratio, dependent on the equilibrium probabilities of both states, is compared to a random number between 0 and 1. If the acceptance ratio is greater than the random number, the move is accepted and the new state is added to the Markov chain. If the move is rejected, a copy of the original state is re-added to the chain. In the above ratio, if the move from y to x is favourable then the ratio will always be greater than 1 and therefore be accepted.

Given that every sample in the chain is at equilibrium, the ensemble average of a property, A, is simply the mean of the property over all microstates in the chain.

$$\langle A \rangle \approx \frac{1}{M}\sum_{i=1}^{M} A_i \tag{2.136}$$

To understand how MC can be applied to a molecular system, let's consider a simple canonical (NVT) system of particles with positions $\mathbf{r}^N$ and apply a random translation ($\delta\mathbf{r}$) to one or more particles such that the system has a new set of coordinates, $\mathbf{r}^N_{new}$. In this case, the probability of proposing the reverse move is equal to that of the forward move and as such these terms cancel out in the acceptance ratio. Therefore, the acceptance ratio only depends on the equilibrium probability of the old and the proposed state. Equation 2.134 becomes:

$$\begin{aligned}\frac{A(\mathbf{r}^N_{new}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^N_{new})} &= \frac{\pi_{NVT}(\mathbf{r}^N_{new})}{\pi_{NVT}(\mathbf{r}^N)} \\ &= \frac{Q^{-1}_{NVT}\Lambda^{-3N}(N!)^{-1}e^{-\beta U(\mathbf{r}^N_{new})}}{Q^{-1}_{NVT}\Lambda^{-3N}(N!)^{-1}e^{-\beta U(\mathbf{r}^N)}} \\ &= e^{-\beta\Delta U}\end{aligned} \tag{2.137}$$

where $\pi_{NVT}(\mathbf{r}^N)$ is the probability density of a particular configuration, given by Equation 2.53. The acceptance ratio is now only dependent on the change in potential energy associated with the translation, ΔU, and the probability of accepting the move, $A(\mathbf{r}^N_{new}|\mathbf{r}^N)$, can then be calculated by subjecting the acceptance ratio to the Metropolis-Hastings test defined in Equation 2.135.

It should be noted that Monte Carlo sampling of this kind is very inefficient for large, condensed systems. This is because randomly translating a particle or particles almost certainly leads to a steric clash, resulting in a large change in potential energy and causing the move to be rejected. This can be circumvented by setting a maximum limit for how far a particle can be translated. This limit would, however, have to be very small, meaning that a very high number of moves would be needed in order to see any major change to the system. Many methods have been proposed to improve acceptance rates and make MC sampling more efficient; one such method is known as Nonequilibrium Candidate Monte Carlo[111] (Sec. 2.5.1).

2.5.1 Nonequilibrium Candidate Monte Carlo

Nonequilibrium candidate Monte Carlo (NCMC) is a method that breaks up large Monte Carlo moves into a series of smaller steps. These steps come in the form of sequential perturbations and relaxations, where the relaxation steps allow the system to adapt to a small perturbation, with the final product being the complete move. These series of small changes and relaxations greatly improve the likelihood of a move being accepted as the relaxation allows the system to adapt and prevent unfavourable steric clashes.[109–111,132,150,210–212]

An NCMC move separates a large move proposal (such as rotation of a dihedral) into a series of smaller perturbation steps connecting the two end states, denoted a_n, where work is done by making a small change to the system, and propagation steps, K_n, where the system releases heat as it relaxes to the perturbation. The "move protocol" refers to the order in which these steps are applied, e.g. $\Lambda_p = \{a_1, K_1, \cdots, a_T, K_T\}$. Applying this protocol to an initial state, x_0, generates a path/sequence of microstates denoted X, where $X = \{x_0, x_1, \cdots, x_T\}$. A forward move applying this protocol can be shown as:

$$x_0 \xrightarrow{a_1} x_1^* \xrightarrow{K_1} x_1 \rightarrow \cdots \rightarrow x_{T-1} \xrightarrow{a_T} x_T^* \xrightarrow{K_T} x_T. \tag{2.138}$$

To maintain a detailed balance, there must be a non-zero probability of selecting the reverse protocol such that when it is applied to state x_T the set of microstates, X, is reversed, thus returning the system to state x_0.

Just as in regular MC, an acceptance ratio can be derived. However, for NCMC, each step in the protocol needs to be taken into account leading to the following:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = \frac{P(\tilde{\Lambda}_p|\tilde{x}_T)}{P(\Lambda_p|x_0)} \frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)} \frac{\pi(\tilde{x}_T)}{\pi(x_0)} e^{-\Delta S(X|\Lambda_p)}, \tag{2.139}$$

where $P(\Lambda_p|x_0)$ is the probability of selecting protocol Λ_p and applying it to state x_0. $\alpha(X|\Lambda_p)$ is the cumulative probability of each perturbation step in the forward move and $\Delta S(X|\Lambda_p)$ is the conditional path action difference. Notations with the $\sim$ represent the reverse move. The latter terms depend on the individual steps within the NCMC move and can be further defined as:

$$\frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)} = \prod_{t=1}^{T} \frac{a_t(\tilde{x}_t^*, \tilde{x}_{t-1})}{a_t(x_{t-1}, x_t^*)} \tag{2.140}$$

$$e^{-\Delta S(X|\Lambda_p)} = \prod_{t=1}^{T} \frac{K_t(\tilde{x}_t, \tilde{x}_t^*)}{K_t(x_t^*, x_t)}, \tag{2.141}$$

where $a_t(x_{t-1}, x_t^*)$ is the probability of generating state x_t^* from state x_{t-1} by applying the protocol a_t. In a similar fashion, $K_t(x_t^*, x_t)$ is the probability of generating state x_t by applying the propagation K_t to the perturbed state x_t^*.

The acceptance ratio shown in Equation 2.139 is very general and seems complex, however, in most use cases it is typically much less so owing to various simplifications. For example, if the probability of selecting the forward and reverse moves is equal, then $P(\Lambda_p|x_0) = P(\tilde{\Lambda}_p|\tilde{x}_T)$ and cancels out. If the system is propagated using a deterministic approach such as molecular dynamics, then $K_t(x_t^*, x_t) = K_t(\tilde{x}_t, \tilde{x}_t^*)$ and as such, the conditional path action difference also drops out. Further simplifications can be made depending on its use case and will be discussed further.[111]

2.5.2 Grand Canonical Monte Carlo

2.5.2.1 Acceptance Criteria

Grand Canonical Monte Carlo (GCMC) is an MC method which can be utilised to sample the μVT ensemble discussed in Section 2.2.3.[135,140–143,145,147,149,151] GCMC provides a way of exchanging particles between the canonical system and the linked ideal gas reservoir in a theoretically rigorous manner that maintains the chemical potential in the system. A GCMC move can come in two forms, an insertion or a deletion, whereby a molecule is added or removed from the system. The acceptance criteria for these moves are derived below.[149,181]

For an insertion move, the number of particles in the system increases from N to $N+1$ by moving a particle from the ideal gas reservoir into the system. The equilibrium probability of a microstate with N particles in the system and $M-N$

particles in the ideal gas is given by:

$$\pi(\mathbf{r}^N, \mathbf{r}^{M-N}) = Q_{MVT}^{-1}\Lambda^{-3N}\Lambda^{-3(M-N)}e^{-\beta U(\mathbf{r}^N)}d\mathbf{r}^N d\mathbf{r}^{M-N}. \tag{2.142}$$

To move a particle from the ideal gas to the system we must select one of the $M-N$ particles in the reservoir and move it to a random point in the system. The probability of selecting this move is given by:

$$P(\mathbf{r}^{N+1}|\mathbf{r}^N) = \frac{1}{2}\frac{1}{M-N}\frac{d\mathbf{r}}{V_{sys}}, \tag{2.143}$$

where the factor of half is simply because there is a 50% chance of selecting an insertion rather than a deletion. The second term is the probability of selecting 1 particle at random from $M-N$ particles and lastly, the final term shows that the probability of picking a position in the system is inversely proportional to the volume of the system. The probability of selecting the same move in reverse is given by:

$$P(\mathbf{r}^N|\mathbf{r}^{N+1}) = \frac{1}{2}\frac{1}{N+1}\frac{d\mathbf{r}}{V_{gas}} \tag{2.144}$$

For a deletion move, the probability of selecting the forward and reverse moves is given by:

$$P(\mathbf{r}^{N-1}|\mathbf{r}^N) = \frac{1}{2}\frac{1}{N}\frac{d\mathbf{r}}{V_{gas}} \tag{2.145}$$

$$P(\mathbf{r}^N|\mathbf{r}^{N-1}) = \frac{1}{2}\frac{1}{M-N+1}\frac{d\mathbf{r}}{V_{sys}} \tag{2.146}$$

With these terms defined, the acceptance ratio for an insertion move can be derived as follows:

$$\begin{aligned}\frac{A(\mathbf{r}^{N+1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} &= \frac{P(\mathbf{r}^N|\mathbf{r}^{N+1})}{P(\mathbf{r}^{N+1}|\mathbf{r}^N)}\frac{\pi(\mathbf{r}^{N+1}, \mathbf{r}^{M-N-1})}{\pi(\mathbf{r}^N, \mathbf{r}^{M-N})} \\ &= \frac{d\mathbf{r}V_{gas}^{-1}(N+1)^{-1}}{d\mathbf{r}V_{sys}^{-1}(M-N)^{-1}}\frac{Q_{MVT}^{-1}\Lambda^{-3(N+1)}\Lambda^{-3(M-N-1)}e^{-\beta U(\mathbf{r}^{N+1})}d\mathbf{r}^{N+1}d\mathbf{r}^{M-N-1}}{Q_{MVT}^{-1}\Lambda^{-3N}\Lambda^{-3(M-N)}e^{-\beta U(\mathbf{r}^N)}d\mathbf{r}^N d\mathbf{r}^{M-N}} \\ &= \frac{M-N}{V_{gas}}\frac{V_{sys}}{N+1}e^{-\beta\Delta U}\end{aligned} \tag{2.147}$$

where ΔU is the change in potential energy of the system owing to the particle insertion. With similar substitutions, the acceptance ratio for a deletion move is given by:

$$\frac{A(\mathbf{r}^{N-1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N-1})} = \frac{N}{V_{sys}}\frac{V_{gas}}{M-N+1}e^{-\beta\Delta U} \tag{2.148}$$

These acceptance ratios can be further simplified, as the total particle number M is infinitely large, the ratios $\frac{M-N}{V_{gas}}$ and $\frac{M-N+1}{V_{gas}}$ can be reduced the number density which is related to the chemical potential of the ideal gas (Eq. 2.96). Making this substitution gives:

$$\lim_{M\to\infty} \frac{M-N}{V_{gas}} = \rho_{ideal}$$

$$\mu^{id} = -k_B T \ln(\rho_{ideal}\Lambda^3)$$

$$\frac{A(\mathbf{r}^{N+1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} = \frac{1}{N+1}\frac{V_{sys}}{\Lambda^3}e^{\beta\mu^{id}}e^{-\beta\Delta U} \tag{2.149}$$

$$\frac{A(\mathbf{r}^{N-1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N-1})} = \frac{N\Lambda^3}{V_{sys}}e^{-\beta\mu^{id}}e^{-\beta\Delta U} \tag{2.150}$$

Some of the terms above can be grouped into what is called the Adams parameter, denoted B, and as such the ratios can be simplified further:

$$B = \beta\mu^{id} + \ln\left(\frac{V_{sys}}{\Lambda^3}\right) \tag{2.151}$$

$$\frac{A(\mathbf{r}^{N+1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} = \frac{1}{N+1}e^{B}e^{-\beta\Delta U} \tag{2.152}$$

$$\frac{A(\mathbf{r}^{N-1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N-1})} = Ne^{-B}e^{-\beta\Delta U}. \tag{2.153}$$

It should be noted that, as with basic Monte Carlo, the insertion and deletion of full molecules is often very inefficient owing to the random placement of the molecule. It is unlikely that an insertion will be perfect and there is a strong chance of steric overlap leading to a high potential energy change for the move. Previous studies have seen acceptance rates as low as 0.03%.[147]

2.5.2.2 GCMC at Equilibrium with a Reference Solution

The theoretical definitions require that the system, or GCMC region, be in equilibrium with an ideal gas. It follows that this ideal gas can also be in equilibrium with an aqueous solution and by coupling these equilibria, it means we can effectively set the chemical potential of the ideal gas to that of some arbitrary reference solution, such as a bulk solvent or a mixture to observe a more physically meaningful equilibrium.[151] Figure 2.11 shows this graphically:

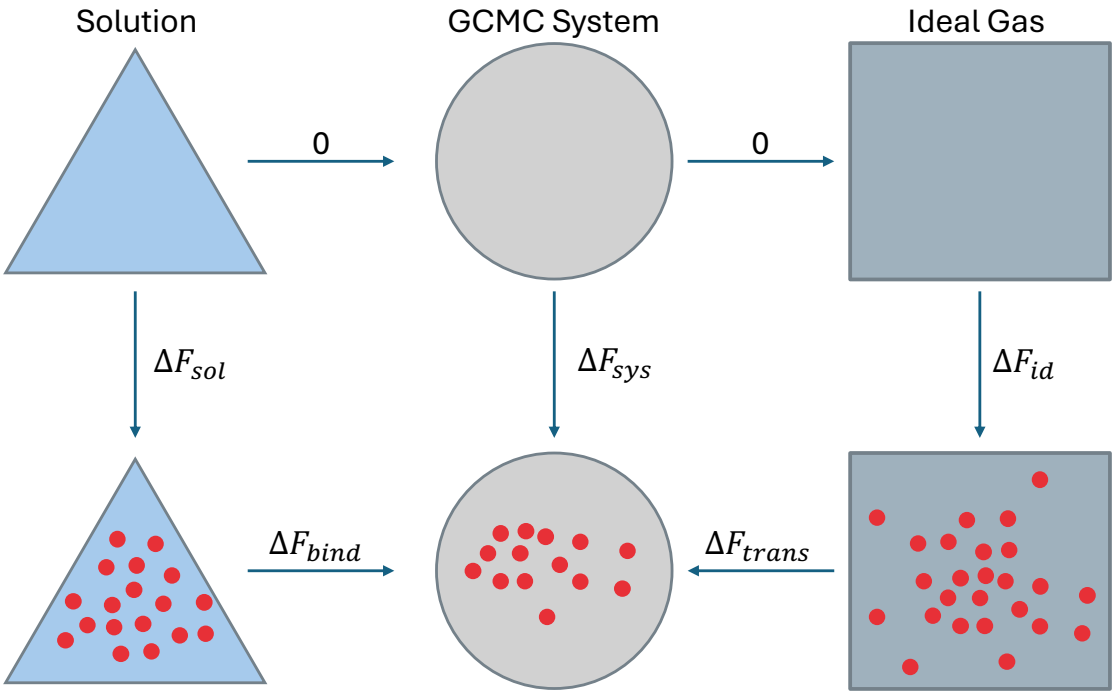


FIGURE 2.11: Thermodynamic cycle linking the binding of molecules from solution to the GCMC system with the binding of molecules from the ideal gas. The left triangles represent a solution phase, the circles represent the GCMC region/system, and the rectangles represent the ideal gas. The top row indicates systems without solute, while the bottom row contains solute particles indicated by the red dots.

This means that our GCMC system is now in equilibrium with a reference solution which is generally more useful when considering binding from solvent. It follows that the chemical potential of the reference solution has both an excess and an ideal component, noting that the ideal chemical potential for a molecule is identical in both the solution and gas phases. The total chemical potential of a molecule in a reference solution is given by:

$$\mu = \mu^{id} + \mu'_{sol}, \tag{2.154}$$

where μ'_{sol} is the excess chemical potential of a molecule in the reference solution. The ideal chemical potential is again defined as:

$$\mu^{id} = k_B T \ln \frac{N\Lambda^3}{V}, \tag{2.155}$$

where N is the number of particles in the reference, Λ is the thermodynamic wavelength, and V is the volume of the reference solution. Under standard state conditions, the number density, N/V is well defined as $1/V^{\ominus}$ giving:

$$\mu = k_B T \ln \frac{\Lambda^3}{V^{\ominus}} + \mu'_{sol} \tag{2.156}$$

Substituting this in the equation for the Adams value, B, gives the standard state Adams value in equilibrium with an arbitrary solution as:

$$\begin{aligned} B^{\ominus}_{eq} &= \beta\left(\mu'_{sol} + k_B T \ln\left(\frac{\Lambda^3}{V^{\ominus}}\right)\right) + \ln\left(\frac{V_{sys}}{\Lambda^3}\right) \\ &= \beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right), \end{aligned} \tag{2.157}$$

where the Adams value now depends on the excess chemical potential of the molecule of interest in a reference solution. Note that we have changed the notation of V_{sys} to V_{GCMC} as the GCMC may not be the whole system (e.g. a sphere).

2.5.2.3 Combining GCMC and NCMC

As alluded to previously, the acceptance rates in instantaneous MC and GCMC are vanishingly low and only get worse as the size of the molecule or system density increases.[147,149,151] In a previous subsection, NCMC was introduced as a method of improving the acceptance rate of large Monte Carlo moves. The same logic applies to GCMC moves. Combining the two means that throughout an insertion or deletion move, the molecule can slowly be switched on/off and the local environment can adapt over time to the move. In practice, these switches use the same λ protocols as described above with a softcore LJ in place. Together these moves are referred to as GCNCMC moves and the acceptance ratio is derived below.

Beginning from the general form of the acceptance ratio defined in Equation 2.139:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = \frac{P(\tilde{\Lambda}_p|\tilde{x}_T)}{P(\Lambda_p|x_0)}\frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)}\frac{\pi(\tilde{x}_T)}{\pi(x_0)}e^{-\Delta S(X|\Lambda_p)}, \tag{2.158}$$

where $P(\Lambda_p|x_0)$ is the probability of selecting protocol Λ_p and applying it to x_0. $\alpha(X|\Lambda_p)$ is the cumulative probability of each perturbation step in the forward move and $\Delta S(X|\Lambda_p)$ is the conditional path action difference. $\pi(x_0)$ is equilibrium probability of microstate x_0. Again, if we model the grand canonical ensemble as a large canonical system linked with an ideal gas reservoir, we can write the equilibrium probability of a microstate with N system particles and $M-N$ ideal gas particles as:

$$\pi_{MVT}(\mathbf{r}^N, \mathbf{r}^{M-N}, \mathbf{p}^M) = Q^{-1}_{MVT} h^{-3M} e^{-\beta E(\mathbf{r}^N, \mathbf{r}^{M-N}, \mathbf{p}^M)} d\mathbf{r}^M d\mathbf{p}^M, \tag{2.159}$$

where:

$$E(\mathbf{r}^N, \mathbf{r}^{M-N}, \mathbf{p}^M) = U(\mathbf{r}^N) + \sum_{i=1}^{M} \frac{|\mathbf{p}_i|^2}{2m}. \tag{2.160}$$

Note that the potential energy term has no dependence on the position of the ideal gas particles.

Similar to instantaneous GCMC, a particle is randomly selected from the ideal gas and moved to a random location in the system of interest, however, in the case of GCNCMC the particle's interactions are initially turned off to the environment, and are then slowly switched on via a lambda scheme between 0 to 1. Each lambda change is a perturbation kernel and is subjected to a relaxation (propagation) step. Since the lambda scheme is predetermined the perturbation kernels are therefore deterministic meaning the cumulative probability of each step in the forward move is equal to that in the reverse move such that $\alpha(X|\Lambda_p) = \alpha(\tilde{X}|\tilde{\Lambda}_p)$ cancelling out of the acceptance ratio. The probabilities of selecting the forward and reverse protocols are the same as regular GCMC, for an insertion move:

$$P(\Lambda_p|x_0) = \frac{1}{2}\frac{1}{M-N}\frac{d\mathbf{r}}{V_{sys}} \tag{2.161}$$

$$P(\tilde{\Lambda}_p|\tilde{x}_T) = \frac{1}{2}\frac{1}{N+1}\frac{d\mathbf{r}}{V_{gas}} \tag{2.162}$$

and for a deletion move:

$$P(\Lambda_p|x_0) = \frac{1}{2}\frac{1}{N}\frac{d\mathbf{r}}{V_{gas}} \tag{2.163}$$

$$P(\tilde{\Lambda}_p|\tilde{x}_T) = \frac{1}{2}\frac{1}{M-N+1}\frac{d\mathbf{r}}{V_{sys}}. \tag{2.164}$$

Substituting these relationships in Equation 2.158 and applying similar simplifications and rearrangements as in Section 2.5.2.1 we arrive at the following for an insertion move:

$$\begin{aligned}\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} &= \frac{P(\tilde{\Lambda}_p|\tilde{x}_T)}{P(\Lambda_p|x_0)}\frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)}\frac{\pi(\tilde{x}_T)}{\pi(x_0)}e^{-\Delta S(X|\Lambda_p)} \\ &= \frac{M-N}{V_{gas}}\frac{V_{sys}}{N+1}e^{-\Delta S(X|\Lambda_p)}e^{-\beta\Delta E(X|\Lambda_p)} \\ &= \frac{1}{N+1}\frac{V_{sys}}{\Lambda^3}e^{\beta\mu}e^{-\Delta S(X|\Lambda_p)}e^{-\beta\Delta E(X|\Lambda_p)} \\ &= \frac{1}{N+1}e^{B}e^{-\Delta S(X|\Lambda_p)}e^{-\beta\Delta E(X|\Lambda_p)}\end{aligned} \tag{2.165}$$

and for a deletion move:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = Ne^{-B}e^{-\Delta S(X|\Lambda_p)}e^{-\beta\Delta E(X|\Lambda_p)}. \tag{2.166}$$

Finally, to further simplify the acceptance ratios, we show that if the propagation kernel maintains the equilibrium distribution, then the conditional path action can be rewritten in terms of the heat change associated with the forward move:

$$\Delta S(X|\Lambda_p) = -\beta q(X|\Lambda_p) \tag{2.167}$$

Throughout this work, the Langevin integrator described in Section 2.1.3.2 is used and has been shown to sufficiently sample the equilibrium distribution.[178] Therefore, the above statement holds for the simulations performed in this work. It is also possible to decompose the energy change into heat released and the work done during the move:

$$\Delta E(X|\Lambda_p) = W_p(X|\Lambda_p) + q(X|\Lambda_p). \tag{2.168}$$

With these two substitutions, the acceptance ratio can be further simplified:

$$\begin{aligned} \frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} &= \frac{1}{N+1} e^{B} e^{-\Delta S(X|\Lambda_p)} e^{-\beta \Delta E(X|\Lambda_p)} \\ &= \frac{1}{N+1} e^{B} e^{\beta \Delta q(X|\Lambda_p)} e^{-\beta (W_p(X|\Lambda_p) + q(X|\Lambda_p))} \\ &= \frac{1}{N+1} e^{B} e^{-\beta W_p(X|\Lambda_p)}, \end{aligned} \tag{2.169}$$

where now the acceptance ratio only depends on the work done during the nonequilibrium switch which can be calculated as the sum of the work done on the system by each perturbation step:

$$W_p(X|\Lambda_p) = \sum_{t=1}^{T} [U(x_t^*) - U(x_{t-1})] \tag{2.170}$$

Following similar logic, the acceptance ratio for a GCNCMC deletion move is shown as:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = N e^{-B} e^{-\beta W_p(X|\Lambda_p)}. \tag{2.171}$$

Note that it turns out that the only difference in the GCNCMC acceptance ratio to instantaneous GCMC is simply the replacement of potential energy with the nonequilibrium work done over the protocol.

2.5.2.4 GCNCMC Implementation and Design Considerations

This project aims to build and develop the application of GCNCMC to small molecules. In previous work, GCNCMC has been shown to efficiently sample buried water molecules within protein systems. This was first introduced in a software package called ProtoMS, a pure Monte Carlo-based engine, and has since been redeveloped by Samways *et al.* into a Python module named *grand*. The *grand* module serves as an add-on to the OpenMM[152,161] molecular dynamics software and has many benefits over ProtoMS. Namely, the ability to combine GCMC insertions and deletions with conventional MD such that a canonical system can be simulated as usual with insertion and deletion moves interspersed. Further, the ability to perform MD is what facilitates the propagation steps in GCNCMC.

In *grandlig*, one can either insert/delete a molecule of interest from the whole system or the user can define a sphere to target the sampling, this sphere is known as the "GCMC region". While in theory any shape with a calculable volume can be used as the GCMC region, a sphere was chosen for its simplicity. This sphere can be 'attached' to the centre of a subset of user-defined protein atoms and will follow the protein around as it is simulated in the MD phase of the protocol. The use of a sphere means that it covers the targeted protein region independent of any protein rotations.

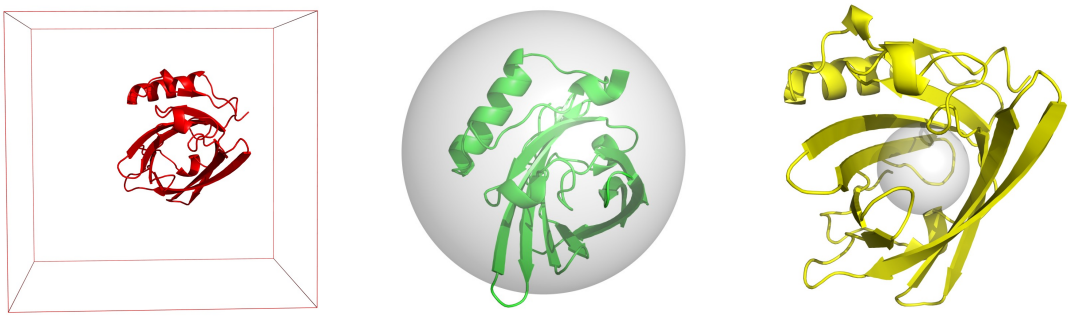


FIGURE 2.12: Examples of GCMC regions. Left: An entire simulation cell. Middle: A large sphere which encompasses a whole protein. Right: A smaller sphere targeted to a specific binding site.

The use of a sphere for non-equilibrium GCMC moves (GCNCMC) requires special care, in that other molecules can potentially diffuse in or out of the sphere throughout the switch, meaning that the acceptance ratios defined in Equations 2.169 and 2.171 must be adjusted slightly:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = \frac{1}{N_T} e^{B} e^{-\beta W_p(X|\Lambda_p)} \tag{2.172}$$

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = N_0 e^{-B} e^{-\beta W_p(X|\Lambda_p)}, \tag{2.173}$$

where N_0 is the number of GCMC molecules in the sphere at the start of a move and N_T is the number of molecules at the end. Additionally, it should be noted that if a molecule that is switched lies outside of the sphere by the end of the move, it should be automatically rejected since the reverse protocol cannot be proposed, breaking the condition for detailed balance.

A final consideration is how long the GCNCMC move should take which is referred to as the switching time, τ, and depends on the number of perturbations and the length of the propagation steps:

$$\tau = (n_{pert} + 1) n_{prop} \delta t, \tag{2.174}$$

where δt is the MD time step employed during the relaxation process. In theory, the switching time should not impact accuracy but can help to improve convergence by enhancing acceptance rates with longer switching times. However, there is a trade-off

between the length of the move and how many moves can get done in a certain amount of computing time.

2.5.2.5 Summary

In summary, the entire GCNCMC/MD protocol involves running regular MD to propagate the system with GCNCMC moves interspersed. An insertion or a deletion move is selected with equal probability. For an insertion move, a "ghost" molecule is randomly placed into the GCMC region, while for a deletion move, a fully interacting molecule within the GCMC region is randomly selected. The nonbonded interactions of the selected molecule are then scaled appropriately throughout the switch. To avoid numerical instabilities as a result of the nonphysical states sampled during a move, a soft-core Lennard-Jones potential is used as described previously.[150] For an insertion move, the Lennard Jones interactions are fully switched on before the electrostatics and vice versa for a deletion move to avoid any naked charges. At the end of the NCMC move, the acceptance test is performed according to Equations 2.169, 2.171, 2.172 & 2.173 and, if the move is accepted, the new state is added to the Markov chain. If the move is rejected, a copy of the previous state is added to the chain and the simulation continues. Choices for the GCMC region, switching time, and number of cycles can be adjusted according to the situation. A graphical summary of the method is shown in Figure 2.13. In Chapters 4 and 5 this GCNCMC method is further developed and validated to sample the binding of small molecules. In the further chapters the method is applied in an SBDD setting to demonstrate its ability to robustly identify ligand binding sites, sample multiple binding modes and calculate binding affinities.

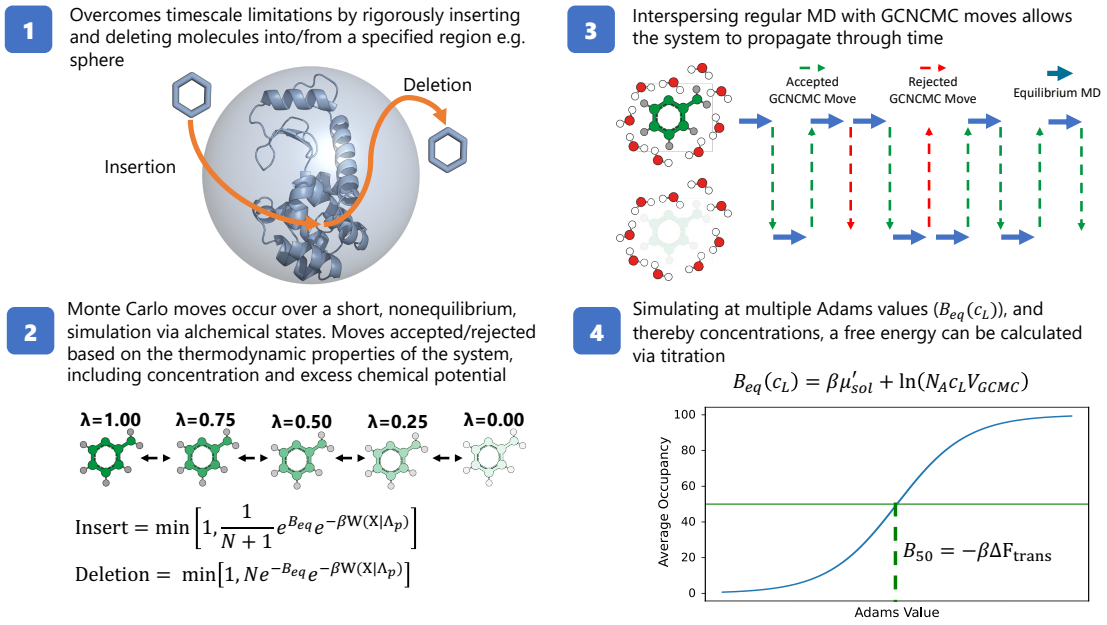


FIGURE 2.13: High level overview of the GCNCMC protocol. 1) Insertion and deletion moves occur within a user-defined region (grey sphere). 2) Moves are performed using a nonequilibrium switch occurring over a short time scale and are accepted or rejected according to the work done on the system over the move, $W(X|\Lambda_p)$, the excess chemical potential of the molecule, μ'_{sol}, the concentration of the molecule, c_L, the number of molecules already in the region, N, and the volume of the defined GCMC region, V_{GCMC}. 3) The resulting simulation is a regular MD simulation with GCNCMC moves interspersed. If a move is rejected (red dashed lines) the simulation restarts from the state before the move. 4) Binding affinities may be calculated by titrating the Adams value, B_{eq}, and thereby concentration. More details can be found in the "Methods" section.

Chapter 3

Preliminary Studies of MiniFrag Binding

3.1 Introduction

The MiniFrag[11] regime takes FBDD to the extreme by using even smaller fragments compared to traditional fragment libraries, however, their characteristically weak binding can limit detection in functional and biophysical assays. In this chapter, we perform some basic preliminary studies to understand how these weakly binding molecules behave *in silico* and how simulations can be used to enhance this drug discovery regime. In particular, we are looking to find methods of reproducing experimental binding poses, such that simulations could be used in cases where experiments are not feasible, as well as strategies that will complement the experimental data by calculating binding affinities of poses. Such tools could be used in industry to save time, effort and cost by guiding drug discovery programmes and avoiding the need to perform unnecessary and sometimes expensive experiments.

Consistent with the philosophy of FBDD, MiniFrag screens have shown improved hit rates compared to a traditional fragment library, though this may also be partly due to the higher concentrations used (1 M) compared to more traditional screens (100 mM).[11] That said, the improved coverage of chemical space and lower molecular complexity are certainly contributing factors by the same reasoning as fragments versus lead-like screening.[11] Interestingly, the primary goal of the MiniFrags was not simply to improve hit rates, but rather to identify novel, weaker points of interaction that cannot be found using larger molecules. These new binding regions could be linked together in the optimisation process to increase affinity by exploiting these previously undiscovered interactions. This 'fragment-linking' approach is common in structure and fragment-based drug design.[35,36] In all the systems tested, the MiniFrag screen identified pockets not found in the more traditional screen. This is exemplified

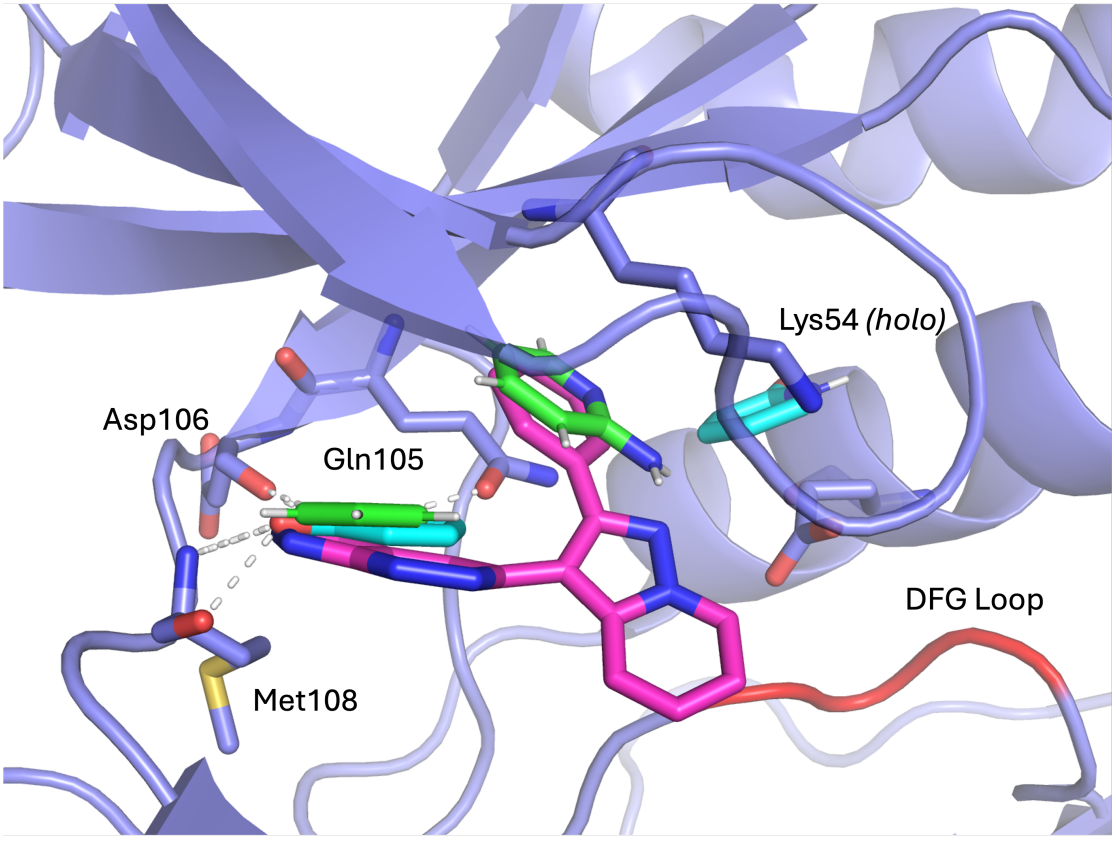


FIGURE 3.1: ERK2 active site (purple, PDB: 6qa1). Pyridin-2-amine (green) is shown bound in sites 1a and 1b, and Pyridin-2-one (cyan) in sites 1a and 1c. Overlaid in magenta is a known ERK2 inhibitor (PDB: 1tvo). Key residues and the DFG loop are labelled.

by the ERK2 kinase system where MiniFrags were found to bind in four sub-sites within the kinase active site (1a-d). In contrast, the more traditional fragment screen only identified sites 1a and 1d. Here, we focus only on the three sub-sites 1a-c as in Figure 3.1. In the case of 1c, no drug molecules are known to bind in this site, which may indicate some unrealised potency for new lead molecules, further highlighting the potential of MiniFrags. Site 1b is a known cryptic pocket and binding to this site requires side chain movements of the Lys54 and Gln105 residues. Cryptic pockets are an ongoing challenge in drug design as they are generally hard to detect both experimentally and computationally, so it is remarkable that molecules of this size can induce such large protein movements.[97,213]

3.1.1 FTMap and Related Methods

FTMap[88] is a static structure computational tool that uses a fast Fourier transform (FFT) algorithm to identify binding hotspots in biological systems. The algorithm distributes 16 different organic molecules over a protein surface and uses an FFT correlation approach to sample probe positions on both a translational and rotational

grid. Using this FFT-based approach significantly decreases the calculation time when compared to the classical docking of probes. The algorithm samples a large number of conformations for each probe, and then energy minimizes the 2000 lowest energy poses. The minimized probe conformations are clustered based on their Boltzmann averaged energies with the lowest average energy clusters retained up to a maximum of six clusters. The centre of these probe clusters are further clustered together with the other probes to seek a consensus on the most favourable binding regions.

FTSite[214] is a second tool that is heavily based on the FTMap algorithm. However, despite being similar, FTSite is designed to identify actual ligand-protein binding sites, rather than hotspots, by ranking clusters based on the number of interactions made with the protein rather than the total number of probes in the cluster.

Here, we apply FTMap and FTSite to the ERK2 system in an attempt to map the same binding pockets as the MiniFrags. Although FTMap has published success, it is a static structure method and may fall short on many problems simply due to the lack of protein dynamics and flexibility.[215] It would be expected that both FTMap and FTSite could predict the sites in the *holo* structures with some degree of accuracy owing to the fact the binding sites are already preformed for the algorithm. However in the case of the *apo* structure, particularly for sites 1b and 1c, the pockets are shut and occluded requiring protein movements to open, and are therefore likely to go undetected.

It is worth noting that there are many web servers in the FTMap family, one of which, FTFlex,[216] does incorporate protein flexibility but requires the user to select hotspots based on an initial FTMap run. Additionally, at the time of writing the authors of FTMap have released two new methods, E-FTMap[217] and FTMove,[218] where the former has extended the probe set from 16 to 119 molecules and the latter scrapes the PDB for similar proteins to the one submitted and performs FTMap analysis on all of them. In the case of FTMove, improved results are only expected if there are *holo* structures, or structures with pre-formed pockets, available in the PDB.

3.1.2 Mixed-solvent Simulations

Mixed-solvent MD (MSMD) methods such as MDMix,[219] SILCS,[91] and MixMD[89,92] were described in depth in the Introduction (Sec. 1.5.2) and have proven to be effective at mapping protein surfaces for interaction hotspots, analogous to the MSCS method.[47] MiniFrag screens follow a similar protocol to MSCS and as such MSMD simulations may be a useful tool for mapping MiniFrag interactions. Typically the probes used in MSMD are carefully selected to represent a wide variety of interactions, for example, methanol and isopropanol are used by SILCS and MixMD as hydrogen bonding probes. Generally, the largest probes used by MSMD algorithms do not exceed simple 6-membered rings making the MiniFrags complementary to this

approach. Here we aim to develop a basic MSMD protocol and apply it to the MiniFrags-ERK2 system to attempt to reproduce the correct binding poses seen in the crystal structures.

3.1.3 Absolute Binding Free Energy Calculations

As discussed in the Introduction (Sec. 1.5.4), MiniFrags are screened using X-ray crystallography which, while being useful at giving a full picture of the binding, does not provide any thermodynamic data such as the binding affinity. To this end, it is beneficial to develop a protocol that can computationally calculate MiniFrag affinity. These calculations can also be used to guide the simulations; for example, weaker binders may require more elaborate techniques to reproduce binding in simulation.

Here, we apply basic absolute binding free energy (ABFE) calculations, using the double decoupling method, to pyridin-2-amine bound to ERK2. Pyridin-2-amine binds to ERK2 in multiple places, two of which sit within the kinase active site (1a and 1b, Fig. 3.1).

3.2 Simulation Details

3.2.1 Static Structure Analysis

To test both FTMap's and FTSite's robustness against ERK2, three structures were submitted to the publicly available web servers - an *apo* structure (3o71), a structure with pyridin-2-amine bound in sites 1a and 1b (6qa1) and a structure with pyridin-2-one bound in sites 1a and 1c (6qa4).

3.2.2 Mixed-solvent Simulations

Mixed solvent MD simulations were performed using the published MiniFrags known to bind to ERK2 (Figure 3.2) as well as four probe molecules: acetonitrile, isopropanol, n-methylacetamide, and pyrimidine. These probe molecules are discussed further in Chapter 7. Simulations were performed using an *apo* structure (PDB: 4gsb) with all crystal waters and small molecules removed. To equilibrate the protein structure, the protein was first solvated in just TIP3P water and 0.15M of sodium chloride ions. Further ions were added to neutralise the system (2 x Na). This protein-water system was heated to 298 K over 0.5 ns of MD with a further 1.5 ns in the NVT ensemble at 298 K. The volume was then equilibrated over 8 ns of MD in the NPT ensemble at 1 bar using a Monte Carlo barostat. This equilibrated protein structure was then

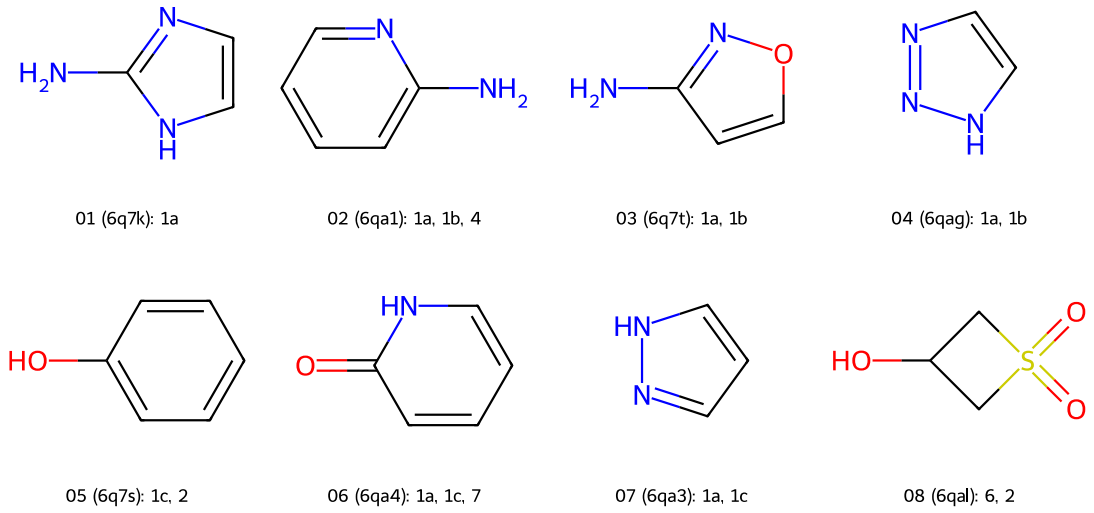


FIGURE 3.2: MiniFrags which bind to ERK2 and their PDB accession codes. The sites in which each MiniFrag binds are also indicated.

removed from the solvent and re-solvated five times in water and 0.5 M of probes; this was to ensure each replicate had an independent solvent configuration. Ions were added to neutralise the system and up to a concentration of 0.15 M.

These protein-cosolvent systems were then equilibrated for 1 ns in NVT and 3 ns in NPT. Production simulations were then run for 50 ns in the NPT ensemble giving a total simulation time per probe of 250 ns. The final 10 ns of each repeat are combined and used for analysis described below.

The AMBER ff14SB[220] and TIP3P[221] forcefields were used to model the protein and water respectively. The general AMBER forcefield (GAFF)[222] with AM1-BCC[223] charges were used to model the ligands. Long-range electrostatic interactions were calculated using PME[183] with a 12 Å cut-off. A switching function was applied to the Lennard-Jones potential between 10 and 12 Å. Simulations were carried out at 298 K using the BOAOB Langevin integrator[178] with a 4 fs timestep using hydrogen mass repartitioning[155] (H mass=4 amu) and a friction coefficient of 1 ps^{-1}. A Monte Carlo barostat was used to keep the pressure at 1 bar with volume changes attempted every 25 timesteps.

To analyse MSMD simulations, we perform a very basic grid analysis similar to previously reported MSMD studies.[89,94] After aligning the trajectory to a single state, a fictitious grid is built in the system with grid voxels spaced 0.5 Å apart. Then, for each frame of the trajectory we loop over all the heavy atoms of all the probes in the system. If a probe atom overlaps with any voxel, the voxel occupancy for that frame is assigned the number 1. The total occupancy of each voxel is then calculated by summing the amount of frames a probe was present. The final summation is then divided by the total number of frames to provide an average occupancy for each voxel:

$$< O >_{x,y,z} = \frac{\sum_i^{N_{frames}} O^i_{x,y,z}}{N_{frames}} \tag{3.1}$$

where $< O >_{x,y,z}$ is the average occupancy of a voxel at positions x, y and z. $O^i_{x,y,z}$ is the occupancy of a voxel at frame i and N_{frames} is the total number of frames in the simulation. A deeper discussion on these grid analyses can be found in Chapter 7.

3.2.3 Pocket Exposure Simulations

Pocket exposure calculations were performed using fpocket[224] with the following equation:

$$Pocket\ Exposure\ (\%) = \frac{PLA_{ligand} \cup PLA_{fpocket}}{PLA_{ligand}} \times 100 \tag{3.2}$$

where PLA stands for pocket lining atoms. PLA_{ligand} are all the heavy atoms within 4.5 Å of the bound ligand in the *holo* structure, and $PLA_{fpocket}$ are the heavy atoms defined by fpocket that make up the lining of the pocket at a given snapshot.

3.2.4 Absolute Binding Free Energy Calculations

The structure of pyridin-2-amine bound to ERK2 in sites 1a and 1b (PDB: 6qa1) was used as a starting point for these simulations. Crystal waters and bound ligands, except for the ligands of interest, were removed. The AMBER ff14SB[220] and TIP3P[221] forcefields were used to model the protein and water respectively. The general AMBER forcefield (GAFF)[222] with AM1-BCC[223] charges was used to model the ligands. Long-range electrostatic interactions were calculated using PME[183] with a 12 Å cut-off. A switching function was applied to the LJ interactions between 10 and 12 Å.

Simulations were carried out at 298 K using the BOAOB Langevin integrator[168] with a 2 fs timestep and a friction coefficient of 1 ps^{-1}. A Monte Carlo barostat was used to keep the pressure at 1 bar with volume changes attempted every 25 timesteps.

Boresch restraints[127] between three ligand, and three protein, atoms were added using a custom force of the form shown in Equation 2.121. The restraints were imposed using a non-linear lambda scheme
($\lambda = 0.00, 0.01, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.85, 1.0$) collecting 1000 samples per lambda window with each sample accounting for 3 ps of simulation time (total 3 ns per lambda).

Decoupling simulations were performed over 32 lambda windows with the first 10 windows turning off the electrostatics in a linear fashion and the final 22 windows

turning off the Lennard-Jones interactions using a non-linear scheme weighted towards the 'off' state. Samples were collected every 3 ps to give 1000 samples per lambda window. The simulations in the solvent system (water + ligand) and the complex system were performed identically. The potential energy samples were analysed using MBAR.

The thermodynamic cycle described in Figure 2.9 is used to calculate the absolute binding free energy of pyridin-2-amine in sites 1a and 1b. The free energies are calculated for both sites with just one ligand present and also while the second ligand is present. This should highlight any cooperative binding that may be in effect. It is also possible to combine all of these results into a thermodynamic cycle which should equal zero upon closure (Figure 3.3).

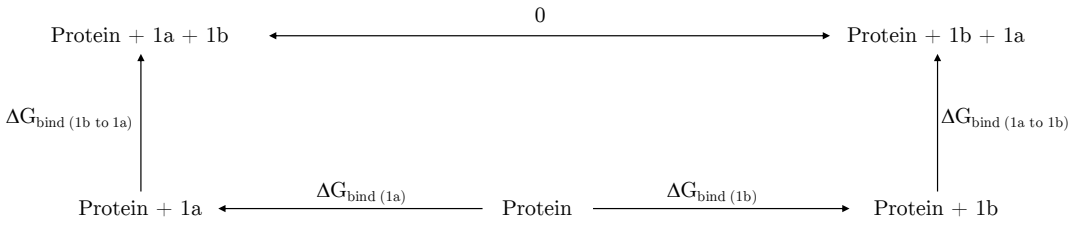


FIGURE 3.3: Thermodynamic cycle showing the different routes taken to reach a final state of "Protein + 1a + 1b". Following the ΔG terms in the cycle clockwise from top right to top left should equal 0 (Equation 3.3).

3.3 Results and Discussion

3.3.1 Static Structure Analysis

As expected, FTMap using the *holo* structures 6qa1 and 6qa4 mostly highlights the correct binding regions with 1a being highlighted as the top ranked cluster in both structures with the expected interactions with the protein reproduced well. Site 1b using 6qa1 is somewhat well reproduced by the second ranked cluster however the overlap of the cluster and the bound ligand is not perfect. Interestingly, site 1c using 6qa4 was identified in the 12th ranked cluster which may suggest this site is weakly binding (*holo* results are not shown).

Using FTMap with an *apo* structure is the bigger test of the algorithm as the sites are not pre-formed. Figure 3.4 shows the results from FTMap using an *apo* structure overlaid with bound pyridin-2-amine and pyridin-2-one molecules in sites 1a and 1c. The results show that FTMap only manages to correctly identify site 1a as the second highest ranked cluster with another cluster close by (ranked 7th). Sites 1b and 1c were not found at all.

The results from FTSite are similar to FTMap with some slight differences. FTSite using the *holo* pyridin-2-amine structure (6qa1) shows the top ranked binding site as a large region which covers both 1a and 1b. Using the pyridin-2-one *holo* structure, FTSite only identifies the 1a site. Likewise, FTSite using the *apo* structure only manages to find site 1a and ranks it the lowest of the three predicted sites (Figure 3.4. The two sites ranked above may be false positives as they do not correspond to any published MiniFrag structure.

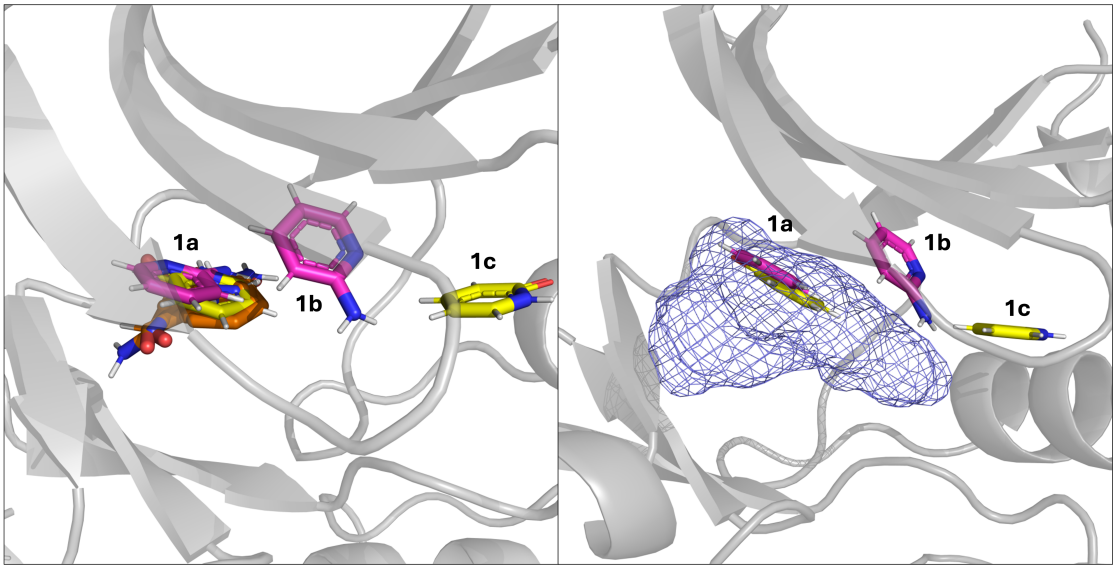


FIGURE 3.4: FTMap (left) and FTSite (right) outputs from using the *apo* ERK2 structure 3o71 (gray). Overlaid are the bound pyridin-2-amines (purple) and pyridin-2-ones (yellow) from their respective structures (6qa1 and 6qa4). The second ranked FTMap cluster is shown in orange and the lowest ranked FTSite is shown in blue mesh.

3.3.2 Preliminary MD Simulations

3.3.2.1 Mixed-Solvent Simulations

MSMD simulations with 0.5 M of each MiniFrag were performed starting from a true *apo* (4gsb) structure. Selected occupancy grid results are shown in Figure 3.5. These simulations were performed to assess the performance of MSMD simulations to sample the kinase substrate binding site using known MiniFrag binders as well as smaller, cosolvent probes. A selection of key results are shown in Figure 3.5 and discussed further below.

What is immediately clear from trajectory visualisation and occupancy analysis is that sites 1a and 1c are sampled by some, but not all, MiniFrags. Site 1a is the most solvent accessible site and experimentally binds six out of the eight MiniFrags as indicated in Figure 3.2. MiniFrags 01 to 07 all sampled the site for at least 10% of frames with only MiniFrags 03 to 06 exhibiting a percentage greater than 30%. The MiniFrag with the

TABLE 3.1: Final maximal occupancies from MSMD simulations of ERK2. Results are judged by visual inspection of the grid analysis.

	01	02	03	04	05	06	07	08	pyr	iso	nme	acn
1a	10	20	50	35	30	30	20	00	40	30	00	20
1b	00	10	00	00	00	00	00	00	00	00	00	00
1c	00	00	00	20	00	00	50	00	00	10	00	20

highest occupancy in site 1a was 03 with an occupancy of 50%. In general, these results are somewhat disappointing as an occupancy of 10% makes it impossible to distinguish the site from the background noise as indicated in Figure 3.5 (left) where it is difficult to pick out the binding regions even at 40%. In the MixMD[225] approach, grids are normalised such that the contour levels represent the number of standard deviations between the raw grid point occupancy and the mean occupancy giving a somewhat better filtering of the noise. This may be a useful avenue to explore in the future.

The cryptic site, 1b, was almost exclusively not sampled by any of the MiniFrags. This result is not unexpected as moving the occluding lysine is likely to require longer timescales than were simulated in this study. That said, however, it was hoped that a MiniFrag could induce this movement. MiniFrag 02 appears to bind in site 1b for 10% of the analysis frames, however, upon closer inspection the binding pose is erroneous and the lysine conformation is still akin to the apo structure.

The third sub-site, 1c, was again poorly sampled. This pocket is positioned towards the rear of the active site and is less exposed than site 1a. Only MiniFrags 04 and 07 are bound to this site for 20% and 50% of frames respectively. Interestingly, both 04 and 07 are the only two 5-membered MiniFrags with no substituents making them the smallest by volume in the set, indicating a possible size dependence for binding in simulation. Experimentally only MiniFrags 05, 06 and 07 bind in pocket 1c, validating the simulation results for 07.

Finally, simulations using more traditional cosolvent probes, acetonitrile (acn), isopropanol (iso), n-methylacetamide (nme), and pyrimidine (pyr) were performed to assess the performance of these slightly smaller molecules at mapping the active site. In all cases, these simulations were relatively disappointing with site 1a being mapped up to 40% by pyr and site 1c 20% by acn. Site 1b was not identified at all by the probe molecules.

3.3.2.2 Pocket Exposure Studies

To investigate the pocket dynamics of the active site, simulations of ERK2 starting from a bound pyridin-2-amine structure with the ligands removed (PDB: 6qa1) were

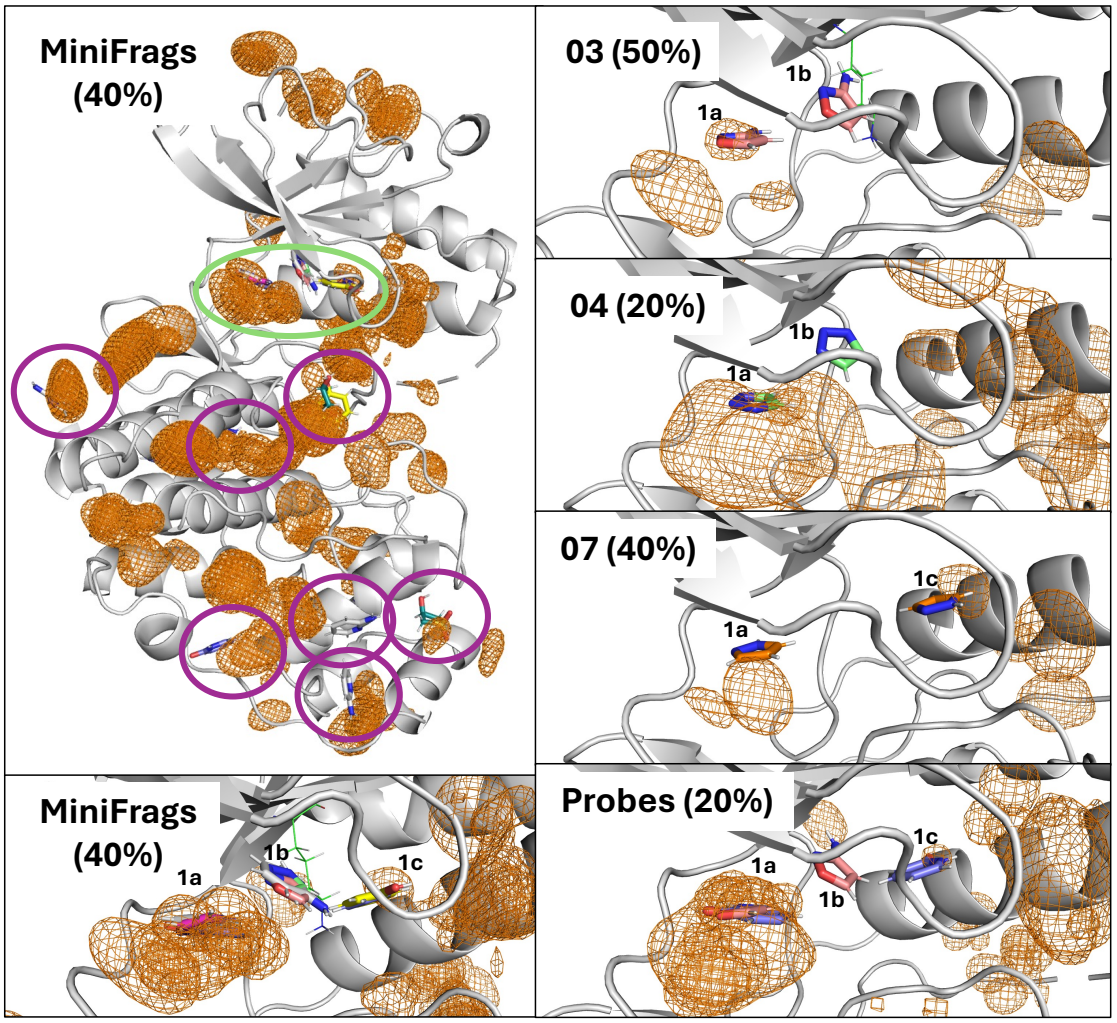


FIGURE 3.5: Results from mixed solvent MD simulations of ERK2 with different MiniFrags. Upper left shows overlaid occupancy maps, for all eight MiniFrags, contoured at 40%, detailing that a MiniFrag was present at a particular grid point for 40% of the analysis frames. The crystal poses of all MiniFrags are shown as sticks and the different binding regions are circled. Lower left is a zoomed view of the traditional kinase active site showing that site 1b was not explored by any MiniFrag, likely owing to the occluding lysine residue shown in green. Figures on the right show individual MiniFrags at differing percentages. Probes refer to simulations of the more traditional cosolvent probes acn, iso, nme, pyr. The protein, shown in grey, is from PDB 6qa1.

performed. Simulations were performed for ∼70 ns and the pocket exposure was measured every 20 ps using fpocket and reported using Equation 3.2. Figure 3.6 shows that the pocket exposure of the 1a pocket remains consistent throughout with the *holo*, 6qa1, pocket exposure akin to that of the *apo* structure (4gsb). The 1b pocket sees a prompt closure during the equilibrium phase to resemble an exposure akin to the *apo* structure (Figure 3.7).

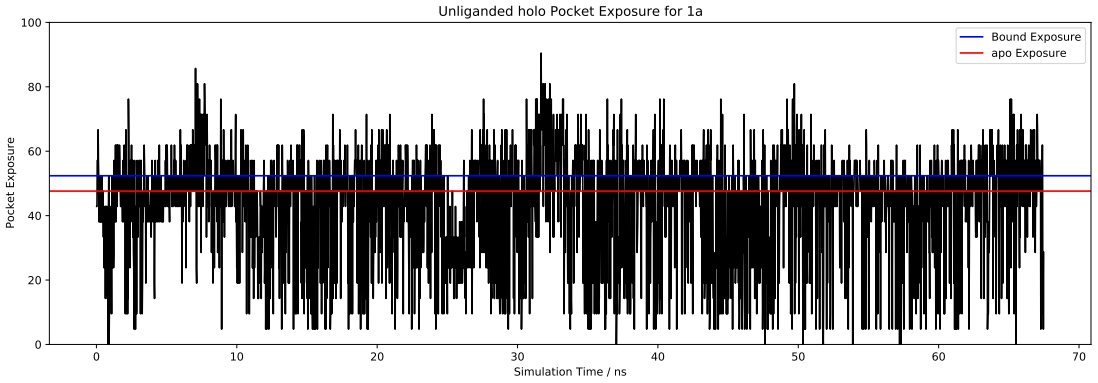


FIGURE 3.6: Pocket exposure (Eq. 3.2) of the **1a** pocket over time for a simulation starting from a bound *holo* structure with the ligand removed (black time series). The 1a pocket exposure measured for the static *apo* (3o71) and *holo* (6qa1) structures are shown in red and blue respectively.

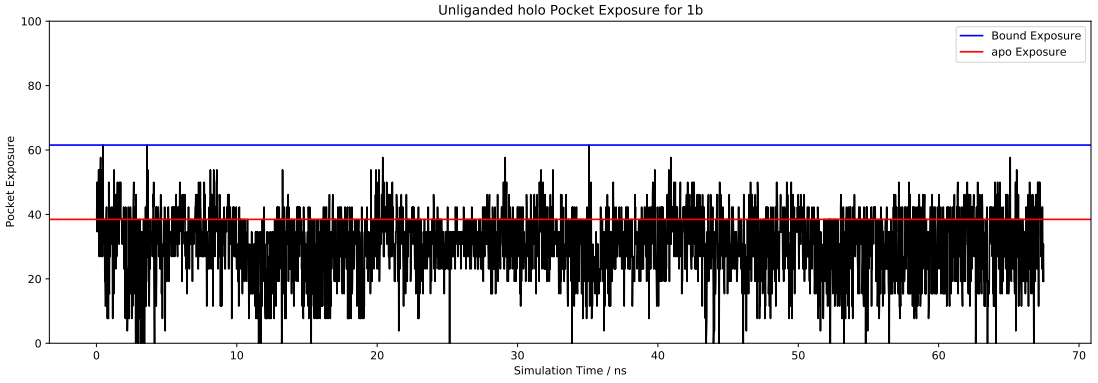


FIGURE 3.7: Pocket exposure (Eq. 3.2) of the **1b** pocket over time for a simulation starting from an unliganded *holo* (6qa1) structure. The 1b pocket exposure measured for the static *apo* (3o71) and *holo* (6qa1) structures are shown in red and blue respectively. This plot represents a prompt closure of the 1b pocket without the presence of a ligand.

3.3.3 Absolute Binding Free Energy Calculations

Table 3.2 gives the calculated absolute binding free energies for the binding of pyridin-2-amine to ERK2.

TABLE 3.2: Absolute binding free energies for pyridine-2-amine to ERK2 in units of $kcal\ mol^{-1}$.

Simulation	1a	1b	1a to 1b	1b to 1a
$\Delta G^{\ominus}_{Restr,on}$	2.09 ± 0.128	2.56 ± 0.27	1.62 ± 0.042	1.98 ± 0.104
$\Delta G^{\ominus}_{Complex,off}$	127.027 ± 0.176	125.262 ± 0.222	129.77 ± 0.254	127.121 ± 0.108
$\Delta G^{\ominus}_{Restr,off}$	-8.519	-8.559	-8.519	-8.559
$\Delta G^{\ominus}_{Solv,off}$	119.161 ± 0.025	119.161 ± 0.025	119.161 ± 0.025	119.161 ± 0.025
$\Delta G^{\ominus}_{Sym}$	-0.41	-0.41	-0.41	-0.41
$\Delta G^{\ominus}_{Bind}$	-1.847 ± 0.219	-0.512 ± 0.350	-4.12 ± 0.302	-1.791 ± 0.152

3.3.3.1 Single Ligand Results

The binding free energies of ligands bound in sites 1a and 1b were calculated individually without the second ligand present. Results show that the binding of 1a is 3-fold more favourable than 1b. This result is in line with the fact that site 1b is a known cryptic pocket which typically has a lower binding affinity for ligands. Cryptic pockets tend to close rapidly when the ligands are removed which is observed in the simulation of 1b decoupling.

3.3.3.2 Double Ligand Results

The free energy of binding of pyridin-2-amine to sites 1a or 1b while the other ligand (co-binder) is present was also investigated. Boresch style restraints were imposed on the co-binder to maintain its binding mode throughout the simulation. Note, that these restraints were not added using a lambda scheme and were fully switched on throughout, including the end states. The data shows that the binding of 1a is approximately twice as favourable when the 1b binder is present and that 1b has a 4 fold increase in affinity when the 1a binder is already present. The hypothesis is that as one ligand is bound, it can fix the conformation of the wider binding site and therefore allow further ligands to bind more easily.

To check if the calculated free energies are approximately correct (within the realms of simulation) a cycle closure on the thermodynamic cycle presented in Figure 3.3 can be performed. By following this cycle, the sum of all its components should equal zero showing that the free energy of achieving the "Protein+1a+1b" state is independent of the route taken. This means that the energy of binding 1a followed by 1b should equal the binding of 1b followed by 1a. Equation 3.3 details this mathematically.

$$
\begin{aligned}
&(\Delta G_{bind(1a)} + \Delta G_{bind(1b|1a)}) = (\Delta G_{bind(1b)} + \Delta G_{bind(1a|1b)}) \\
&(\Delta G_{bind(1a)} + \Delta G_{bind(1b|1a)}) - (\Delta G_{bind(1b)} + \Delta G_{bind(1a|1b)}) = 0 \\
&(-1.847 + -1.791) - (-0.521 + -4.12) = 1.003\ kcal\ mol^{-1} \\
&Error = \sqrt{0.219^2 + 0.350^2 + 0.302^2 + 0.152^2} = 0.534\ kcal\ mol^{-1}
\end{aligned}
\tag{3.3}
$$

The cycle closure does not quite reach the desired value of zero even within the margin of error indicating possible sampling issues. Future work should focus on converging this towards zero with protocol improvements.

3.4 Summary

In this preliminary section, we aimed to build up a picture of how the ERK2-MiniFrag system behaves in simulation and to ascertain whether certain basic methods would be suitable for predicting binding. We first showed that, as expected, the FTMap and FTSite algorithms can predict hotspots and binding sites when using *holo* structures with pre-formed binding sites. However, when using an *apo* structure, only the more exposed 1a site was found. Owing to the static structure nature of FTMap, sites 1b and 1c were not found in the *apo* structure as side chain rearrangements are required. Furthermore, this project is focused on predicting the binding poses of MiniFrags, and while FTMap and FTSite are somewhat useful as a complementary method, it is limited to using the 16 pre-selected probes, of which only one can be found in the MiniFrag library (phenol). However, there are some similar MiniFrags to the FTMap probes and there may be more in the E-FTMap implementation.[217] Overall the results presented here highlight both the advantages and limitations of static structure methods. One binding site that is well reproduced is a good result, and the time required to obtain these results is orders of magnitude faster than simulation methods. As such, these techniques are very useful in conjunction with other methods and certainly have a purpose and a role to play in FBDD and SBDD.

Second, we performed basic MSMD simulations using known MiniFrags binders as probes. To make the test as realistic as possible, these simulations used a fully equilibrated, true *apo*, structure to ensure no templating and full solvation of the pocket. Almost all of the simulated probes sampled pocket 1a to some degree, however, in some cases the occupancy maps are not too convincing in that it can be hard to distinguish binding in site 1a from the background noise. However, it is fair to say that these occupancy maps are not very sophisticated and the maps used by MixMD[226] may be a future avenue for exploration.

Sampling of pockets 1b and 1c was less successful. This is to be expected as binding in pocket 1b must compete with an occluding lysine residue. Overall, it is clear that

predicting MiniFrag binding with basic MSMD simulations is going to be difficult. While MSMD methods such as MixMD and SILCS have shown success on multiple occasions,[89] they typically use very small probes, such as methanol, to map interaction hotspots on protein surfaces, not complete ligand binding.

In future attempts, it may be worth trying to simply increase the length of the simulation. A suggested value would be 100 ns as this provides a good balance between simulation time and compute time, however, as these are weak binders it would not be surprising to observe no binding in that time frame either. A second suggestion is to increase the concentration of the ligand up to 1 M to align with the experiment.[11] Astex found that at 1 M the MiniFrags are highly soluble in water, however for some fragments, this property does translate into simulation, leading to aggregation, which is a point for consideration in future development.

Finally, ABFE calculations were performed to investigate the strength of MiniFrag binding. These calculations aimed to firstly develop a protocol that could provide some thermodynamic data to complement the experimental screening and secondly investigate approximately the order of magnitude with which particular MiniFrags bind within the limits of the simulation forcefields. The second point is the most important because, in the absence of experimental data, it is unknown how strongly the MiniFrags bind, and if it were found that the binding was so weak, it may be impossible to reproduce in a basic MD simulation. The results show that for pyridin-2-amine, the binding of both ligands is favourable in simulation but the binding to site 1b, without a ligand present in 1a, seems to be very weak. Future work will also be able to give insight into the binding mechanism which could help to explain the predicted values. This should involve a much deeper exploration into the ERK2 system itself.

Overall, the ABFE results suggest that binding is weak and potentially cooperative which is in line with the fact that MSMD simulations struggled to resolve the binding modes of the MiniFrags. Together this implies there is scope for enhanced sampling algorithms in this context. In the following chapters we develop an enhanced sampling approach, GCNCMC, to sample the binding of MiniFrag and MiniFrag-like molecules in an attempt to improve the results observed in this chapter. The ERK2-MiniFrag system is revisited in Chapter 7.

Chapter 4

GCNCMC/MD Development, Implementation and Application to Small Molecules

Some of the text, theory and results presented in this chapter have been published in the paper: Accelerating Fragment Based Drug Discovery using Grand Canonical Nonequilibrium Candidate Monte Carlo authored by WP (DOI: 10.26434/chemrxiv-2024-q9l5z). Another publication is in preparation.

4.1 Introduction

The success of grand canonical methods in enhancing water sampling has been beneficial in structure-based applications. Particularly useful is the ability to place and determine the positions of stable water molecules in occluded binding sites. This was first exemplified by Woo *et al.* who placed water molecules into the KcsA potassium channel successfully reproducing the crystallographically known water sites.[141] Since then, a retrospective study by Samways *et al.* used GCMC to predict non-bulk crystallographic water sites for 108 unique structures with FDA-approved small molecule drugs with an 81.4% success rate.[146] Bodnarchuk *et al.* showed the use of water-based GCMC in a lead optimisation setting by identifying water molecules which could be easily displaced by a modification to a ligand to increase that ligand's binding affinity.[145] The application of GCMC water sampling is now commonplace in free energy calculations and implemented in the popular FEP+ software developed by Schrodinger.[148] Finally, Thomaston *et al.* applied the method to mutants of the influenza A M2 proton channel to understand the effects of these mutants on the generally well-conserved water network.[227]

Using the same philosophy as water-based GCMC sampling, we aim to build a method which can place small molecule fragments. This has been done before to varying degrees. Clark *et al.* and Kulp *et al.* used a purely Monte Carlo sampling scheme to insert a large number of ligands to a gas phase system of T4 Lysozyme; they then annealed the chemical potential such that only the ligands making the strongest interactions are left.[144,228] The negatives of such an approach include a lack of protein flexibility, low acceptance rates and no competition with solvent molecules. Lakkaraju *et al.* developed a GCMC-like approach to enhance SILCS simulations where solvent is explicitly included and MD is used to incorporate protein flexibility. However, in this work, the value of the chemical potential is allowed to fluctuate to maintain a desired concentration, so despite having some success, this method breaks detailed balance and the grand canonical ensemble.[94] In both cases, the acceptance rate of these instantaneous GCMC moves is low.

Here, we further develop the water-based GCNCMC protocol such that our fragment-based GCMC moves use a nonequilibrium switching process to significantly improve acceptance rates.[150] Combining this with explicit solvent MD allows for both protein flexibility and competition with water. In this chapter, we first explain the minor differences between water and fragment-based GCNCMC, we then validate the method by reproducing a bulk thermodynamic quantity, namely concentration, as in previous studies.[147,150] We then apply the technique to a simple test system, β-cyclodextrin, to further understand how the method behaves.

4.1.1 Host Guest Systems

Host-guest complexes, such as β-cyclodextrin (βCD), are convenient and tractable test systems that are often used to test new simulation methods and notably exhibit many of the same characteristics as fragment binding to proteins.[229] Namely, host guest systems present the same molecular interactions, such as hydrogen bonds, desolvation effects and even the ability to restrict ligand conformation to a particular binding mode. Owing to their small size, simulations of host-guest systems generally converge quickly.[229–233] It is for these reasons that host guest systems are becoming more widely used for method and forcefield development, particularly in the context of free energy calculations.[229,230]

Interestingly, it has been reported that guests with a single polar group bind to βCD in two distinct orientations where the polar group points out of either end of the host. Ligands have been shown to bind more favourably in the secondary alcohol orientation (Fig. 4.1).[229]

In this section, we apply various ligand GCNCMC/MD protocols to para-cresol binding to βCD to investigate the effect, if any, of the switching time (τ) on acceptance rates, convergence and desolvation effects.

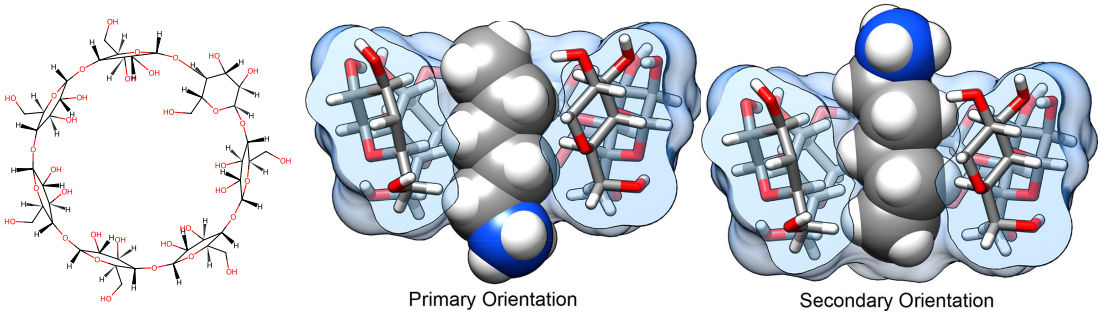


FIGURE 4.1: Left: 2D depiction of the β-cyclodextrin host. Right: Example of a ligand bound in two poses to β-cyclodextrin. The orientation refers to the positioning of the guest's polar group (blue spheres) at the CD opening, which comprises primary or secondary alcohols. Adapted from Henriksen *et al.*[229]

4.2 Theory and Implementation

4.2.1 Application of GCNCMC to Small Molecules

The application of GCNCMC to small molecules is almost identical to that of water molecules and follows the same derivation outlined in the Theory chapter (Sec. 2.5.2.3). In that chapter, the standard state Adams value, that defines the equilibrium to a standard state reference solution, $B^{\ominus}_{eq}$, was defined as:

$$B^{\ominus}_{eq} = \beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right), \tag{4.1}$$

where μ'_{sol} is the excess chemical potential of the molecule of interest in the reference solution, V_{GCMC} is the volume of the GCMC region and $V^{\ominus}$ is the standard state volume.

The standard states for water and small molecules are well-defined as 55 and 1 M respectively. However, in many cases, simulating a molecule, such as a fragment, at a concentration that is not the standard state is more experimentally relevant. For example, fragment-like molecules tend to bind to their targets in the micromolar to millimolar range.[32] In such situations where the molecule in the reference solution (the solution with which our simulated system is in equilibrium) deviates from the standard state concentration, we define the Adams value with a specific concentration dependence to reflect equilibrium with a solution of concentration, c:

$$B_{eq}(c) = \beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V(c)}\right), \tag{4.2}$$

where $V(c)$ is now the average volume occupied by a molecule at concentration, c and can be trivially calculated as:

$$V(c) = \frac{1}{N_A c_L}, \tag{4.3}$$

where N_A is Avogadro's number and c_L is the ligand concentration.

4.2.2 Excess Chemical Potential

A pre-requisite for any GCNCMC simulation is the calculation of the excess chemical potential, μ'_{sol}, of the molecule of interest in a reference solution with which the GCMC region is desired to be in equilibrium. In other words, how favourable that molecule is in a given solution will ultimately decide on the equilibrium between a binding site and said solution. Conceptually, a hydrophobic molecule in water is likely to be 'unhappy' meaning that it would be easier to insert and harder to delete from a hydrophobic protein region.

As described in the Theory, the excess chemical potential of a molecule is formally defined as the change in the excess free energy ($\Delta F'$) when adding a molecule (ΔN) to a given solution ($\frac{\Delta F'}{\Delta N}$) and as such has a concentration dependence whereby the number of existing molecules of the same species can affect μ'_{sol}, particularly at higher concentrations. In other words, the excess chemical potential of a molecule in a solution is equivalent to the solvation free energy of that molecule in that solution. The mentioned concentration dependence is often neglected by traditional solvation free energy calculations which are performed at "infinite dilution" where a single molecule is coupled/decoupled from a box containing only water. At sufficiently low concentrations, such as those at which molecules bind to proteins, this approximation holds particularly well since the probability of interacting with another molecule of the same kind is low and thus acts as if it is in water alone. Further, calculating the excess chemical potential for very low concentrations is difficult in practice, as large simulation boxes are required to achieve such a low concentration of molecules.

At higher concentrations, the probability of a single molecule interacting with another of its kind is greater and thus contributes to the excess chemical potential. The effects of these interactions are complex, hard to predict, and differ greatly between molecules. As an example, the free energy of adding an apolar benzene molecule to a box already containing N molecules of benzene would be much more favourable than adding to a pure box of water.

In this work, similar to traditional free energy calculations, we often approximate the excess chemical potential of a given fragment to be equal to the "infinite dilution"

hydration free energy of that fragment, or in other words, the free energy of adding a fragment molecule to a box of water. For sufficiently dilute concentrations such as those used in some of these studies, it is assumed that the impact of interactions with other molecules of the same species is negligible and that μ'_{sol} is independent of concentration. In previous works using water,[147,150,151] this approximation holds as we have previously been interested in sampling the relationship between a binding site and bulk water, where the excess chemical potential of water in water is indeed its hydration free energy. In summary, for a given molecule, only one value of μ'_{sol} requires calculation, and this value can be applied to any dilute concentration of that molecule. However, there are some exceptions to this when a higher concentration of molecule is used, such as in the following concentration plots. Throughout this report, the value of μ'_{sol} used will be labelled appropriately as infinitely dilute or otherwise.

With μ'_{sol} fixed, the Adams value becomes dependent only on the concentration of fragment in the reference solution, where it is now intuitive to say that a higher reference solution concentration would lead to more binding in the GCMC region of the protein, as shown in Figure 4.2.

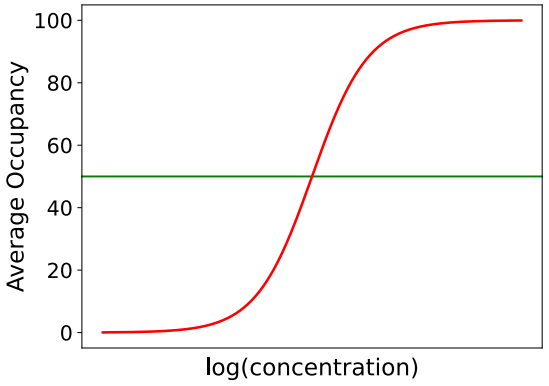


FIGURE 4.2: Typical binding site occupancy for an increasing probe concentration in the reference solution.

4.3 Simulation Details

4.3.1 General Procedure for Excess Chemical Potential Calculations

The general procedure for calculating the excess chemical potential is described here and will be referred to throughout the rest of this thesis. The excess chemical potential, μ'_{sol}, for any particular ligand was calculated using a basic hydration, or solvation, free energy FEP calculation. The molecule of interest is placed in the centre of a simulation box containing only water. If a certain concentration is required for a solvation free energy calculation, ligands are added to the box up to the desired concentration. The system is then equilibrated for 3 ns in the NPT ensemble at 298 K. A Monte Carlo

barostat is used to maintain the pressure at 1 bar with volume updates every 25 steps of MD. The ligand is then decoupled from the box over 30 lambda values with the first 10 turning off the electrostatics and the final 20 the Lennard-Jones interactions. At each lambda value, the system is equilibrated for 1 ns before being run for a further 4 ns with potential energy samples collected every 400 timesteps. Free energies were calculated using the multistate Bennett acceptance ratio (MBAR)[207] as implemented in *pymbar*.[207] Generally, four repeats per ligand are performed with the mean average taken forward. Each lambda window can either be run as an individual simulation in parallel or one simulation can be performed where the lambda schedule is performed serially.

All simulations are performed at 298 K and the MD is performed using the Langevin BAOAB integrator.[178] PME[183] is employed to calculate the long-range electrostatics with a 12 Å cutoff and a switching function applied to the LJ interactions between 10 Å and 12 Å. Note, as methods throughout this thesis evolved, the number of lambda windows, size of the box and ionic concentration varied. However, this is the most general outline of the procedure.

4.3.1.1 Acetone Calibration Curve

To investigate the effect of higher concentrations of ligand on the excess chemical potential, we take inspiration from a publication by Ross *et al.*[212] who calibrated the chemical potential of exchanging water molecules with salt pairs as a function of ionic strength. Here, we use the insertion and deletion functionality of our code to sequentially add acetone molecules to a box of 887 TIP3P waters up to a maximum of 20 molecules, before then deleting each acetone, in turn, to ultimately return to a pure water box. We repeat this cycle 250 times recording the nonequilibrium works of each addition/removal of acetone. These works are then used to calculate an equilibrium free energy using the Bennett Acceptance Ratio.[200] The free energy of adding a molecule to the system is then plotted as a function of N. In an attempt to get an accurate average, 15 repeats were performed. Acetone was parametrized using GAFF[222] with AM1-BCC charges.[223] Insertion and deletions to the system were performed using a switching time of 50 ps.

4.3.2 Concentration Simulations

In the concentration simulations performed here, the target concentrations (0.5 M acetone and 0.1 M pyrimidine) are no longer sufficiently dilute to approximate the excess chemical potential using an infinitely dilute hydration free energy calculation, therefore we require a rigorous parametrization of both species at this specific ligand concentration. Using the same protocol as above, we decouple a molecule of acetone

or pyrimidine from a box already containing 0.5 M or 0.1 M of acetone or pyrimidine respectively. Further, to fully control the concentration of our test systems, we must also perform GCNCMC moves of the water molecules in the box, and therefore we also parametrize the excess chemical potential of water in these two solutions. We found that in both cases the excess chemical potential of water does not differ from that of bulk water since it is still the dominant species in the solution. In this case, the effect of other ligands on the water is likely too sensitive to measure.

For 0.5 M acetone in water, the calculated μ'_{sol} values were -3.25 $\pm$ 0.03 and -6.09 $\pm$ 0.01 kcal mol^{-1} for acetone and water respectively. The average volume per acetone and water molecule was 3360 $\pm$ 0.9 and 31.5 $\pm$ 0.01 Å^3. For 0.1 M pyrimidine, the μ'_{sol} values were -4.49 $\pm$ 0.02 and -6.09 $\pm$ 0.01 kcal mol^{-1} respectively. The average volume per pyrimidine and water molecule was 16312 $\pm$ 9 and 30.6 $\pm$ 0.01 Å^3. The average volume per ligand was calculated by recording the ratio of the number of ligands to box volume throughout a 5 ns NPT simulation at the appropriate concentrations.

The starting points for these tests were equilibrated boxes of pure water containing no other species and boxes containing solutions of 1 M acetone and 0.5 M pyrimidine. We then alternate between GCNCMC moves of the ligand and water to control the concentration of the system in the grand canonical ensemble (μVT). For every 20 ps of MD, one ligand move and three water moves were performed. The switching times for the ligand and water moves were 50 (n_{pert} = 499, n_{prop} = 50) and 10 (n_{pert} = 99, n_{prop} = 50) picoseconds respectively. The GCMC region's volume was the system's total volume: 62 nm^3.

Ten repeats of each ligand from each starting point were performed. The data plotted in Figure 4.5 are the mean averages across all 10 repeats with standard error of the mean shown in the shaded regions. Full details of the simulations performed including the initial concentrations, parameters and final results can be found in Table 4.1. Acetone and pyrimidine were parametrized using GAFF[222] with AM1-BCC charges.[223]

4.3.3 Host Guest Simulations

The structure of the host, βCD, was taken from a benchmark review by Mobley *et al.*[234] and solvated in an 8 Å buffer of TIP3P[221] water and parameterised using the Q4MD-CD forcefield, designed specifically for cyclodextrins.[235] It uses parameters from the GLYCAM04 and Amber99SB forcefields to describe the geometrical and dynamical aspects of cyclodextrins.[236–238] Charge assignment for this forcefield is derived from RESP calculations.[239]

GCNCMC insertion and deletion moves of para-cresol were performed at various switching times (50 ps to 500 ps in 50 ps intervals) while keeping n_{prop} constant (50

steps). The GCMC region was defined as a sphere with a 5 Å radius centred on the host. Simulations were performed at 298 K and the MD was performed using the Langevin BAOAB integrator.[178] PME[183] was employed to calculate the long range electrostatics with a 12 Å cutoff and a switching function applied to the LJ interactions between 10 Å and 12 Å. The parameters of para-cresol were derived from the general amber forcefield (GAFF)[222] with AM1-BCC[223] charges and has an excess chemical potential of -5.13 kcal mol^{-1}.

For these simulations, to collect data efficiently, the simulation was forced to alternate between insertion and deletion moves with valid moves (ones which do not leave the GCMC sphere) being automatically accepted by setting a large B value for insertions and a low B value for deletions. Of course, this is not how a normal simulation would be performed, but it is a useful way to quickly generate concentration independent nonequilibrium work values for diagnostics. Or, in other words, we are simply using our code functionality to generate work distributions independent of the GCNCMC acceptance criteria by alternating insertion and deletion moves.

In another set of simulations, performed with switching times 50, 250 and 500 ps, moves were not cycled between insertion and deletion and moves were accepted or rejected *in situ* according to a B value of -15.5. This is a regular GCNCMC simulation and is used to compare to the above method.

4.4 Results

4.4.1 Effect of Concentration on the Excess Chemical Potential

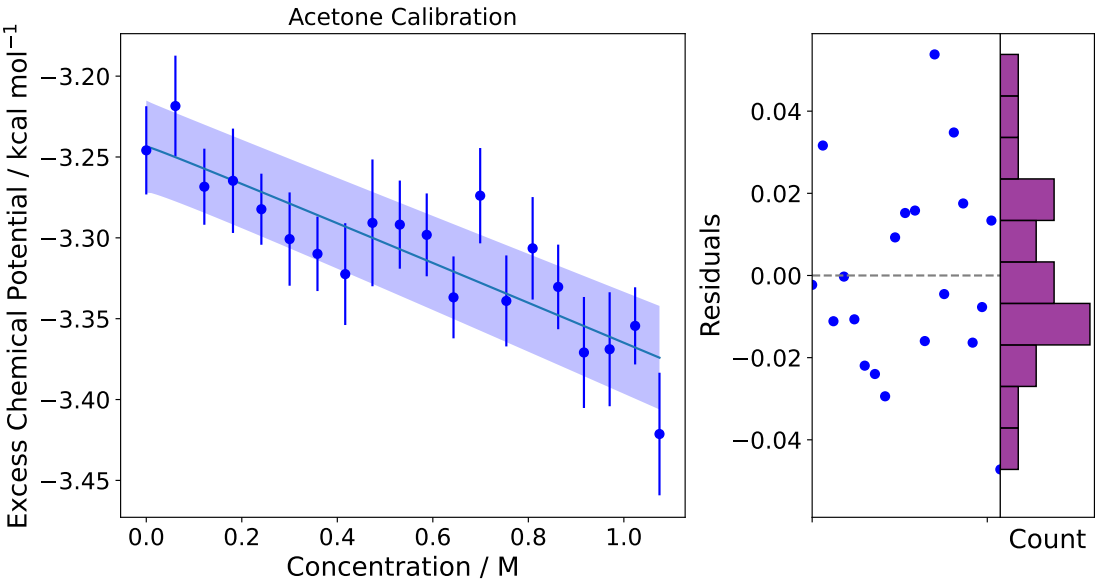


FIGURE 4.3: Excess chemical potential of acetone as a function of ligand concentration. The data were fitted to a function of the form $y = a + b\sqrt{x} + dx$. The shaded region represents the error of the fit.

The number of acetone molecules in the system was cycled up from 0 to 20 and then down from 20 to 0 corresponding to a concentration range of 0 to 1.2 M. The nonequilibrium work for each insertion or deletion was recorded and used to calculate free energy using BAR such that the number of acetones can be plotted as a function of free energy. Fitting a curve to this function means that, in principle, the excess chemical potential for any concentration could be found, though in practice this would require many high quality simulations at a range of concentrations to give precise data. What is clear from Figure 4.3 is that the excess chemical potential is indeed dependent on the concentration, though, at least for acetone, this dependence is small and within the range of approximately 0.2-0.3 $kcal\ mol^{-1}$. Typically, comparing free energies with such a low dynamic range is not recommended, as random noise in the data can effect the trend. Figure 4.4 shows the results from 15 independent repeats on the same plot to stress this element of random noise. Further, in a traditional protein-ligand binding free energy calculation, one would generally not worry about such a small difference in free energy but as we have seen previously, small differences in μ' can have a profound impact on the overall ligand concentration in GCNCMC simulations of bulk ligand-water solutions.[147]

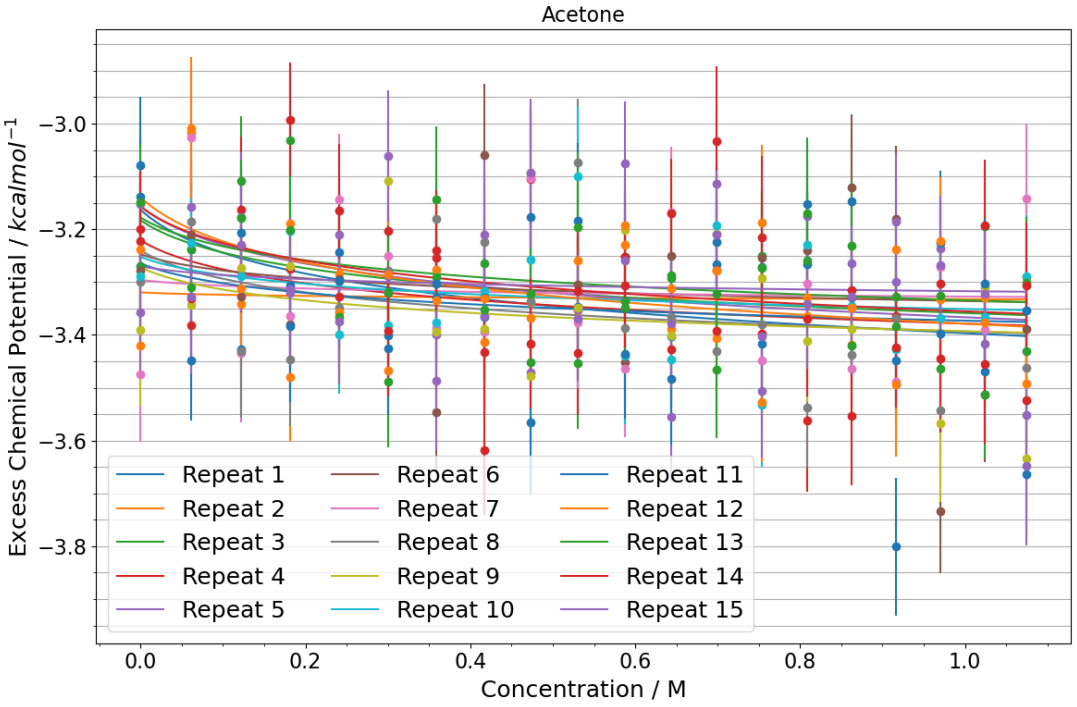


FIGURE 4.4: Excess chemical potential of acetone as a function of concentration for 15 independent repeats. The error bars represent the uncertainty estimation for the BAR estimator in pymbar.

Note that this sort of calibration can be expensive and therefore not viable for a large data set of molecules. As such, it makes more sense to only do so if necessary. Generally, it is easier to perform a single hydration free energy calculation for dilute concentrations or a single solvation free energy calculation at the appropriate concentration.

4.4.2 Effect of Excess Chemical Potential on Concentration

In previous work, we have validated our GCMC and GCNCMC methods by reproducing the mass density of TIP3P water boxes.[147,150] For fragment-water mixtures it is more appropriate to measure the bulk concentration of the fragment in water. We selected solutions of 0.5 M acetone and 0.1 M pyrimidine for this test as they do not aggregate at these concentrations, both experimentally and computationally. The starting concentrations for these tests were equilibrated boxes of pure water containing no other species, and boxes containing solutions of 1 M acetone and 0.5 M pyrimidine. To perform these simulations, we set the GCNCMC parameters appropriate for the target concentrations (Table 4.1). To fully control the concentration of the system we must perform ligand GCNCMC moves and also water moves to maintain the balance between the two species. For example, if we start from a pure water box at the correct density, we would need to be able to delete water molecules to

make space for the ligand molecules. In other words, without water moves, the maximal concentration achievable by the ligand is limited by the size of the box.

Figure 4.5 shows the variation in concentration over simulation time for both fragments. In each case, after an appropriate equilibration period, the concentration fluctuates around the target value, demonstrating that not only can GCNCMC simulations maintain a defined concentration but also rapidly equilibrate the system. The final mean concentration for acetone was 0.55 ± 0.02 M and 0.56 ± 0.02 M when starting from 0 M and 1 M concentrations respectively. While slightly higher than the desired concentration of 0.5 M, the consistency between the two systems is reassuring and indicates that the values of the excess chemical potential of either the ligand or water may not be sufficiently accurate to sample 0.5 M exactly. In the case of pyrimidine, a lower concentration of 0.1 M was selected and well reproduced.

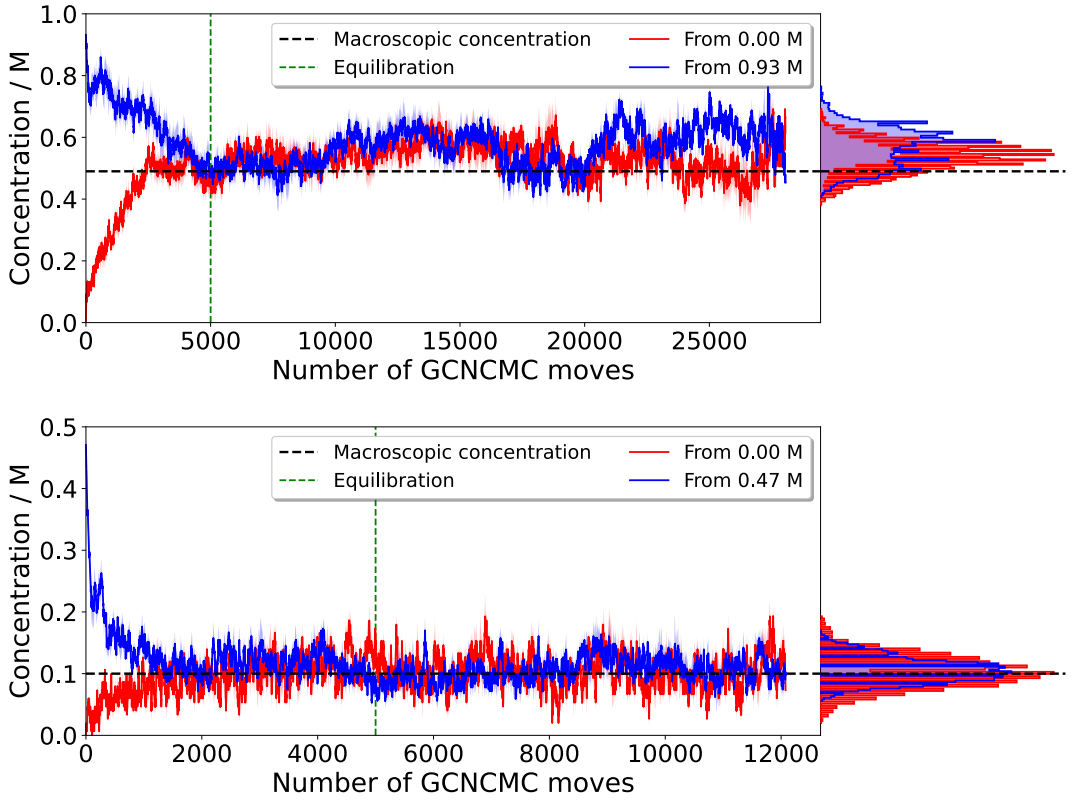


FIGURE 4.5: Fragment concentrations as a function of time. Top: GCNCMC simulation concentrations of acetone starting from a pure water box and a 1 M solution. Bottom: GCNCMC simulation concentrations of pyrimidine starting from a pure water box and a 0.5 M solution. Data points represent the mean concentration at each step over 8 repeats. The shaded regions represent the standard error of the mean. Histograms are binned mean concentrations from after the equilibration point, decided by eye, indicated by the green dashed line.

It has been shown previously that maintaining bulk ensemble properties, such as concentrations, via GCMC and GCNCMC can be very sensitive to the parametrization

of the excess chemical potential, and the large values of N and V_{GCMC} greatly magnify any errors in the calculated parameters.[147,148,150,212] We refer the reader to a publication by Ross *et al.* for a deeper understanding of this issue of sensitivity and fluctuations.[212] Note that for typical protein-ligand applications, this sensitivity becomes less of an issue as there are fewer interacting GCMC molecules, the volume of the GCMC region is smaller, and simulations are performed at more dilute concentrations where the difference in μ'_{sol} values becomes negligible.

TABLE 4.1: Simulation parameters, starting concentrations and final results for the bulk concentration simulations of acetone and pyrimidine.

Ligand (Initial conc.)	Ace. (0 M)	Ace. (0.93 M)	Pyr. (0 M)	Pyr. (0.47 M)
$\mu'_{sol,L}$ / kcal mol^{-1}	-3.25 $\pm$ 0.03		-4.49 $\pm$ 0.02	
$V_L(c_L)$ / Å^3	3360 $\pm$ 0.9		16312 $\pm$ 4	
$\mu'_{sol,W}$ / kcal mol^{-1}	-6.09 $\pm$ 0.01		-6.09 $\pm$ 0.01	
$V_W(c_W)$ / Å^3	31.5 $\pm$ 0.01		30.6 $\pm$ 0.01	
Desired [L] / M	0.49		0.1	
Average [L] / M	0.55 $\pm$ 0.02	0.56 $\pm$ 0.02	0.10 $\pm$ 0.01	0.10 $\pm$ 0.01
Median [L] / M	0.53	0.55	0.11	0.10

To illustrate this point we fixed the excess chemical potential of acetone to its hydration free energy (-3.17 ± 0.02 *kcal mol*$^{-1}$) and performed simulations starting from a pure water box setting the desired concentration to 0.1 M, 0.5 M, and, 1.0 M. These results are shown in Figure 4.6. The results show that as we aim for higher concentrations using the same μ'_{sol} value, we find a greater over estimation of the concentration. This makes sense when we consider the equilibrium and the role of the excess chemical potential. At 1 M, the excess chemical potential of acetone in water (-3.31 ± 0.01 *kcal mol*$^{-1}$) is more negative than what is predicted at infinite dilution owing to favourable interactions with other acetone molecules. By using the infinitely dilute value, the equilibrium is being shifted to the simulated system driving insertion moves away from the supposedly less favourable reference solution into the more favourable, higher concentration, simulated system.

In line with our previous statements, the overestimation becomes less pronounced at lower concentrations, with 0.5 M ($\mu'_{sol} = -3.25 \pm 0.02$ *kcal mol*$^{-1}$) having been slightly over estimated but with 0.1 M being well reproduced where the excess chemical potential is -3.21 ± 0.02 *kcal mol*$^{-1}$. It is clear that at lower concentrations where the value of the excess chemical potential tends towards the infinitely dilute value, the concentrations become well reproduced.

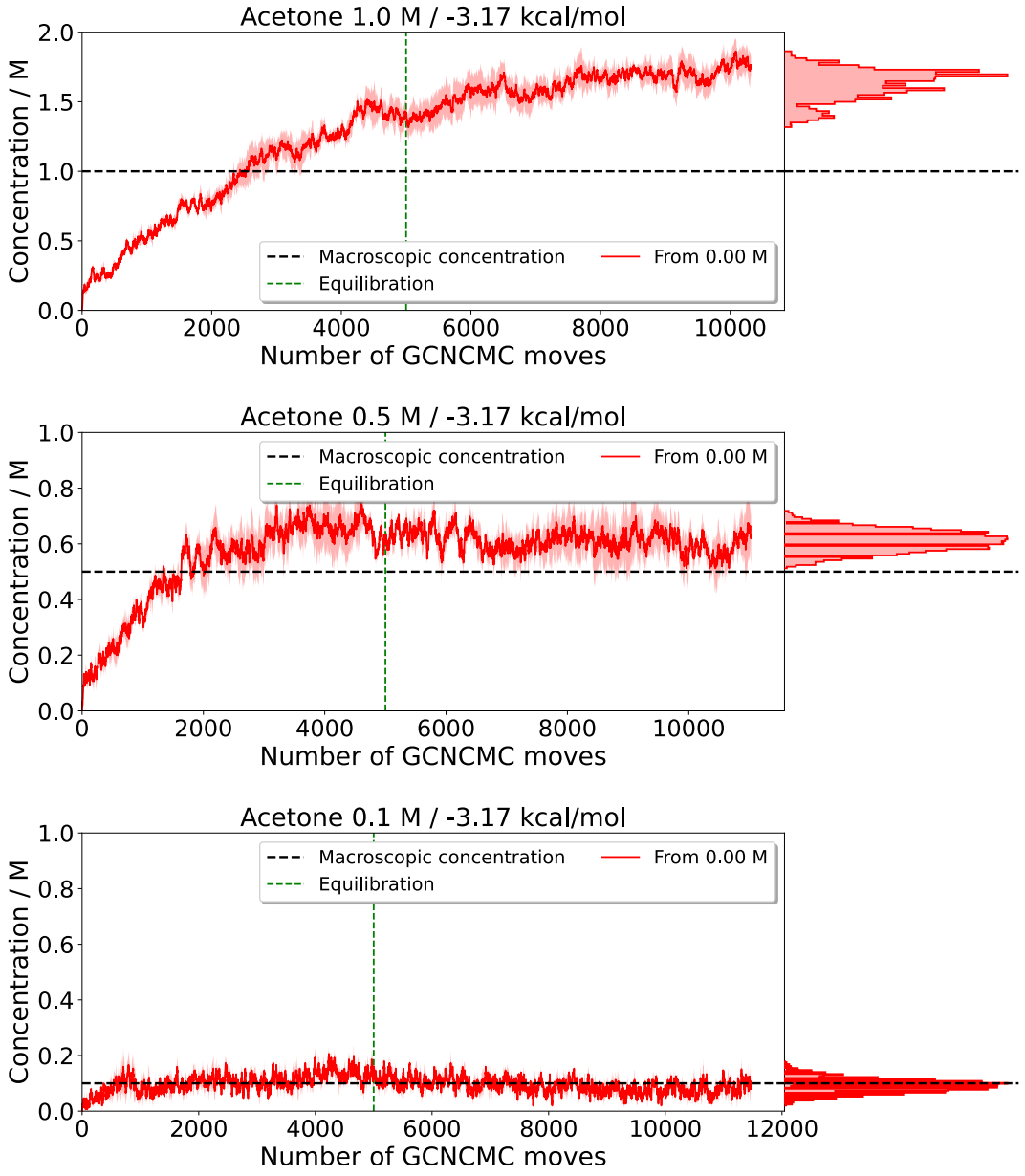


FIGURE 4.6: Acetone concentration as a function of time with a fixed chemical potential. The dashed black line indicates the target concentration. Top: 1.0 M, Middle: 0.5 M, Bottom: 0.1 M.

4.4.3 Host Guest Simulations

4.4.3.1 Work Distributions

Work distributions obtained from GCNCMC insertion and deletion moves at different switching times (τ) are plotted in Figure 4.7. Note, these works are insertion and deletion works where initial N was 0 and 1 respectively and the works plotted have not been accepted or rejected as this would bias the distributions.

In GCNCMC, the acceptance ratio is dependent on the work done throughout the move and, as such, lower work values will encourage a greater proportion of moves to be accepted. It is clear from the plots that the work distributions shift to lower values at higher switching times. Crucially, the plots become more narrow and symmetric around the 'true' value thus lowering the standard deviation of the distribution. This in turn means that each proposed move is more consistent and is accepted or rejected based on more precise and consistent work measurements.

Given these results, it is expected that the acceptance rate at higher switching times will be improved, however, it should be noted that longer switching times result in fewer moves that can be performed in a specified time which can be detrimental to the overall convergence of a particular result. For insertions, it seems as though there is little difference in the work distributions for any switching time greater than 150 ps (Figure 4.8). Although, these distributions are likely to be heavily system-dependent and therefore to fully understand and optimise the protocol, further tests of efficiency will need to be performed in different settings such as pure solvent boxes and protein systems.

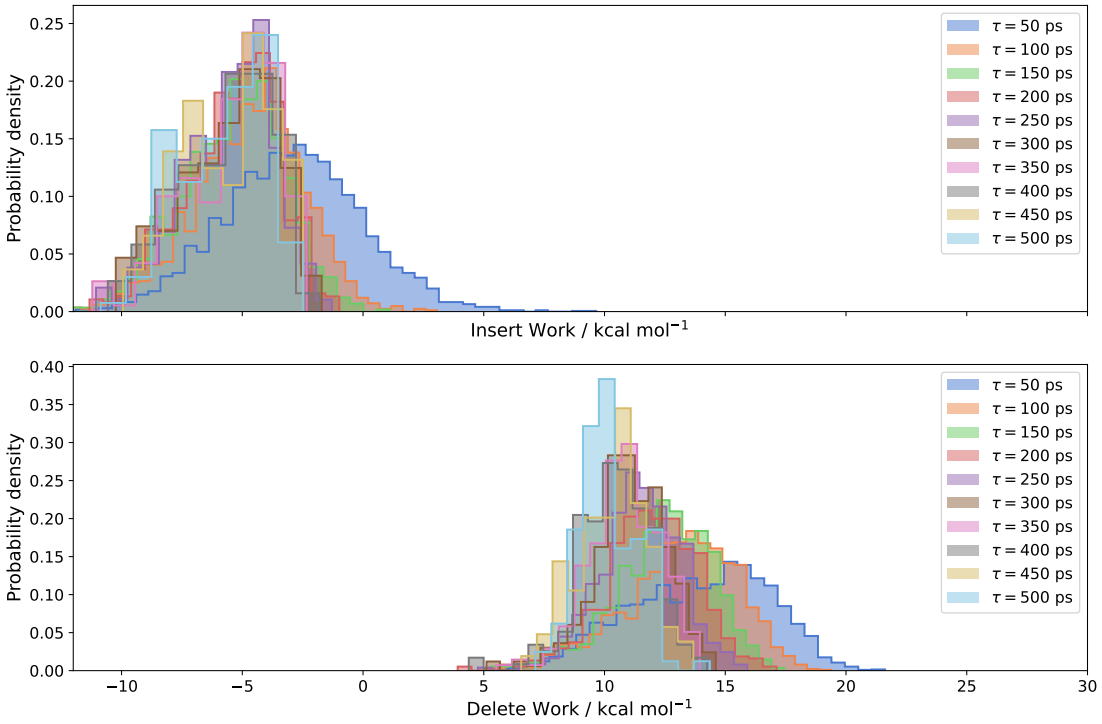


FIGURE 4.7: Work distributions of GCNCMC moves at different switching times. The value of n_{prop} is fixed at 50.

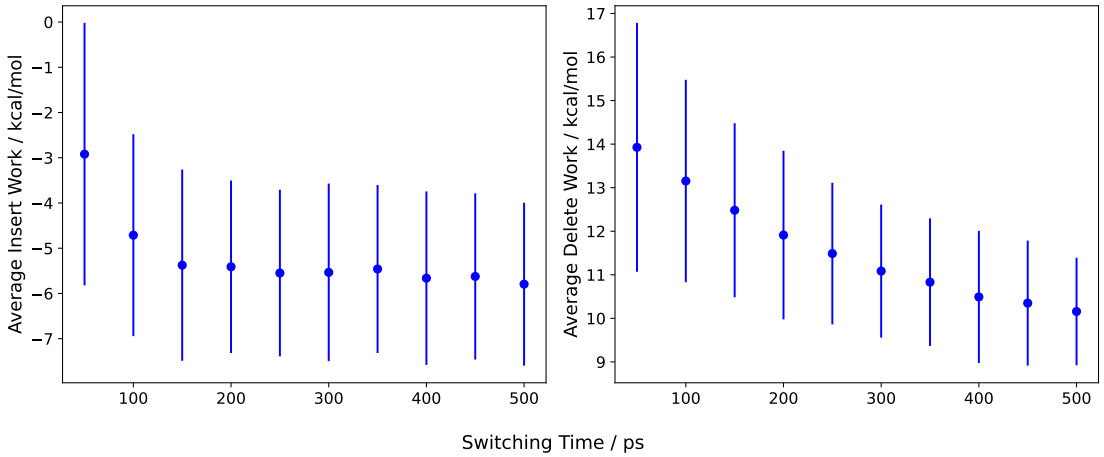


FIGURE 4.8: Average work done at different switching times. Error bars represent one standard deviation. The value of n_{prop} is fixed at 50.

4.4.3.2 Desolvation

The cavities in βCD and other host systems are completely open to solvent and are fully hydrated when there is no guest bound. This will also be the case for many protein binding sites which may be solvent exposed. Therefore it is useful to test the ability of GCNCMC insertions to displace these water molecules as the guest is switched on to see if there is any dependence on the switching time. Figure 4.9 shows the rate at which the waters within the cavity are displaced throughout valid insertion move with switching times of 50, 300, and 500 ps. This is measured by counting the number of waters in the GCMC sphere after each propagation step in the switch. Notably, all three switching times seem to displace the same number of waters throughout the switch. Given the shape of the curves, it is implied that individual waters are displaced once the ligand interactions are roughly switched on by a third, which of course happens later in the move for longer switching times. These results show that water molecules can be displaced during a GCNCMC move, however, in this test case the waters are very solvent exposed and bind weakly. In the following chapters, the ability of GCNCMC to displace buried water molecules or tightly bound waters in protein systems is evaluated.

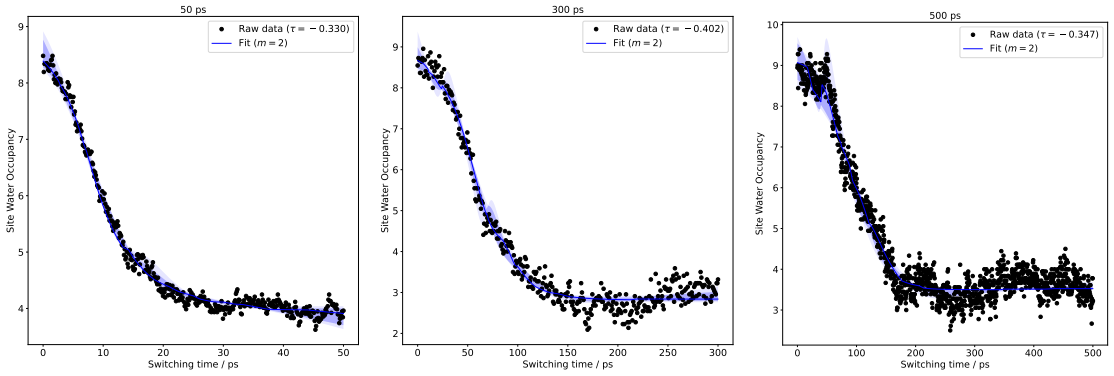


FIGURE 4.9: The effect of switching time on water displacement from the βCD cavity.

4.4.3.3 Convergence

As mentioned, these simulations were set up in such a way that if there is no ligand bound an insertion move is forced and vice versa a deletion move. Each valid move is automatically accepted, and as such, the resulting average occupancy of these simulations should converge to 0.5 exactly assuming all moves are valid. Figure 4.10 (top) shows the convergence at different switching times where it is immediately clear that the simulations do not converge to 0.5, rather each value of τ seems to converge on a different value. This can be explained by the fact that many of our moves are automatically rejected owing to leaving the GCMC sphere as exemplified by Figure 4.11. Unsurprisingly, at longer switching times, where the GCMC molecule spends more time in weakly interacting states, results in a larger proportion of moves being rejected owing to leaving the GCMC sphere. Interestingly, shorter switching times result in occupancies of less than 0.5 implying that the GCMC molecule leaves the sphere more during insertion moves than deletion moves. The opposite can be said for the longer switching times (e.g. 500 ps) where the average occupancy is much greater than 0.5 implying that more deletion moves are becoming invalid. A switching time of 150 ps seems to converge to approximately 0.5 but this is likely system specific. Finally, with increasing switching time comes increased computational cost meaning there is a balance to be found between computational efficiency and high quality moves.

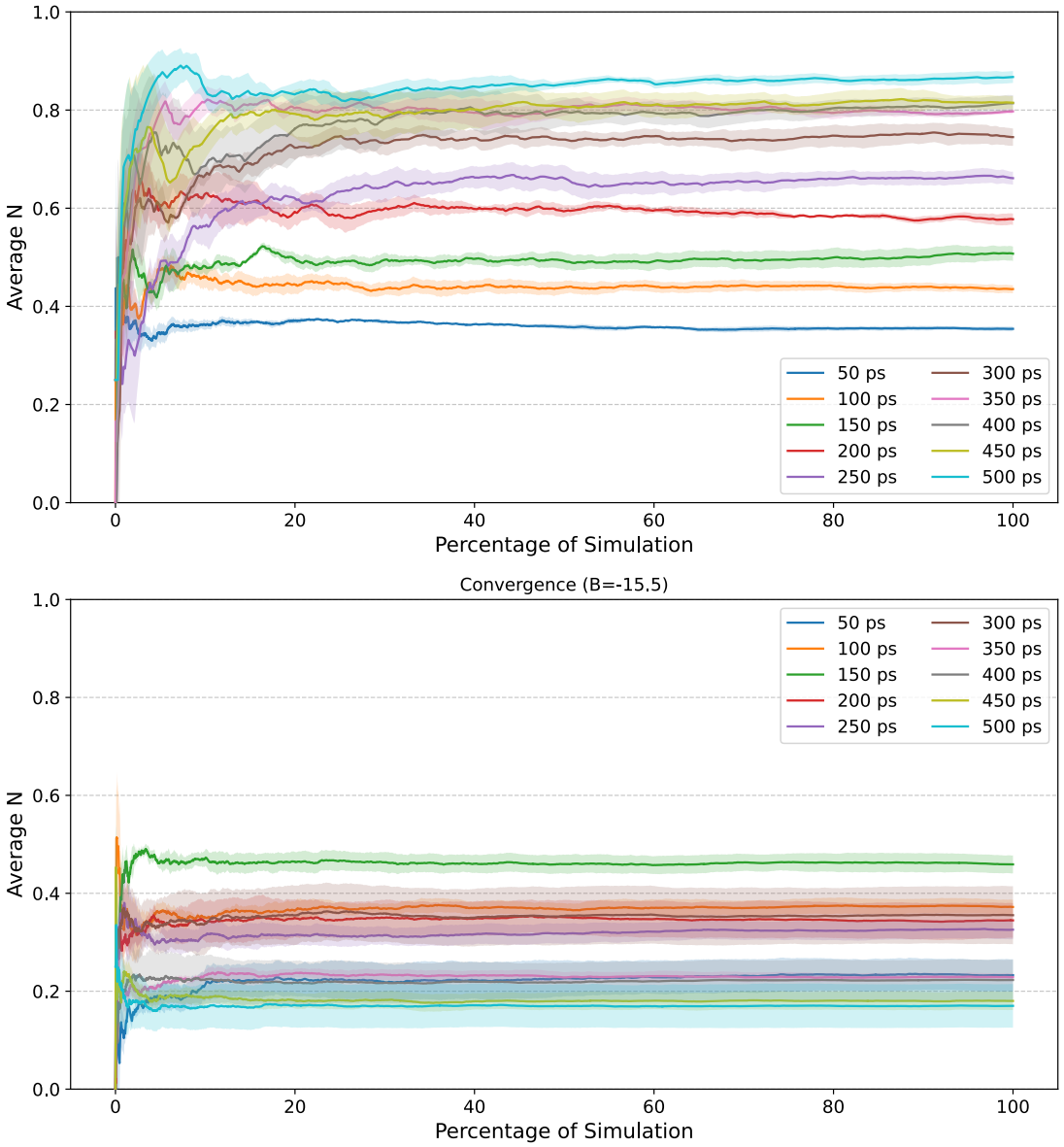


FIGURE 4.10: Convergence of the engineered GCNCMC simulations at different switching times. Top: Before applying acceptance criteria Bottom: After applying the acceptance criteria with a B value of -15.5

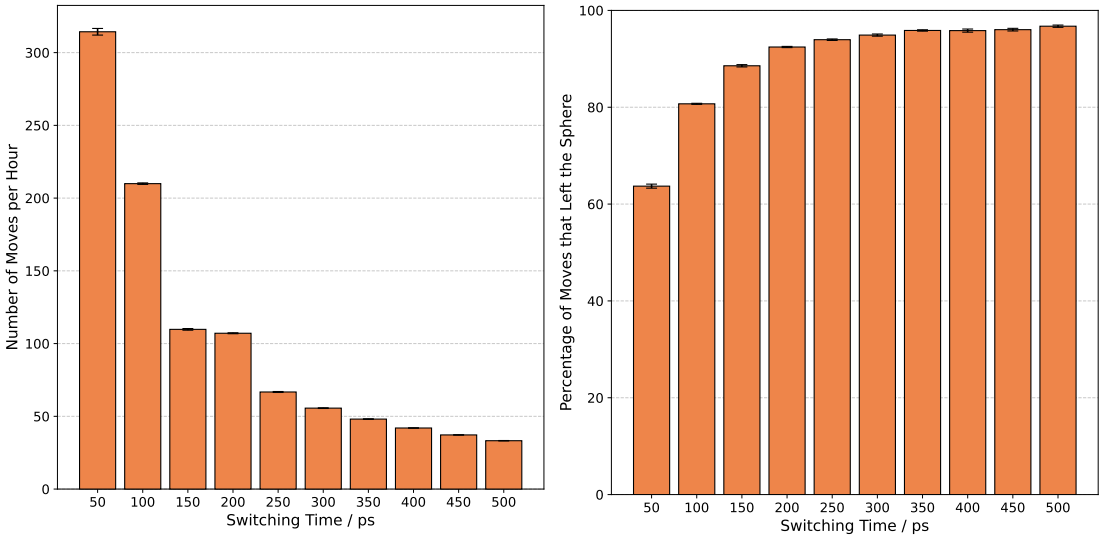


FIGURE 4.11: Left: Number of moves performed per hour on a GTX1080 GPU. Right: Proportion of GCNCMC moves automatically rejected owing to leaving the GCMC sphere.

Regardless of switching time, applying the acceptance criteria to these work distributions should result in the average occupancy converging to the same value (Fig. 4.10 Bottom). This can be done post-simulation by iteratively selecting work measurements from the distributions and applying the acceptance criteria in the same way as if the simulation were being performed. This is shown in the bottom row of Figure 4.10 where it seems that different switching times converge to different values. This is unexpected but may be a result of the finite number of valid work measurements available, particularly at the high switching times. Again, as up to 95% of all moves are discarded, the data in the final work distribution is limited and may cause some bias in the results. More investigation in this area is required.

Furthermore, at higher switching times the ligand may unbind and rebind throughout the switch which could have an adverse effect on the final work measurement as irrelevant regions of the configurational space may be sampled. This was discussed in Section 2.4.2.1. This effect can be seen by using the Bennett acceptance ratio to calculate binding affinities using the nonequilibrium work measurements. Figure 4.12 shows a slight upward trend with increasing switching time, although the range on the plot is only approx. 1.4 kcal mol^{-1}. This is again unexpected as increasing the switching time should converge to one true free energy estimate. It is therefore likely that higher switching times result in the insertion or deletion move sampling more irrelevant regions of configurational space. This is revisited in Chapter 9.

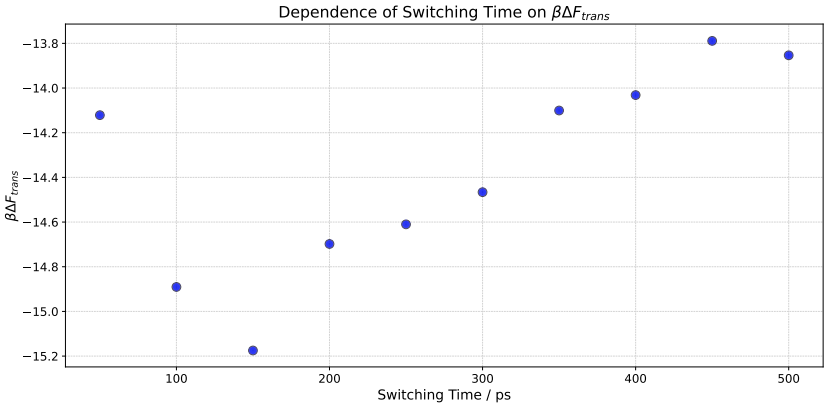


FIGURE 4.12: Effect of switching time on the free energy of transfer from gas to complex.

Crucially, performing regular GCNCMC simulations **without** cycling insertion and deletion moves, and accepting the moves *in situ*, rather than post simulation, results in convergence across all switching times as shown in Figure 4.13. The main difference here to the prior results is that the nonequilibrium works are filtered *in situ* meaning only moves that sample the relevant region of the configurational space are accepted and are therefore not included in the final work distribution. Indeed, using all the valid works, including rejected works, from these simulations and applying the acceptance criteria post simulation results in the same trend as Figure 4.10 (results not shown). Future work should aim to understand this behaviour more deeply.

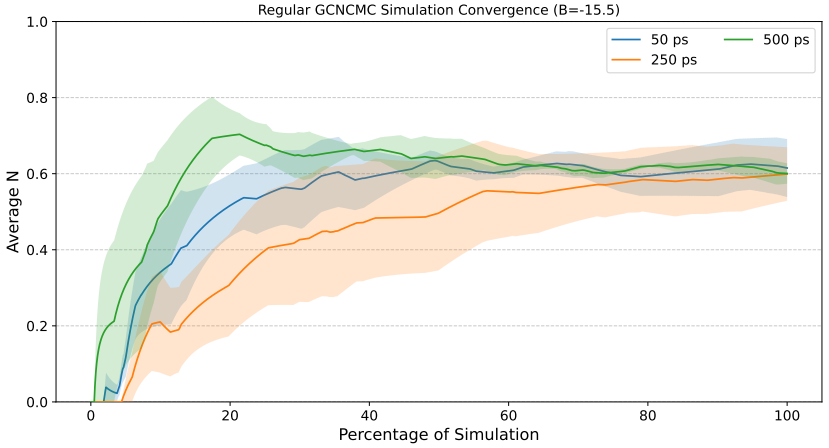


FIGURE 4.13: Average N as a function of simulation time. These are regular GCNCMC simulations with moves accepted *in situ*.

4.5 Summary

In this chapter, we have formulated, tested and validated the development of our GCNCMC method to sample small molecules. Code-wise, this included a large

rewrite of how the forces are distributed and interfaced. The implementation is validated by reproducing ensemble concentrations for two molecules, acetone and pyrimidine. This involved running long GCNCMC/MD simulations, allowing the number of both the molecule of interest and water molecules to fluctuate. It was observed that when using an excess chemical potential accurate for the desired concentration, then that concentration is well sampled in GCNCMC simulation.

A large discussion about the effects of concentration on the excess chemical potential followed, and it was argued that at low concentrations, such as those at which molecules bind, then the excess chemical potential can be assumed to be infinitely dilute. However, for the concentration simulations performed here, that assumption no longer holds and the effect of wrongly calculating the excess chemical potential on the concentration of our systems was demonstrated. In line with previous work, the concentration is remarkably sensitive to the parameterization of μ'.

The method is then applied to a simple host-guest system, namely para-cresol binding to β-cyclodextrin. This test system was used to generally outline the method and study the effect of switching time on the measured non-equilibrium work values. It is demonstrated that longer switching times result in more narrow work distributions centred around the true value, however, this is at the cost of longer compute times and more rejected moves owing to the molecule leaving the GCMC sphere. A potential solution to this is discussed further in Chapter 9. In the following chapter we build on the GCNCMC method to calculate the binding affinities of small fragment molecules.

Chapter 5

Development of a Titration Protocol to Calculate Binding Affinities using GCNCMC

Some of the text, theory and results presented in this chapter have been published in the paper: Accelerating Fragment Based Drug Discovery using Grand Canonical Nonequilibrium Candidate Monte Carlo authored by WP (DOI: 10.26434/chemrxiv-2024-q9l5z). Another publication is in preparation.

5.1 Introduction

As stressed previously, the ability to calculate binding affinity *in silico* is a key factor in computational aided drug discovery and molecular design. In recent years, significant efforts have been focused on improving and designing computational methods for predicting binding affinities.[112,121–123,125,126,234,240]

GCNCMC is a theoretically rigorous method based on statistical mechanics and thermodynamics with a large degree of overlap with non-equilibrium free energy calculations (Sec. 2.4.2.2), whereby the work measurement required for the GCNCMC acceptance criteria is the same work measurement required for the Bennett Acceptance Ratio (Eq. 2.131). Here, we show that the two methods are intrinsically linked and related.

Traditional ABFE approaches, including FEP, often require extensive sampling and the use of restraints to prevent ligands from dissociating during simulations. This can lead to inaccuracies, especially if the chosen restraints enforce incorrect binding modes. The GCNCMC framework provides an alternative that eliminates these restraints by allowing ligands to sample all potential binding orientations freely.

In this chapter, we explore the use of GCNCMC simulations to calculate binding affinities using a novel titration-based protocol. To validate, the method is applied to calculate both basic hydration free energies and binding free energies for the host-guest system β-cyclodextrin.

5.2 Theory and Development

5.2.1 Grand Canonical Integration

Grand canonical integration (GCI) is a method first introduced by Ross *et al.* and details how water-based GCMC simulations can be used to calculate the binding free energy of water networks.[143,151] In a similar vein, we can use this framework to calculate the binding affinities of small molecules using GCNCMC simulations. To aid this derivation, Figure 2.11 is revisited:

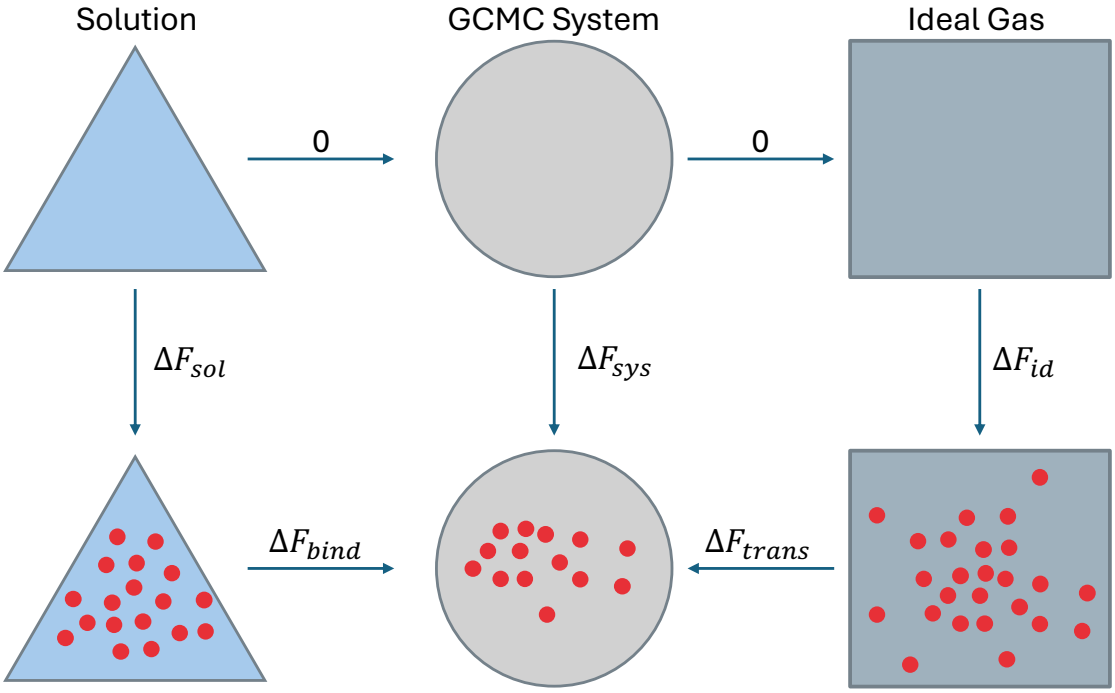


FIGURE 5.1: Figure 2.11 revisited. The thermodynamic cycle links the binding of molecules from solution to the GCMC system with the binding of molecules from the ideal gas. The left triangles represent a solution phase, the circles represent the GCMC region/system and rectangles are the ideal gas. The top row indicates fully empty systems while the bottom row contains some particles indicated by the red dots.

It follows that the free energy of transferring molecules from the ideal gas into the GCMC system can be separated into the following:

$$\Delta F_{trans} = -\Delta F_{id} + \Delta F_{sys} \tag{5.1}$$

where ΔF_{id} is the free energy difference associated with changing the number of molecules in the ideal gas and ΔF_{sys} is the Helmholtz free energy difference of changing the number of molecules in the GCMC system. Using a Legendre transformation we can relate the Helmholtz free energy to the grand potential (Ω):[187]

$$F_{NVT} = \Omega_{\mu VT} + N\mu \tag{5.2}$$

such that:

$$\Delta F_{sys}(N_i \rightarrow N_f) = \Delta\Omega_{sys}(\mu_i \rightarrow \mu_f) + N_f\mu_f - N_i\mu_i \tag{5.3}$$

Equation 2.81 showed that the partial derivative of the grand potential with respect to chemical potential gives the negative of the number of particles in the system. As such the change in grand potential is given by:

$$\Delta\Omega_{sys}(\mu_i \rightarrow \mu_f) = -\int_{\mu_i}^{\mu_f} N(\mu)d\mu \tag{5.4}$$

and substituting this into the above and rewriting in terms of the Adams value (Eq. 2.151) gives:

$$\Delta F_{sys}(N_i \rightarrow N_f) = k_B T\left[N_f B_f - N_i B_i - (N_f - N_i)\ln\left(\frac{V_{sys}}{\Lambda^3}\right) - \int_{B_i}^{B_f} N(B)dB\right] \tag{5.5}$$

The free energy of adding particles to the ideal gas can be calculated analytically. Evaluating Equation 2.93 for N_i and N_f we get:

$$\Delta F_{id}(N_i \rightarrow N_f) = k_B T\ln\left(\frac{N_f!}{N_i!}\right) - k_B T(N_f \rightarrow N_i)\ln\left(\frac{V_{ideal}}{\Lambda^3}\right) \tag{5.6}$$

Combining the two results gives an equation for the transfer free energy:

$$\beta\Delta F_{\text{trans}}\ \left(N_i \rightarrow N_f\right) = N_f B_f - N_i B_i + \ln\left(\frac{N_i!}{N_f!}\right) - \left(N_f - N_i\right)\ln\left(\frac{V_{\text{sys}}}{V_{\text{ideal}}}\right) - \int_{B_i}^{B_f} N(B)\mathrm{d}B \tag{5.7}$$

However, we are often not interested in calculating the free energy of transfer, instead, we are more interested in the free energy of binding from solution to the GCMC system. Using the same thermodynamic cycle we can say that:

$$\Delta F_{bind} = -\Delta F_{sol} + \Delta F_{id} + \Delta F_{trans} \tag{5.8}$$

Note that this is the same relationship that allows us to define an excess chemical potential for a reference solution, rather than the ideal gas in Section 2.5.2.2.

ΔF_{sol} is the free energy change associated with adding molecules to a solution, given by:

$$\Delta F_{sol}\left(N_i \rightarrow N_f\right) = (N_f - N_i)\mu_{sol} \tag{5.9}$$

where

$$\mu_{sol} = \mu'_{sol} + k_B T \ln(\rho_{sol}\Lambda^3) \tag{5.10}$$

Again, combining this with the terms defined above gives an equation for the binding free energy:

$$\beta\Delta F_{bind}\left(N_i \rightarrow N_f\right) = N_f B_f - N_i B_i - \left(N_f - N_i\right)\left\{\beta\mu'_{sol} + \ln\left(\rho_{sol} V_{sys}\right)\right\} - \int_{B_i}^{B_f} N(B)\mathrm{d}B \tag{5.11}$$

Finally, under standard state conditions, we take $\rho^{\ominus}_{sol} = 1/V^{\ominus}$, where $V^{\ominus}$ is the standard state volume. The above can then be rewritten as:

$$\beta\Delta F^{\ominus}_{bind}\left(N_i \rightarrow N_f\right) = N_f B_f - N_i B_i - \left(N_f - N_i\right)\left\{\beta\mu'_{sol} + \ln\left(\frac{V_{sys}}{V^{\ominus}}\right)\right\} - \int_{B_i}^{B_f} N(B)\mathrm{d}B \tag{5.12}$$

In the limit of infinite sampling, the value of N should increase monotonically with B and as such can be represented by a set of sigmoid functions:

$$N(B) \approx \sum_{i=1}^{m} \frac{n_i}{1 + \exp(\omega_{0,i} - \omega_i B)} \tag{5.13}$$

where m is the number of sigmoids, and n_i, $\omega_{0,i}$ and ω_i are the parameters of the i^{th} sigmoid. The integral of this function can be calculated analytically:

$$\int_{B_i}^{B_f} N(B)dB \approx \sum_{i=1}^{m} \frac{n_i}{\omega_i} \ln\left(\frac{e^{\omega_i B_f} + e^{\omega_{0,i}}}{e^{\omega_i B_i} + e^{\omega_{0,i}}}\right) \tag{5.14}$$

5.2.2 GCNCMC/MD Titrations in the Context of Small Molecules

While GCI remains valid for small molecules, we have reframed the method with the aim of making it more intuitive by exploiting the concentration dependence in the Adams value:

$$B_{eq}(c) = \beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V(c)}\right) \tag{5.15}$$

where,

$$V(c) = \frac{1}{N_A c} \tag{5.16}$$

The binding process of a ligand to a protein can be defined by the following equilibrium: L + P $\rightleftharpoons$ LP where the equilibrium constant for the unbinding process, known as the dissociation constant, K_D, is given as a ratio of the concentrations of the species:

$$K_D = \frac{[L][P]}{c^{\ominus}\,[LP]} \tag{5.17}$$

where $[P]$, $[L]$ and $[LP]$ are the molar concentrations of the protein, ligand, and complex respectively and $c^{\ominus}$ is the standard state concentration, taken to be 1 M by convention for a ligand in solution. In the simple, and most common case, of one ligand binding in one binding site, we can calculate K_D as the dimensionless ligand concentration, $\frac{[L]}{c^{\ominus}}$, at the point at which the concentration of the bound protein is equal to the concentration of the free protein, $[LP] = [P]$.

This corresponds to the ligand concentration that binds half of the receptor and manifests itself in a GCNCMC simulation as the ligand concentration, or B value, required so that the receptor is bound for half the simulation (50% occupancy). This also happens to be the concentration that gives equal acceptance probabilities for both insertion and deletion moves, resulting in maximal binding and unbinding events.

Titrating over a range of B values means that the concentration of the solution with which the protein is in equilibrium, is also being titrated, and as such the concentration at which 50% of the ligand binds can be found. The dissociation constant, K_D, is then easily related to the standard binding free energy (Eq. 5.18).

$$\Delta F^{\ominus} = k_B T \ln K_D \tag{5.18}$$

It is now useful to define the Adams value, B, which corresponds to K_D as B_{50} such that:

$$B_{50} = \beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V(K_D)}\right) \tag{5.19}$$

$$B_{50} = \beta\mu'_{sol} + \ln(N_A K_d V_{GCMC}) \tag{5.20}$$

$$K_D = \frac{e^{B_{50} - \beta\mu'_{sol}}}{N_A V_{GCMC}} \tag{5.21}$$

For one ligand binding to one protein we can define the following sigmoid curve (logistic function):

$$N(B) = \frac{1}{1 + \exp(B_{50} - B)} \tag{5.22}$$

where the value of B_{50} is a parameter to be fit by performing a curve optimization. The equivalence of this function to the GCI formulation is presented in Section 5.2.2.3. It is also possible to write this logistic function in terms of the dissociation constant:

$$N(\log_{10}(c)) = \frac{1}{1 + \exp(\log_{10}(K_d) - \log_{10}(c))} \tag{5.23}$$

However, for simplicity, it is easier to plot the occupancies on the B scale to calculate B_{50} and convert to K_D using Equation 5.21.

Finally, it is also useful to state that the value of B_{50} is also the dimensionless free energy of transfer, the proof of which is in a following section (Sec. 5.2.2.2):

$$B_{50} = \beta \Delta F_{trans} \tag{5.24}$$

Figure 5.2 depicts this titration protocol graphically.

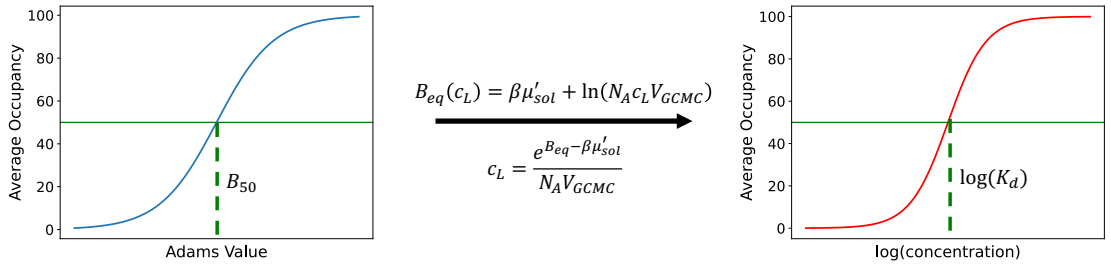


FIGURE 5.2: Graphical representation of the titration protocol. The occupancy at a range of B values is measured and then converted to a concentration.

5.2.2.1 Grand Canonical Acceptance Ratio

Another relationship that is also worth mentioning is that for a system at equilibrium with an ideal gas reservoir, the acceptance of insertion moves must equal deletion moves such that:

$$P_{insert} = P_{delete} \tag{5.25}$$

$$\frac{1}{N+1} e^{B_{50}} e^{-\beta W_{AB}} - N e^{-B_{50}} e^{-\beta W_{BA}} = 0 \tag{5.26}$$

Performing a sum over all the measured work values gives:

$$\sum_{i}^{n_{insert}} e^{B_{50}} e^{-\beta W_i} - \sum_{j}^{n_{delete}} e^{-B_{50}} e^{-\beta W_j} = 0 \tag{5.27}$$

where the value of B_{50} can, like in BAR, be found using numerical methods for any given work distribution of GCNCMC insertion and deletion moves. Note, to simplify the notation, we have assumed that N in the insertion move is 0, and N for the deletion move is 1. The value of B_{50} can then be used to calculate the free energy of transfer from the gas phase using Equation 5.24 or binding affinity using Equation 5.21. One could also substitute Equation 5.24 directly into Equation 5.27 to give a somewhat analogous result to the BAR estimator in Equation 2.131:

$$\sum_{i}^{n_{insert}} e^{\beta(\Delta F - W_i)} - \sum_{j}^{n_{delete}} e^{-\beta(\Delta F + W_j)} = 0 \tag{5.28}$$

Again, Equation 5.28 can be solved numerically for ΔF providing another alternative formulation. We hereafter refer to this approach as the "GCMC Acceptance Ratio" (GAR) owing to its likeness to the Bennett Acceptance Ratio (BAR). Further development and understanding of this relationship are ongoing. We are especially interested in equating GAR and BAR to understand if they are equivalent, or just similar.

5.2.2.2 Equivalence of B_{50} to the Thermodynamic Cycle

Another way of showing that B_{50} is indeed the dimensionless free energy of transfer, is to compare it to the thermodynamic cycle in Figure 5.1. This proof also shows why K_D can be directly calculated from B_{50}.

First, we restate that:

$$\Delta F_{trans} = -\Delta F_{id} + \Delta F_{sol} + \Delta F_{bind} \tag{5.29}$$

and by evaluating each term on the right for $N_i = 0$ and $N_f = 1$, and simplifying, we get:

$$\begin{aligned}
\Delta F_{trans} &= k_B T \ln\left(\frac{V_{ideal}}{\Lambda^3}\right) + \mu_{sol} + k_B T \ln(K_D/c^{\ominus}) \\
&= k_B T \ln\left(\frac{V_{ideal}}{\Lambda^3}\right) + \mu'_{sol} + k_B T \ln\left(\frac{\Lambda^3}{V^{\ominus}}\right) + k_B T \ln(K_D/c^{\ominus}) \\
&= \mu'_{sol} + k_B T \ln\left(\frac{V_{ideal} K_D N_A V^{\ominus}}{V^{\ominus}}\right) \\
\beta\Delta F_{trans} &= \beta\mu'_{sol} + \ln\left(V_{sys} K_D N_A\right)
\end{aligned} \tag{5.30}$$

where the final result of Equation 5.30 is equal to the value of B_{50} given in Equation 5.21. Note that for this derivation, we assume that the volume of the ideal gas and the volume of the system are equivalent in terms of their ideal component as in Ross *et al.*[143]

5.2.2.3 Equivalence of GCI to Logistic Function

Finally, we show that for the case of one ligand binding, the GCI equation and the logistic function (Eq. 5.22) are equivalent. For one molecule binding we find that the GCI equation becomes:

$$\beta\Delta G^{\ominus}_{gci}\,(0 \to 1) = B_f - \left[\beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right)\right] - \int_{B_i}^{B_f} N(B)\,\mathrm{d}B. \tag{5.31}$$

Substituting the logistic function (Eq. 5.22) into Eq. 5.31 we find:

$$\beta\Delta G^{\ominus}_{gci}\,(0 \rightarrow 1) = B_f - \left[\beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right)\right] - \int_{B_i}^{B_f} \frac{1}{1+\exp(B_{50}-B)}\mathrm{d}B \tag{5.32}$$

The integral can now be rewritten as:

$$\begin{aligned}\int_{B_i}^{B_f} N(B)\mathrm{d}B &= \int_{B_i}^{B_f} \frac{1}{1+\exp(B_{50}-B)}\mathrm{d}B \\ &= \int_{B_i}^{B_f} \frac{\exp(B-B_{50})}{1+\exp(B-B_{50})}\mathrm{d}B\end{aligned} \tag{5.33}$$

and by using the following relationship:

$$\frac{\mathrm{d}}{\mathrm{d}x}\ln(1+\exp(x)) = \frac{\exp(x)}{1+\exp(x)} \tag{5.34}$$

the integral can be evaluated analytically as:

$$\begin{aligned}\int_{B_i}^{B_f} N(B)\mathrm{d}B &= \ln[1+\exp(B-B_{50})]\,|_{B_i}^{B_f} \\ &= \ln\left[\frac{1+\exp(B_f-B_{50})}{1+\exp(B_i-B_{50})}\right].\end{aligned} \tag{5.35}$$

For the logistic equation to be valid between 0 and 1, the concentrations, and therefore B values, must be chosen such that $N(B_i) = 0$ and $N(B_f) = 1$. We therefore require that $B_i - B_{50} \ll 0$ and $B_f - B_{50} \gg 0$.

When $(B_i - B_{50}) \rightarrow -\infty$ we get:

$$1+\exp(B_i-B_{50}) = 1 \tag{5.36}$$

and as $(B_f - B_{50}) \rightarrow \infty$ we get:

$$1+\exp(B_f-B_{50}) = \exp(B_f-B_{50}) \tag{5.37}$$

Putting these two limits into Eq. 5.33 shows that:

$$\int_{B_i}^{B_f} N(B)\mathrm{d}B = B_f - B_{50} \tag{5.38}$$

Substituting this result into Eq. 5.31 we get:

$$\beta\Delta G^{\ominus}_{gci} = B_f - \left[\beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right)\right] - B_f + B_{50} \tag{5.39}$$

Simplifying and substituting in Eq. 5.21 for B_{50} we get:

$$\begin{aligned}
\beta\Delta G^{\ominus}_{gci} &= -\beta\mu'_{sol} - \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right) + B_{50} \\
&= -\beta\mu'_{sol} - \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right) + \beta\mu'_{sol} + \ln(N_A K_d V_{GCMC}) \\
&= \ln(N_A K_d V_{GCMC}) - \ln\left(\frac{V_{GCMC}}{V^{\ominus}}\right) \\
\beta\Delta G^{\ominus}_{gci} &= \ln(N_A K_d V^{\ominus}\ )
\end{aligned} \tag{5.40}$$

given $N_A V^{\ominus} = 1M^{-1}$ we finally arrive at a well known equation showing that binding affinity can be calculated either via the GCI equation or from B_{50} directly.

$$\begin{aligned}
\beta\Delta G^{\ominus}_{gci} &= \ln(N_A K_d V^{\ominus}\ ) \\
\Delta G^{\ominus}_{gci} &= \beta^{-1}\ln(K_d/c^{\ominus}\ ) \\
\Delta G^{\ominus}_{gci} &= \Delta G^{\ominus}_{log}.
\end{aligned} \tag{5.41}$$

5.2.2.4 Summary of Key Results

In this theory section, we have derived two novel methods for calculating binding affinity from GCNCMC simulations. The first, in line with previous work, uses a titration protocol where a given ligand can be simulated at multiple B values as a proxy for concentration. From this, a logistic function can be fitted such that the value of B_{50}, or K_D, can be determined and easily related to the free energy of binding. The key equations for the titration protocol are 5.21, 5.22, 5.23 and 5.18.

The second formulation, named "Grand Canonical Acceptance Ratio", or GAR, builds from the fact for a system in equilibrium with a gas, the acceptance probability of insertion moves must equal that of deletion moves from which we derive Equation 5.28. This equation takes a similar form to the Bennett acceptance ratio and can be solved analogously using numerical methods. In both cases, we show how they are intrinsically linked to the thermodynamic cycle shown in Figure 5.1.

5.2.3 Practical Details and Sampling

In the above sections, the mathematical relationships detail various ways in which binding affinities can be calculated in GCNCMC simulations. Here, the sampling scheme is discussed in further depth.

In traditional free energy calculations, there is often concern about the sampling of the 'correct' binding event. In other words, is the ligand and/or protein sampling the relevant configurational space while the alchemical transformation is applied? For

example, absolute binding free energy calculations often use restraints to prevent the ligand from unbinding as it is decoupled. Suppose these restraints restrict the ligand to a pose that is not relevant, or just simply wrong, then the final free energy estimate will be inaccurate. In bi-directional non-equilibrium free energy calculations, there is a requirement that the forward and reverse work distributions overlap, or are symmetric. If the forward move samples one binding mode, and the reverse samples another, the final estimate will be erroneous. As another example, Baumann *et al.*[197] showed that if the forward transition sampled a particular side chain rearrangement which is failed to be captured in the reverse transition, then the work distributions are no longer symmetric and the free energy estimate by the BAR equation will be wrong.

With this in mind, any sampling scheme which restricts a fragment to one particular binding mode, either intentionally with restraints or unintentionally owing to kinetic barriers, would require each relevant binding mode to be sampled in separate simulations. Mobley *et al.*[128] showed that the final free energy estimate for ligands with multiple binding modes can be calculated using a Boltzmann average:

$$\Delta G^{\ominus} = -\beta^{-1} \ln \left[\exp(-\beta \Delta G_1^{\ominus}) + \exp(-\beta \Delta G_2^{\ominus}) \right] \tag{5.42}$$

where $\Delta G_1^{\ominus}$ and $\Delta G_2^{\ominus}$ is the free energy estimate for the binding modes 1 and 2.

This argument can be extended to symmetrically equivalent binding poses. Using a benzene molecule as an example, benzene could bind identically in 12 different ways. These binding modes need to either be explicitly simulated, or corrected for, in the final free energy estimate. For a molecule with 2 or more symmetric binding poses of equal binding affinity, the final free energy estimate can be corrected with the following:

$$\Delta G_{sym}^{\ominus} = -\beta^{-1} \ln(\sigma), \tag{5.43}$$

where σ is the symmetry number of the molecule.

Within the GCNCMC framework, there is no orientational or binding mode fixing restraints, making each insertion move free to sample any potential binding mode. It is then up to the acceptance criteria to discern whether that binding mode is thermodynamically likely at a given concentration. Intuitively, less favourable binding modes will not be accepted at lower concentrations. Interestingly, deletion moves are different as they generally occur after an insertion move is accepted and as such each deletion move would only be from the thermodynamically relevant binding modes, but could in principle sample other modes during the decoupling.

5.3 Simulation Details

5.3.1 Hydration Free Energies

To test all of the above relations in a fast and tractable manner we have calculated solvation free energies for a sub-set of the FreeSolv database.[153] We compare results to equilibrium FEP and nonequilibrium FEP calculations.

All hydration systems were set up by placing a single molecule into the centre of a 30x30x30 Å box of TIP3P water.[221] The ligands (Fig. 5.3) were parametrized using GAFF[222] with AM1-BCC charges.[223] All simulations were performed using OpenMM 8.0[152] using the *grandlig* python module to set up, and control, the required CustomForces for the free energy calculations. In all simulations, PME[183] was used to calculate long-range electrostatics. Lennard Jones interactions were cut off after 12 Å using a switching function between 10 and 12 Å. Simulations were performed at 298 K. Where appropriate (NPT simulations) a Monte Carlo barostat was used at a pressure of 1 bar. Bonds to hydrogen atoms were constrained using OpenMM's default settings and a 2 fs timestep was used.

For equilibrium free energy calculations 20 lambda values were used with each window starting from the same configuration. Each window was minimized and equilibrated for 1.5 ns (0.5 ns NVT, 1 ns NPT). Production simulations (NPT) were run for a total of 4 ns per lambda recording the configuration's potential energy at each lambda value, every 0.2 ps. The trajectories of simulations at the end states (λ = 0 and λ =1) were recorded and snapshots from these trajectories were used as the starting points for the nonequilibrium switches. The multistate Bennett Acceptance Ratio, as implemented in *pymbar*, was used to estimate the free energy of hydration.[207]

For the nonequilibrium calculations, 100 equally spaced snapshots were extracted from the equilibrium trajectories of the two end states. The ligand is then switched from its starting state to the opposite state (on to off or off to on). Each switch occurs over 1500 lambda values with 0.1 ps of MD between each change of lambda giving a total switching time of 150 ps. The work done over the course of each switch is recorded and the free energy is estimated using the various approaches discussed above; GCNCMC titrations, GAR and BAR.

Given the simplicity of this system, GCNCMC titrations were not performed as individual simulations. Instead, the nonequilibrium works were post-processed to titrate over a large range of B values by simply calculating the average occupancy a given B value would have by randomly selecting works, with replacement, from the distribution and repeatedly evaluating the acceptance criteria (Eqs. 2.169 & 2.171).

All simulations were performed and analysed as 3 independent replicates and all the data are reported as the mean free energy with errors reported as the standard error of the mean.

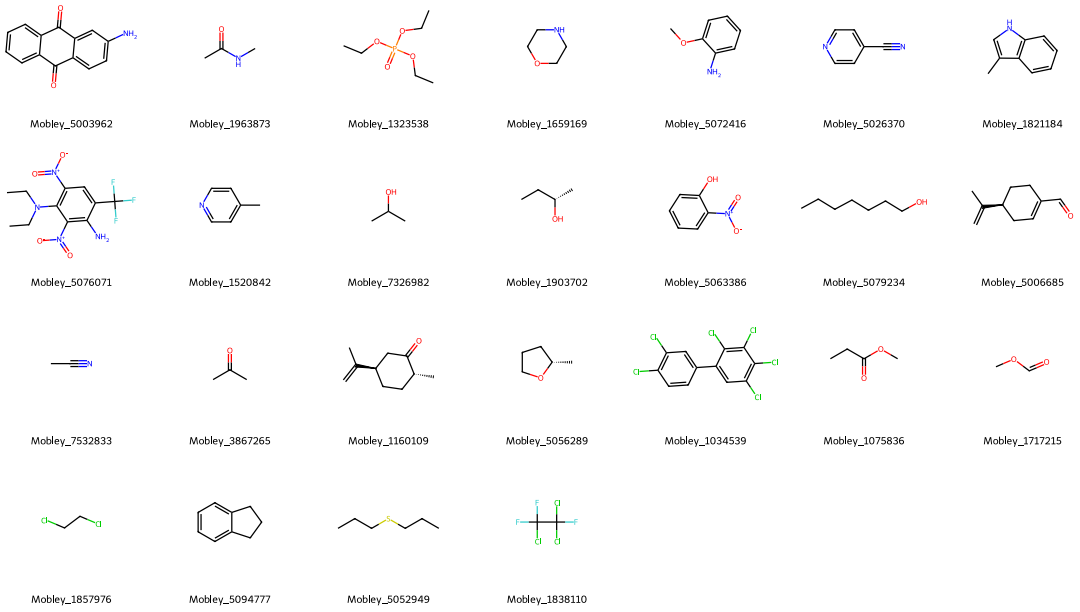


FIGURE 5.3: Subset of molecules from the FreeSolv database used for this study.

5.3.2 Host Guest Titrations

The coordinates of the host, βCD, were taken from a review by Mobley *et al.*[234] and solvated in TIP3P[221] water with an 8 Å buffer. GCNCMC/MD titrations were performed with a switching time of 50 ps ($n_{pert} = 499$ and $n_{prop} = 50$). The GCMC region was defined as a sphere with a 5 Å radius centred in the host cavity midway between two carbon atoms on either side of the host. A total of 22 ligands, from two different studies, with experimental binding affinities were selected for titration and are shown in Figure 5.4.[229,241] Excess chemical potentials for the ligands were calculated as described before (Sec. 4.3.1). Titration B values (n=20) were chosen to loosely surround the experimental dissociation constant. 1700 cycles of GCNCMC/MD were performed at each B value with the first 200 being discarded as equilibration. In these simulations, each cycle consisted of a GCNCMC move attempted for every 1 ps of MD. Each B value was simulated for four repeats.

The ligands (Fig. 5.4) were parametrized using GAFF[222] with AM1-BCC charges.[223] All simulations were performed using OpenMM[152] using the *grandlig* python module. In all simulations, PME[183] was used to calculate long-range electrostatics. Lennard Jones interactions were cut off after 12 Å using a switching function between 10 and 12 Å. Simulations were performed at 298 K.

Final average occupancies from each B value were used to plot the titration curve (Eq. 5.22). The nonequilibrium works resulting from GCNCMC moves were stored and used to calculate affinities using GAR (Eq. 5.28).

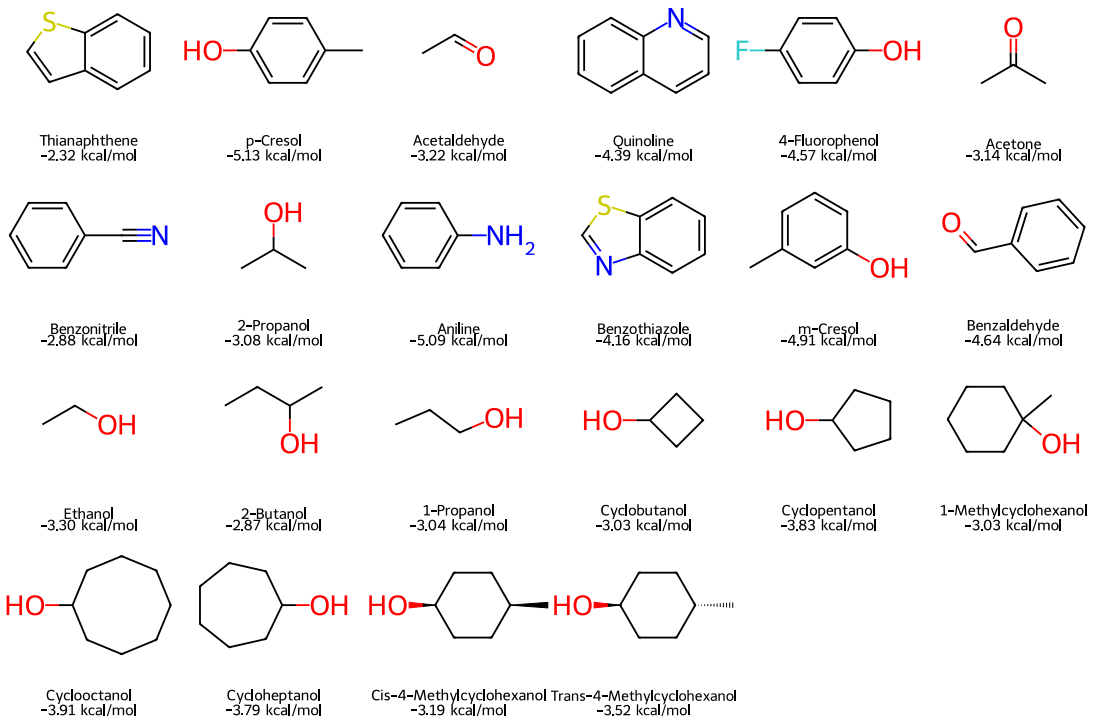


FIGURE 5.4: Guest ligands for the binding to β-cyclodextrin and their calculated values of excess chemical potential, μ'_{sol}.

To validate the free energies obtained by titration, we use a more traditional FEP approach. For each ligand binding to βCD, we selected the most stable pose for each of the two binding modes (primary and secondary) from our GCNCMC simulations as the starting coordinates for the free energy calculations. Two repeats for each binding mode were performed, giving a total of four simulations per ligand. As we found it difficult to define stable Boresch restraints for this host-guest system, we used a spherical flat bottom restraint, with a radius of 5 Å and a force constant of 0.6 kcal mol^{-1} Å^{-2}, to keep the ligands bound. The ligands were then decoupled over 40 lambdas recording potential energy samples every 3 ps. The bound leg free energy of the two binding modes was calculated individually using MBAR and combined using a Boltzmann average. The analytical standard state correction for the restraint was calculated as -0.68 kcal mol^{-1}. The final free energy was calculated according to the thermodynamic cycle in Figure 2.9, reusing the excess chemical potential for the solvent leg.

5.4 Results and Discussion

5.4.1 Hydration Free Energies

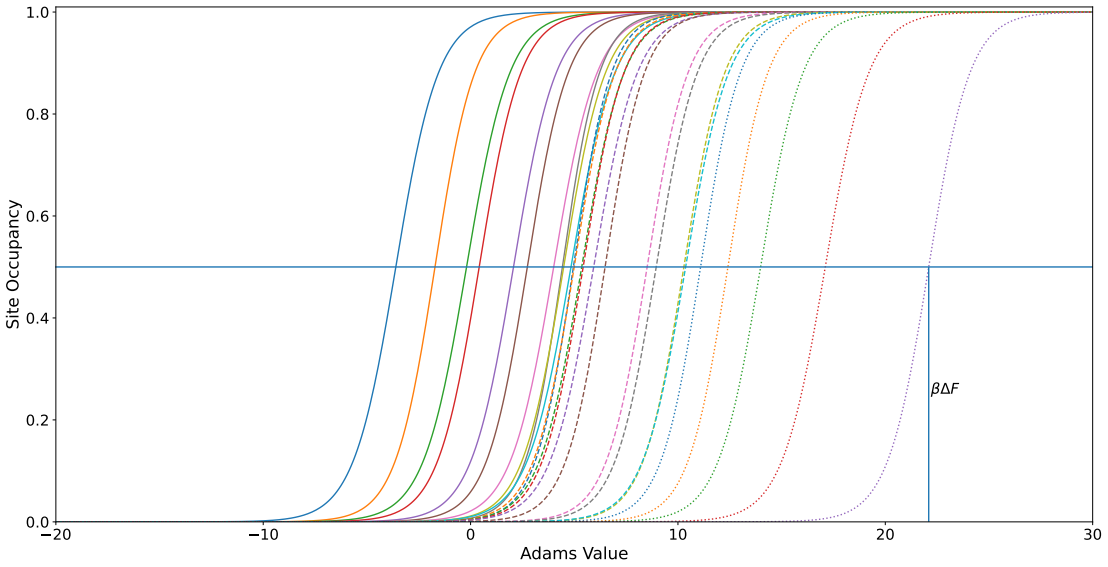


FIGURE 5.5: Modelled titration curves for the FreeSolv molecules tested. Final free energy estimates are derived from the Adams value that gives 50% system occupancy (B_{50}).

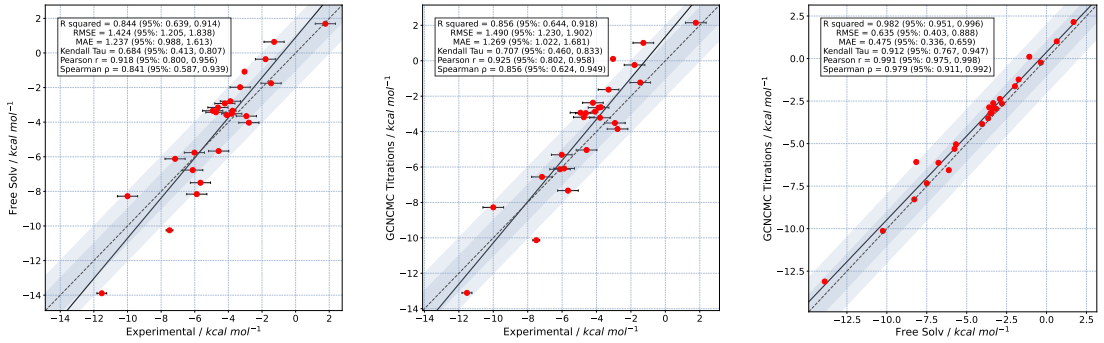


FIGURE 5.6: Hydration free energies for the subset of FreeSolv molecules studied. Left: Published computational FreeSolv data versus experiment. Middle: GCNCMC titrations versus experiment. Right: GCNCMC titrations versus published computational FreeSolv data

Figure 5.5 shows the modelled titration curves for all of the ligands studied, and Figure 5.6 shows the agreement between the titration results, the experimental data, and computational data published in the FreeSolv[153] database. In both cases, a good correlation is observed with low average errors. Crucially, the rank order of ligands, indicated by the Kendall tau (τ), with respect to experiment is particularly high (0.707) and marginally improves upon previously published data (0.684).

To assess the validity of the proposed estimators we also compare the calculated free energies from each estimator using the same nonequilibrium works to a more traditional equilibrium FEP calculation, using potential energy samples and the MBAR estimator. These results are shown in Figure 5.7 which first shows the agreement between the three nonequilibrium methods (using the same works) BAR, the GCMC acceptance ratio (GAR) and GCNCMC titrations. Crucially, consistency with the equilibrium MBAR estimator is observed in all cases. This further confirms that equilibrium free energies can be derived from a quality set of bi-directional work distributions using any of the expressions.

In summary, these results highlight that using the same set of calculated nonequilibrium works but processed in different ways, achieves consistency between all methods. This is shown empirically in Figure 5.7 and confirms the relationships derived earlier in the chapter.

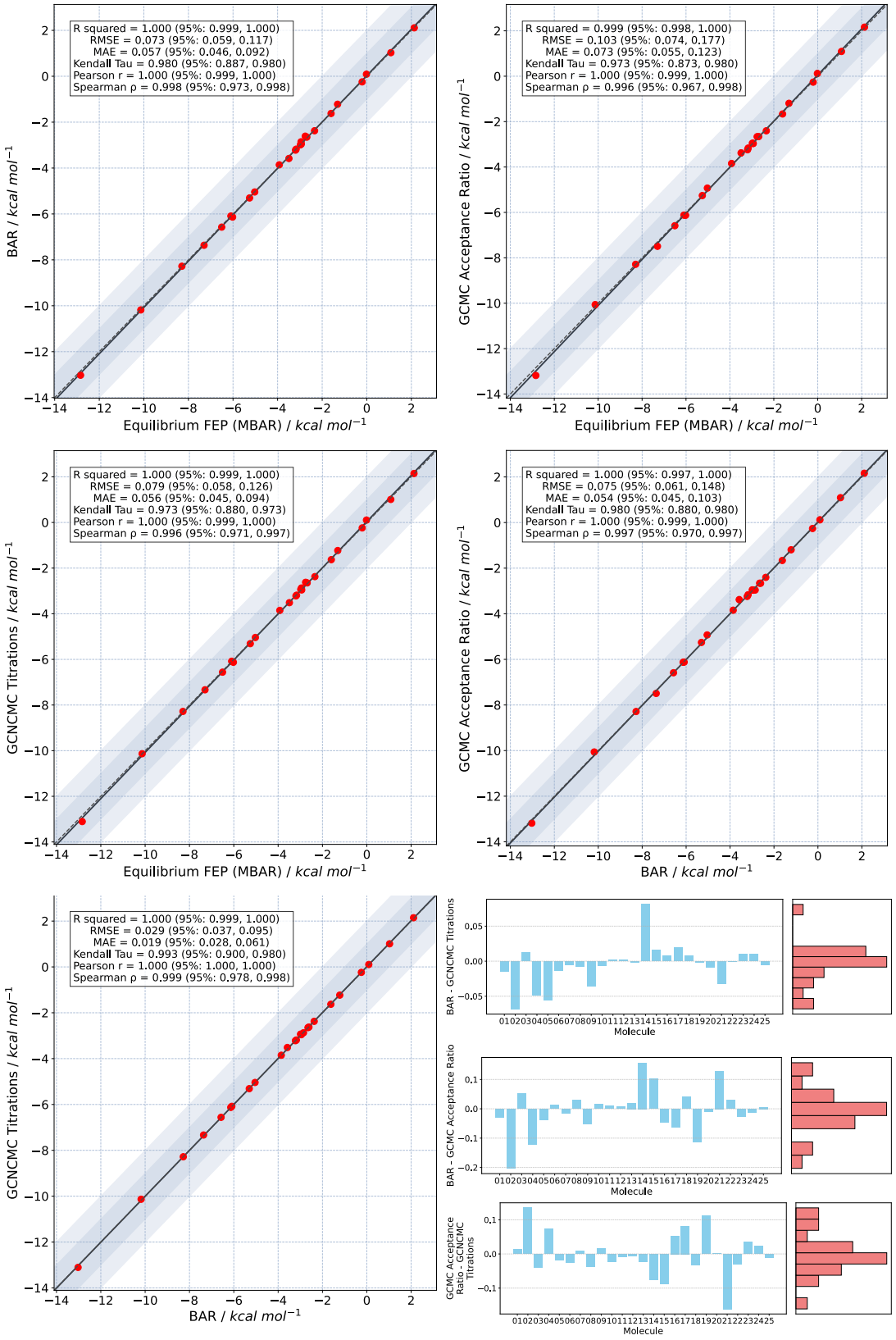


FIGURE 5.7: Correlation plots of hydration free energies for a subset of the FreeSolv database. Amongst all methods and estimators, a perfect correlation is observed. The three figures in the bottom right show the residuals between various methods as indicated on the y axes. A normally distributed error is indicative of random noise.

5.4.1.1 Validation of Excess Chemical Potential Calculations

Although not directly related to this chapter, it is a good place to further validate our hydration free energy and excess chemical potential protocol. Throughout this work, a large number of hydration free energy calculations have been performed as a means of calculating the excess chemical potential of the ligands studied with GCNCMC. Here, we have plotted our data for all those with an entry in the FreeSolv database. In total, 51 of the 140 ligands studied have an entry based on matching the IUPAC names of our ligands to the database. This provides reassurance that our calculations of the excess chemical potential are rigorous and accurate.

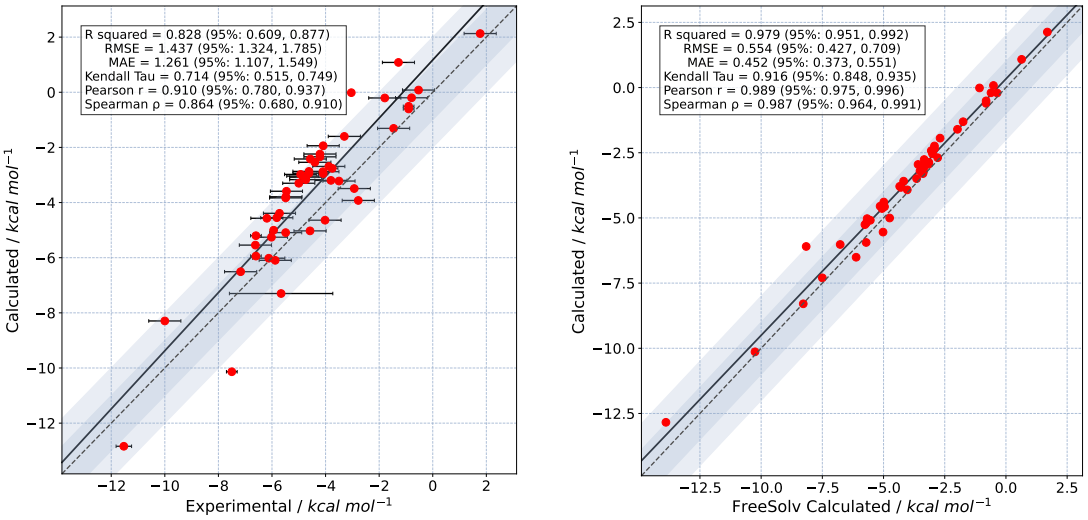


FIGURE 5.8: Calculated values of μ'_{sol} for all ligands studied in this thesis with an entry in the FreeSolv database (51/140). Left: Calculated excess chemical potentials compared to experiment. Right: Our calculated values against those calculated using equilibrium FEP from the FreeSolv database.

5.4.2 Host Guest Simulations

Using a pre-calculated excess chemical potential for each β-CD ligand, as would be necessary for any binding free energy calculation, we simulate over a range of ligand concentrations via the Adams value. We then measure the average occupancy of the simulation at each value of B_{eq} and fit a logistic function, Eq. 5.22, to these data. The fragment concentration for which the corresponding average occupancy is 50% is the dissociation constant, K_D, and is easily related to the free energy of binding (Eq. 5.18).

Individual titration curves for 22 guests binding to βCD can be plotted together, giving a quick and easy indication of the strongest (far left) and weakest (far right) binding fragments (Fig. 5.9). These plots give valuable information concerning the binding process and are readily interpreted.

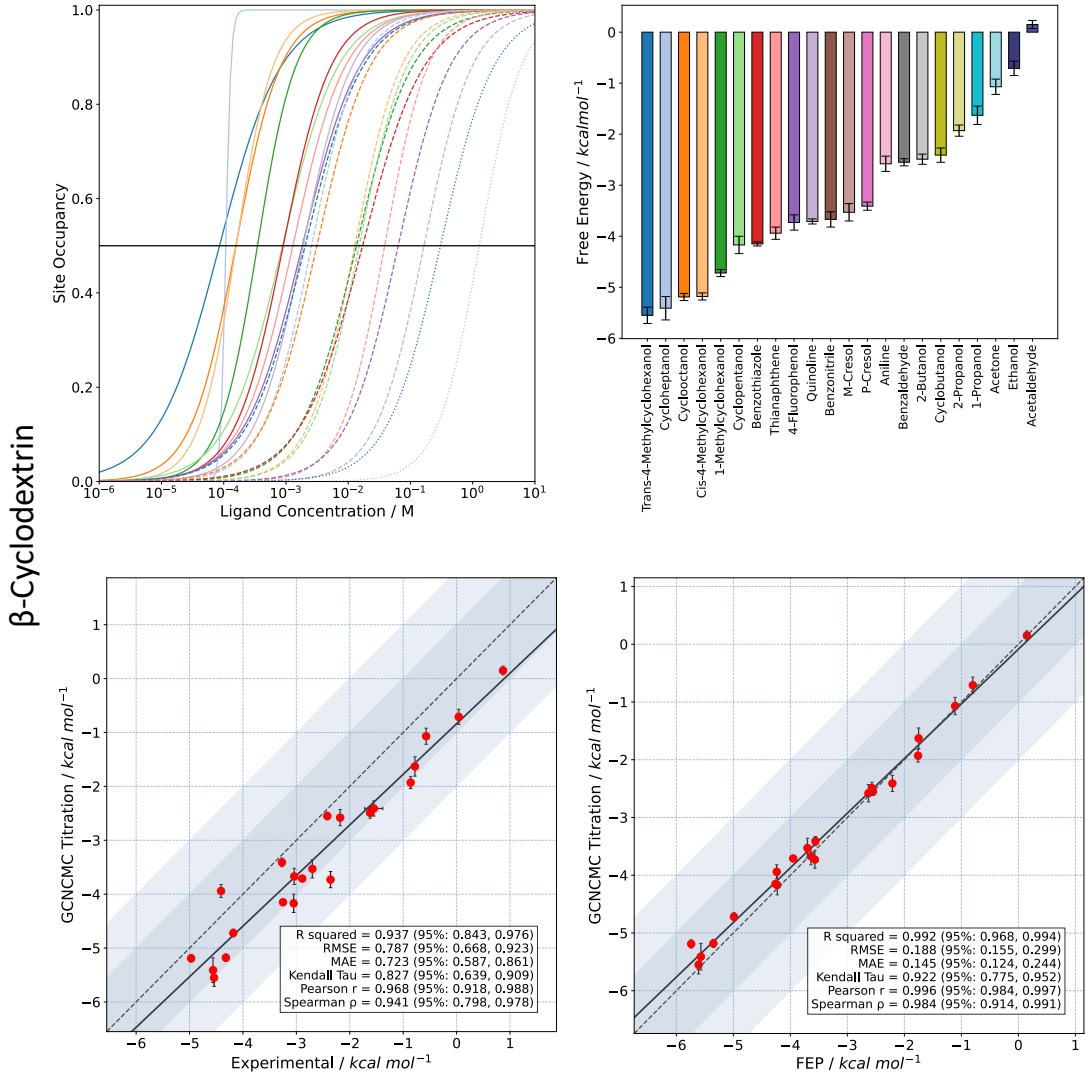


FIGURE 5.9: Binding free energy data for the 22 tested fragment molecules binding to βCD. Top: Titration curves (left) and binding free energy (right), the latter derived from the mean K_D from four simulation repeats, each fitted to a sigmoid curve, and is reported in units of kcal mol^{-1}. The error is the standard error of the mean of the four K_D values obtained from these fits. Bottom: Calculated absolute binding free energies from titration calculations vs. experiment and FEP results, the latter obtained using a flat bottom restraint. The error on the ABFE results are the standard error of the mean of 4 individual repeats (2 starting from each binding mode). Raw data can be found in the Appendix A.1.

Figure 5.9 also depicts the host-guest binding free energies extracted from the GCNCMC titrations, compared to experimental data and a basic ABFE approach which uses a flat bottom restraint to keep the guest bound as it is decoupled, as described in the simulation details. In general, a slight overestimation of the binding affinities relative to experiment was observed, prompting speculation that the forcefield parameters used may not be optimal, with similar trends having been reported previously.[229] Despite this, the calculations give a mean absolute error (MAE) and root mean squared error (RMSE) with respect to experiment of 0.7 and 0.6

kcal mol^{-1} with almost all the data points falling within 1 kcal mol^{-1} of the experimental value. Furthermore, the correlation ($R^2 = 0.94$) and ranking ($\tau = 0.83$) with respect to experiment shows that the method can reliably and accurately rank fragments in terms of their binding affinities. An almost perfect correlation with the more well-established FEP approach gives promising validation.

The nonequilibrium works collected during the GCNCMC titrations were also used to evaluate binding affinities using the GAR equation (Eq. 5.28). These results are shown in Figure 5.10, and in line with the solvation free energies, the results appear almost identical to the titration results prompting speculation as to whether there is any advantage to running titration calculations rather than collecting work data and using the acceptance ratio equations to calculate free energy after the simulation. This will need to be investigated in the future.

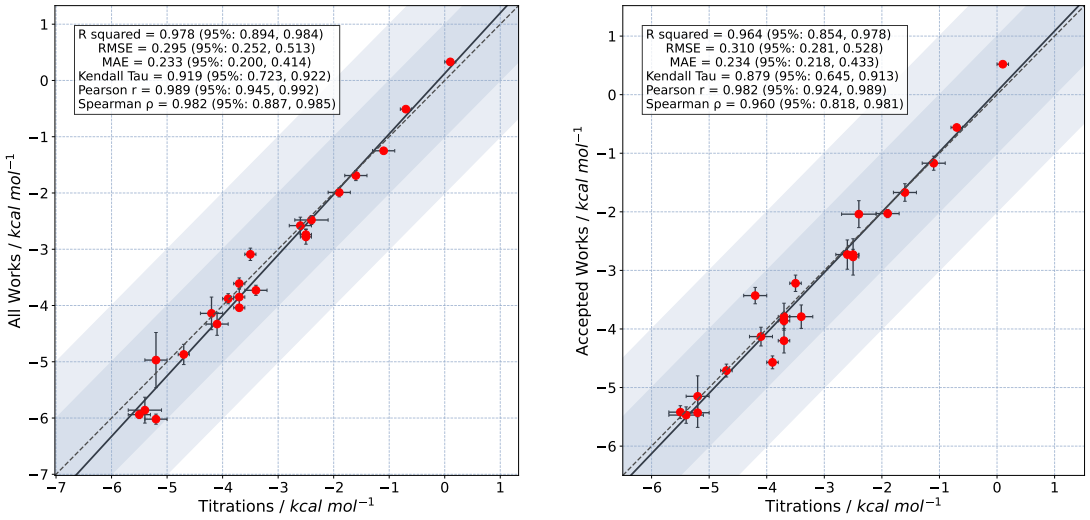


FIGURE 5.10: Host-Guest binding affinities calculated using the "GCMC Acceptance Criteria" (GAR: Eq. 5.28), versus GCNCMC titrations. Left: Using all the collected work values. Right: Using only work values for accepted moves.

5.4.2.1 Host Guest Binding Modes

As mentioned, within the GCNCMC framework, all possible binding modes should be sampled and the acceptance criteria will filter any binding modes which are not statistically likely at a particular concentration.

The β-cyclodextrin system was introduced in a previous chapter (Chap. 4) showing that simple guests with a single polar group, such as the ones titrated here, generally bind in two possible orientations (Fig. 4.1). To assess this, we used frames from the above titration studies for two representative fragments, benzonitrile and para-cresol (Fig. 5.11). Specifically, we look at simulations with B values that give approximately a 50% occupancy (B_{50}), as these are the B values corresponding to the dissociation

constant, K_D. We overlay these frames in Figure 5.11 and see that the polar group for both ligands, as expected, favourably points out the wider secondary alcohol opening of βCD. In the next chapter we perform a more quantitative study of binding modes in a protein-ligand system.

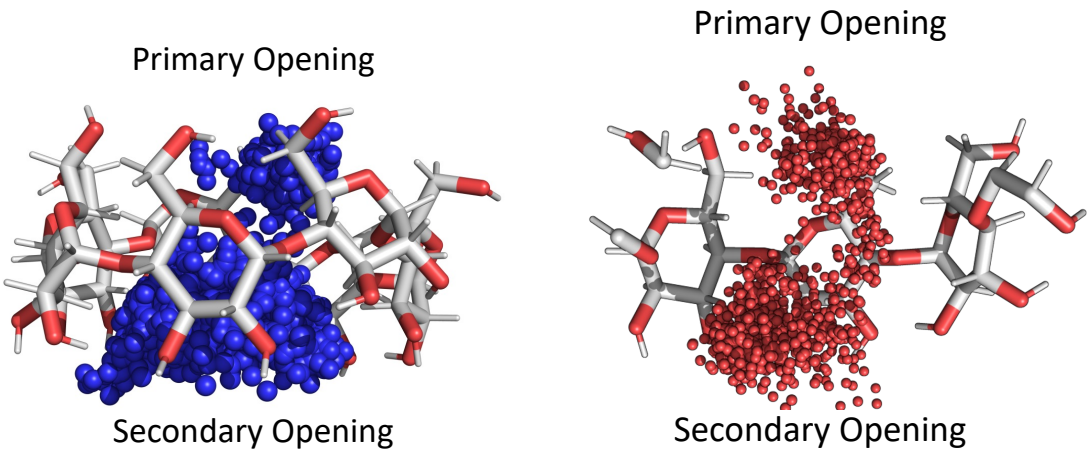


FIGURE 5.11: Overlaid frames from GCNCMC simulations of benzonitrile (left) and p-cresol (right) binding to β-cyclodextrin. GCNCMC simulations show a preference for the polar group of the guest (blue and red spheres) to point out the wider secondary opening. Note that the depiction of the host is that of the first frame only.

5.5 Summary

In this chapter, two novel methods for calculating binding affinities using GCNCMC simulations are presented and successfully applied to calculate hydration free energies and binding free energies with the host-guest system, β-cyclodextrin. The results show strong agreement with both experimental data and traditional FEP approaches, demonstrating the accuracy of this new protocol.

We first present two formulations detailing how GCNCMC simulations, and the associated nonequilibrium work measurements, can be used to calculate binding affinity. The first method determines K_D by titrating over a range of concentrations and fitting the resulting data to a logistic curve (Eq. 5.22). The B value, or concentration, which returns an average occupancy of 0.5 can then be used to calculate binding affinity. The second, denoted GAR, uses the nonequilibrium works directly and is based on the fact that at equilibrium the acceptance of insertion moves should equal deletion moves. In this formulation, Equation 5.28 can be solved for ΔF_{trans} using numerical methods which can then be used in the GCMC thermodynamic cycle (Fig. 5.1) to calculate a binding affinity. Further study into the relationship between GAR and BAR is ongoing.

The results in this chapter show that each of the presented formulations can reproduce simple hydration free energy estimates in line with experimental data, previously published data, and crucially between different methods. Second, we apply the titration method to a slightly more complex, but still simple, β-cyclodextrin system. We find that again, free energy estimates derived from GCNCMC titrations and related methods agree well with experimental measurements and with the more established equilibrium FEP method using MBAR.

Unlike conventional methods, which require user-defined restraints, the GCNCMC titration approach in principle allows for all relevant binding modes to be sampled, thereby eliminating the need to perform multiple simulations for each binding mode. This is confirmed qualitatively by overlaying simulation frames and seeing a clear preference for the secondary binding pose in line with previous publications.[229]

These results provide a robust validation of the titration protocol and provide an alternative to traditional equilibrium-based methods. Furthermore, the derivations presented here alongside the empirical evidence provide a more general validation of the overall GCNCMC method. In the following chapter, we expand on the methods presented here and apply them to protein-ligand complexes.

Chapter 6

Enabling Structure Based Design: Applications of GCNCMC to Protein-Ligand Systems

Much of the work presented in this chapter has been published in the paper: Accelerating Fragment Based Drug Discovery using Grand Canonical Nonequilibrium Candidate Monte Carlo authored by WP (DOI: 10.26434/chemrxiv-2024-q9l5z). Some figures have been reused and cited where appropriate. Some text has been reused and adapted.

6.1 Introduction

Now that the formulations and use cases of GCNCMC have been outlined, this chapter will focus on two protein-ligand systems and demonstrate how GCNCMC can be used in an SBDD setting. The key to SBDD is determining where and how ligands bind to a target. In this retrospective study, we will go from apo structures to binding site identification and finally binding affinities, all using GCNCMC simulations.

6.1.1 T4 Lysozyme

T4L99A is an extensively studied test system with a wealth of experimental data and is commonly used in the development of enhanced sampling and free energy methods.[1,192,234,242–244] In T4, a single point mutation (L99A) artificially creates a small, hydrophobic, cavity that binds a range of very simple ligands. The T4L99A system, while relatively simple, also has some complexities that make it interesting for method development. First, the binding site is fully occluded from the solvent making pocket detection difficult in basic mixed solvent MD (MSMD) simulations owing to the

timescales required for ligand diffusion.[94,95,245] Larger ligands, such as p-xylene, induce a rotamer flip of the Valine-111 side chain and many studies have struggled to reproduce this movement in standard MD simulations without enhanced sampling.[112,132] Lastly, some ligands, such as toluene, bind to T4L99A in two or more binding modes (Fig. 6.2) which can, again, be problematic to sample in simulations owing to kinetic trapping in one binding mode, or with the use of orientational restraints in free energy calculations.[1,109,128]

6.1.2 Major Urinary Protein-1

Like T4L99A, Major Urinary Protein-1 (MUP1) is another protein system with an occluded binding pocket and has been used as a test system for relative binding free energy calculations.[123,240] In this study, we perform titration calculations for 14 structurally diverse small molecules binding to MUP1 (Fig. 6.12). Crystal structures of these molecules show the presence of 1-2 bound water molecules in the binding site in various configurations, depending on the ligand. Previous studies have deemed these water molecules to be irrelevant to the final free energy estimate and removed them before simulation.[123,240] However, in previous studies using GCNCMC, we have found that both the ligands and water molecules adapt their configurations in response to changes in the local environment, and we therefore retain the crystallographic waters.[150] The set of primary alcohols binding to MUP1 all form a hydrogen bond to Tyr120, either directly or via a bridging water, indicating the need to retain the crystal waters.

The binding modes of these alcohols fit into two categories with the smaller pentan-1-ol and hexan-1-ol binding in a similar orientation and heptan-1-ol, octan-1-ol, and nonan-1-ol binding in another, due to steric hindrance. Interestingly, pentan-1-ol also binds, with weaker electron density, in the second orientation.[246] We will assess the sampling of these different modes in our titration calculations.

6.2 Simulation Details

6.2.1 System Setups

The apo structure of T4L99A (PDB: 4w51) was protonated according to a pH of 7.0 using PDBFixer.[247] Missing loops were added where appropriate and protein termini were capped using N-methyl and acetyl caps. Each system was then solvated in a box of TIP3P[221] water with a buffer of 12 Å around the protein. NaCl ions were added to neutralize the system and up to a salt concentration of 0.15 M.

All simulations of MUP1 start from a protein-ligand complex with PDB code 1i06. The crystal ligand was removed but crystallographic waters were retained. The protein was protonated according to a pH of 7.0 using PDBFixer.[247] Missing loops were added where appropriate with PDBFixer and the protein termini were capped using N-methyl and acetyl caps. The system was then solvated in a box of TIP3P[221] water with a buffer of 12 Å around the protein. NaCl ions were added to neutralize the system and further added to a concentration of 0.15 M.

6.2.2 General Simulation Details

Unless otherwise stated, all simulations were performed using OpenMM[152] with the *grandlig* Python module to set up the custom forces and to perform GCNCMC simulations and free energy calculations. Simulations were performed at 298 K and all MD was performed using the Langevin BAOAB[168] integrator with a friction coefficient of 1 ps^{-1} and a time step of 2 fs or, for MUP1, 4 fs with hydrogen mass partitioning (Hydrogen mass = 2 Da). Where appropriate, a Monte Carlo barostat is used to maintain a system pressure of 1 bar. The cut-off for nonbonded interactions was 12 Å with a switching function applied at 10 Å for the Lennard-Jones interactions. Particle mesh Ewald[183] was used to calculate the effect of the long-range electrostatics. Owing to software limitations, the long-range dispersion correction is neglected, as per our previous work.[147] The proteins T4L99A and MUP1 were modelled using the AMBER ff14SB forcefield.[220] All simulations use TIP3P[221] waters and all ligands are parameterized using GAFF[248] with AM1-BCC charges.[223] Ions, wherever present, were modelled with Joung-Cheatham parameters.[249]

6.3 T4 Lysozyme

6.3.1 Binding Site Identification

Firstly, we must determine where our ligands bind to T4L99A. The binding site in T4L99A is fully occluded from the solvent, making sampling an issue in basic MD simulations. Some publications state that the binding of a benzene molecule can take tens of microseconds.[192,245] Clearly, this is unfeasible in a real world setting and therefore we aim to improve the sampling using GCNMC. To do so, we assume no knowledge concerning the position of the binding site and solvate our apo structure in a 0.5 M solution of benzene. We then use a GCMC sphere encompassing the entire protein to allow potential insertion moves in and around the protein. We compare the results to a basic MSMD simulation of an equivalent simulation time.

6.3.1.1 Specific Simulation Details

In enhanced MSMD simulations looking for the benzene binding site, the protein was solvated in a 0.5 M benzene and water solution. The GCMC sphere was centred on the middle of the protein at the midpoint between the CA atoms of Phe104 and Glu11 with a radius of 26.5 Å to cover the whole protein. The infinitely dilute excess chemical potential of benzene was taken to be -0.68 kcal mol^{-1} calculated as described previously (Sec. 4.3.1). The average volume per ligand was taken as 3321 $Å^3$ to define a concentration of 0.5 M. Using a switching time of 50 ps, we ran six repeats of 700 GCNCMC/MD cycles (1 move per 50 ps of MD) with the first 200 cycles discarded as equilibration, giving a maximal simulation time of 50 ns (25 ns of MD with 25 ns of switches). For a fair comparison, GCNCMC simulations were compared to 50 ns simulations of conventional NPT MD on the same system.

6.3.1.2 Results

As with ERK2 in Chapter 3, FTMap[88] was used to try and predict the small aromatic binding site deeply buried under the surface of T4L99A. Although buried and inaccessible to solvent, the site is rigid and preformed and therefore should be a simple test for FTMap. Surprisingly, the binding site was not detected by any FTMap probe, even benzene. Given this result, it is clear we must use a more elaborate method such as MSMD and GCNCMC-enhanced MSMD.

Across six repeats of GCNCMC-based MSMD simulations, using 0.5 M probe concentration, the benzene binding site was readily found within an average of 34 GCNCMC moves. For context, 700 moves per repeat were run in 24 hours of wall time on a GTX1080 GPU. Once the site was found, the ligand remained for the rest of the simulation as every deletion proposal was deemed unfavourable at this concentration. By binning the coordinates of sampled benzene heavy atoms onto a grid with 0.5 Å resolution, we can count the number of frames a benzene atom was present at each grid point and then average based on the total number of frames. This grid-based analysis is described in depth in Section 7.2.1. Contouring this grid to represent an occupancy of at least 90% of our frames (Fig. 6.1), shows a clear signal around the crystallographic binding pose indicating that a benzene molecule was present in this site for at least 90% of our simulation. As no other grid points are occupied at such high percentages, the binding site is clear with a lack of false positives.

These results are compared to a basic MSMD simulation of T4L99A in a 0.5 M solution of benzene and are also shown in Figure 6.1. The MD grid is contoured at 30% indicating the grid points where a benzene atom had resided for 30% or more of the simulation frames, showing that the benzene binding site was not sampled at all. Turning this contour level all the way down to 0% still shows no binding in the buried

binding site highlighting the difficulty MSMD faces when binding into occluded pockets. Again, this is unsurprising as studies have shown that benzene binding to this cavity is only observed in tens of microseconds long MD simulations.[192]

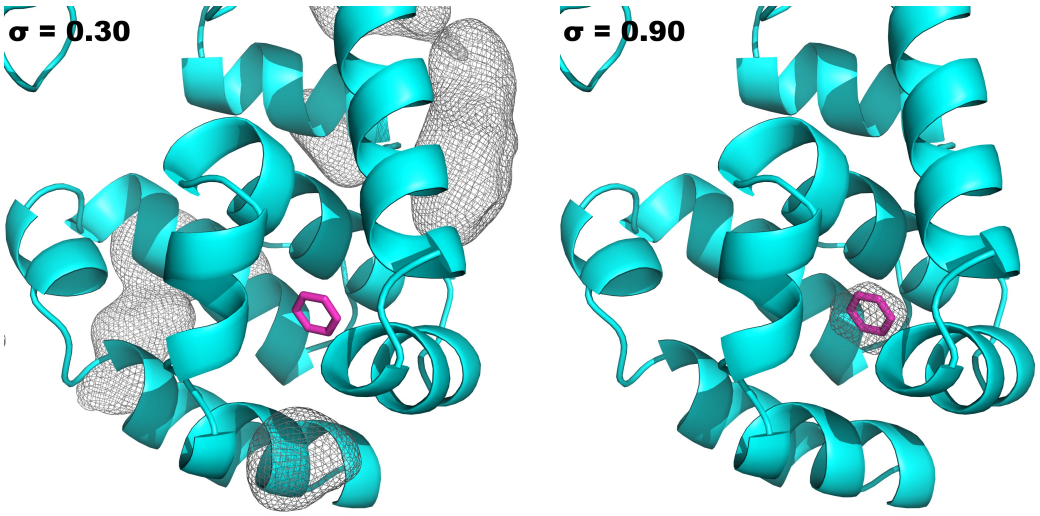


FIGURE 6.1: Occupancy grids of MD (left) and GCNCMC simulations (right) contoured at a value of 0.30 and 0.90 respectively. Grids represent a minimum of 30% and 90% of the frames for which a benzene atom visited a given grid point. The benzene crystal pose is shown in magenta (PDB: 181l).

These results are in line with previously published data.[94,95,245] The most similar approach, SILCS,[94] uses a GCMC-like sampling scheme which oscillates the chemical potential of the system to drive insertions and deletions of molecules into a system with cavity bias. The GCMC moves are alternated with regular MD to propagate the system similarly to the present study. However, as the chemical potential is allowed to fluctuate the simulations do not fully obey the grand canonical ensemble and the acceptance rates of these GCMC moves are typically low. Clark *et al.*[144] also used GCMC to study the binding of small hydrophobic molecules to T4L99A. However, this study was performed in the absence of solvent molecules and did not incorporate any MD. Ligand-mapping MD,[95] another MSMD protocol, used long accelerated MD simulations with a low concentration of benzene probes to identify the buried binding pocket. However, it is unclear if these simulations were reweighted to remove the bias of the accelerated potential. Finally, another accelerated MD-based approach, Gaussian accelerated MD, studied the binding of various small molecules to different mutants of T4 to estimate their kinetic properties and thus free energies with high accuracy.[245]

6.3.2 T4L99A-Toluene Binding Modes

While looking qualitatively at the binding modes in β-Cyclodextrin gave reassuring evidence that GCNCMC naturally samples multiple binding modes, we wanted to perform a more quantitative test to ensure this is the case. As already mentioned, ligands, particularly fragments, can bind in more than one orientation, and usually, x-ray crystallography shows only the most stable. To calculate accurate binding affinities, all binding modes must be sampled, but in traditional ABFE calculations, transitions between modes are rarely seen owing to either the use of orientational restraints or kinetic trapping. As such, FEP protocols must simulate each binding mode in turn, increasing the number of calculations that need to be performed as in the study by Mobley *et al.*[128]

Toluene binding to T4L99A provides a simple test case with previously published simulated data from Gill *et al.*[109] In that study, two toluene binding poses were identified: the crystal pose (denoted A1/A2) and a secondary pose (B1/B2). Note, A1/A2 corresponds to the symmetrically equivalent poses of A and likewise for B, owing to the C_2 symmetry axis of toluene. The authors also identify a dihedral angle between the C-alpha atom of Arg119 and 3 toluene atoms, which, when measured, can be used to easily distinguish the four binding modes. In that same study, free energy calculations revealed a free energy difference of 0.6 k_BT, which translates to a population ratio of 65:35 at 300 K for poses A and B. This is in line with our own free energy calculations, described below, which revealed a free energy difference of 0.73 k_BT at 298 K giving a population ratio of 68:32.

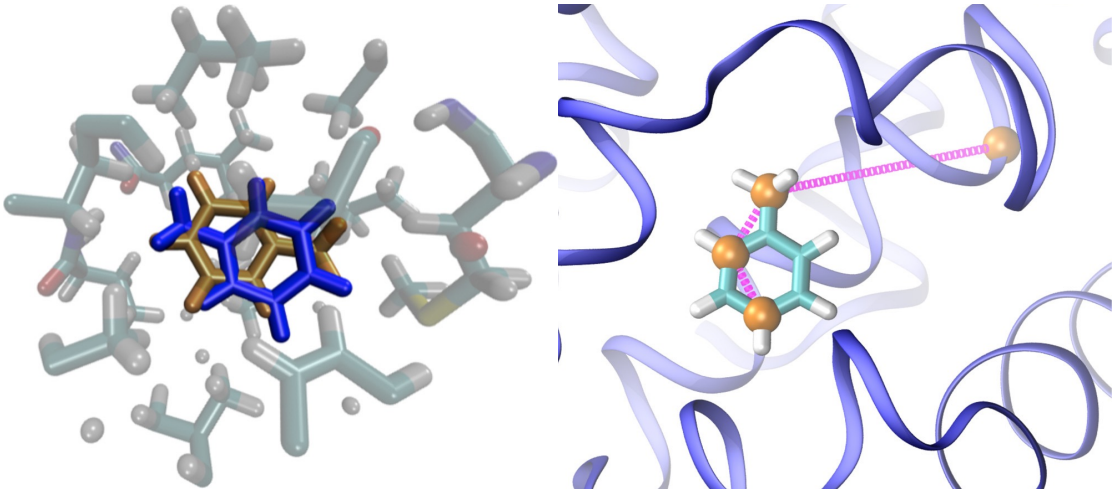


FIGURE 6.2: Left: The two binding modes of toluene to T4L99A. The toluene in orange corresponds to the observed pose in the crystal structure while the blue toluene is a secondary pose. Right: The dihedral angle between 3 toluene atoms at the Cα atom of Arg119 to distinguish between the binding modes. Adapted from Gill *et al.*[109]

6.3.2.1 Specific Simulation Details

In GCNCMC simulations, using the Adams value which has an average occupancy of 50%, B_{50}, results in the maximal number of binding and unbinding events to maintain that 50% occupancy. As a result, simulations performed at this value will give the most optimized sampling of the multiple binding modes. This value also corresponds to K_D and is thus directly comparable to the BLUES results of Gill. In other words, insertion and deletion moves are accepted or rejected at a level appropriate for this concentration. The value of B_{50} was taken to be -7.34 as determined by titration calculations in the following section. The GCMC sphere was centred on the binding site at the midpoint between the CA atoms of Leu85 and Ala100 with a radius of 8 Å. The dihedral angle, identified by Gill *et al.*[109] to distinguish between the bound configurations, between the CA of Arg119 and three toluene atoms was measured at each frame and binned onto a histogram. Dihedral angles between $-\pi$ to -1.5 and 0 to 1.5 were assigned to binding modes the crystal poses A1 and A2 respectively. The secondary poses, B1/B2 were assigned to angles between -1.5 to 0 and 1.5 to π.

6.3.2.2 Results

The populations obtained from GCNCMC simulations are shown in Figure 6.3. We observed a ratio of 67:33 between poses A1/A2 and B1/B2, which is in remarkably good agreement with our free energy estimates (69:31) and that of Gill *et al.* Reassuringly, we also observe population ratios of 33:34 and 17:15 between symmetry-equivalent poses A1/A2 and B1/B2 respectively, indicating thorough sampling. An interesting distinction between these results and those published using BLUES is that when using GCNCMC, we did not have to define a transformation between the poses (taken as a centre of mass rotation by Gill et al.[109]). Further, with GCNCMC, prior knowledge of multiple binding modes is not required as we have shown that these are intrinsically sampled without additional bias. That said, an interesting approach for the future would be to combine both methods such that the sampling of binding modes could be enhanced during the MD portions of the GCNCMC/MD protocol.

In an alternative analysis, using CLonE,[250] we have clustered the pairwise RMSD of the ligand positions throughout the simulation. The data are then projected onto a latent space using PCA. The populations of the four binding modes are in good agreement with the populations obtained using the histograms. This analysis is more general as knowledge of a dihedral angle that discriminates the binding mode is not required. These results provide further validation that all binding modes and symmetrically equivalent modes are sampled within a GCNCMC simulation and as such are naturally included in titration calculations.

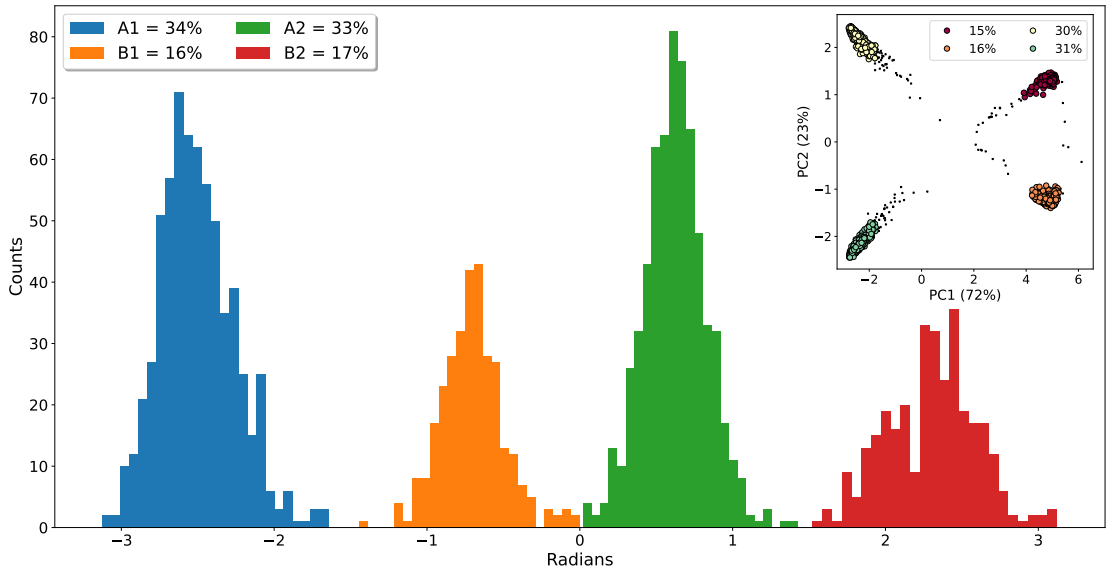


FIGURE 6.3: Distribution of toluene binding modes observed in GCNCMC/MD simulations. Dihedral angles between $-\pi$ to -1.5 and 0 to 1.5 were assigned to binding modes A1 and A2 respectively. B1/B2 were assigned to angles between -1.5 to 0 and 1.5 to π. Inset: Pairwise RMSD between ligand poses projected onto PCA space and coloured by the four clustered binding modes.

It is worth highlighting that to get these results, we first needed to calculate a value for B_{50} from titration calculations. This may generally be unfeasible as it requires a full titration, or absolute FEP, calculation prior. However, the goal of this part of the study is not to identify multiple binding modes, but rather prove that they are accurately sampled in GCNCMC simulations without bias. Consequently, this has implications for our titration calculations, where binding modes are inherently sampled and are thus accounted for in the final free energy estimates, meaning there is no requirement for prior knowledge of the binding modes nor separate calculations for each mode. Further, as symmetrically equivalent poses are explicitly sampled, there is also no need for symmetry corrections.

6.3.3 Titrations

6.3.3.1 Specific Simulation Details

Titrations were performed over 20 B values loosely centred around the experimental binding free energy, although knowledge of the experimental binding affinity is not necessary and the titration could be performed over any B range. B values were calculated using Equation 5.15 with a fixed μ'_{sol} value for each ligand, calculated using a basic hydration free energy calculation as described prior (Sec. 4.3.1). Insertions and deletions were performed with a switching time of 150 ps. In each cycle, a GCNCMC move was attempted for every 1 ps of MD. 1700 cycles were performed, with the first

200 discarded as equilibration, giving a maximal production time of 76.5 ns per B value. As before, the GCMC sphere was centred on the binding site at the midpoint between the CA atoms of Leu85 and Ala100 with a radius of 8 Å. The ligands used for this study are shown in Figure 6.4.

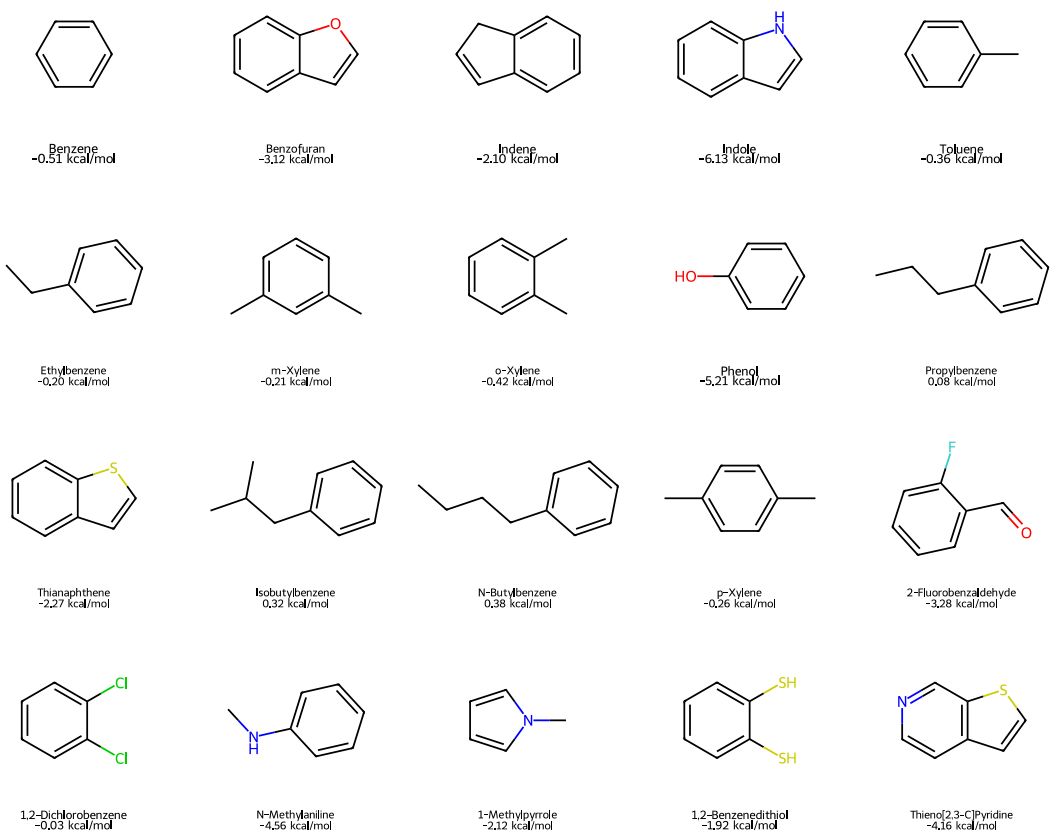


FIGURE 6.4: T4L99A ligands and their calculated values of excess chemical potential, μ'_{sol}.

Equilibrium FEP calculations were performed to compare the titration results. We selected the most stable poses from our GCNCMC simulations as the starting coordinates for the FEP calculations. For ligands where significant multiple binding modes ($\geq$10% population) were identified in titrations, by CLoNE clustering, a full FEP calculation was performed for each mode. This highlights how the number of simulations can quickly get out of control for fragment molecules that bind in multiple orientations. For each mode, Boresch restraints[127] were used to keep the ligand bound using a force constant of 10 kcal mol^{-1} Å^{-2} for distance restraints and 10 kcal mol^{-1} rad^{-2} for angle and dihedral restraints. The individual restraint atoms were chosen automatically using MDRestraints Generator.[126] To calculate the restraint contribution to the free energy, the restraints were slowly applied to the fully coupled ligand over 15 lambda values. The analytical standard state correction was calculated using:

$$\Delta G^{\ominus}_{\text{restr, off}} = -k_B T \ln \left[\frac{8\pi^2 V^{\ominus}}{r_0^2 \sin\theta_{\text{A},0} \sin\theta_{\text{B},0}} \frac{\left(K_r K_{\theta_A} K_{\theta_B} K_{\phi_A} K_{\phi_B} K_{\phi_C}\right)^{1/2}}{(2\pi kT)^3} \right] \tag{6.1}$$

where r_0, θ_A, and θ_B are the reference values for the distance and two angle restraints. K_X are the force constants for X restraint.[127]

For the bound leg, using the same restraints, the ligands were decoupled over 40 lambdas recording a potential energy sample every 3 ps for a total of 1000 sampled. The free energy of the bound leg was then calculated using MBAR. The final free energy estimate was calculated using the thermodynamic cycle shown in Figure 2.9. The final free energy estimate for ligands with multiple binding modes was calculated using a Boltzmann average of all the calculated modes.[128] Finally, symmetry corrections were applied where appropriate.[251]

6.3.3.2 Results

Once a binding site is known, we can shrink the GCMC sphere to cover only the binding site so that we can focus our insertion and deletion moves on just this region. GCNCMC simulations were performed over a range of B values and the resulting average occupancy was used to plot the titration curves shown in Figure 6.5.

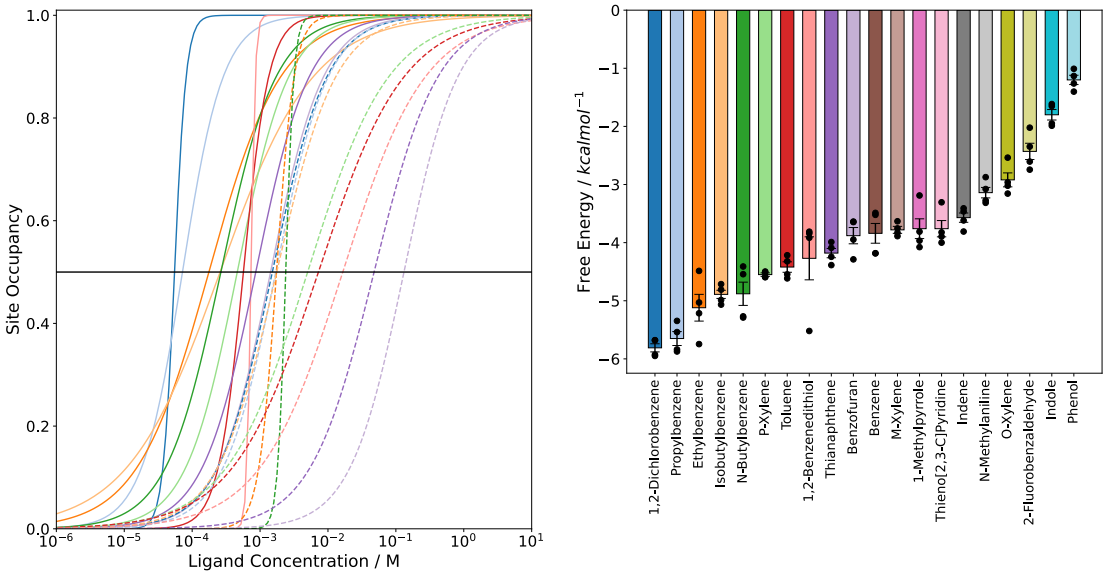


FIGURE 6.5: Titration curves for ligands binding to T4L99A. Values given in the legend are the final calculated free energy, derived from the ligand concentration which gives 50% bound occupancy (K_d), and Kendall Tau values detailing the quality of the fit. Reported errors represent one standard deviation. Raw data can be found in the Appendix A.2.

Figure 6.6 shows the calculated affinities versus experimental data and ABFE calculations. For T4L99A, the correlation with experiment ($R^2 = 0.562$) is comparable

to other methods[1] including to our FEP protocol ($R^2 = 0.522$). However, the average error and RMSE with respect to experiment are particularly high, highlighting the added complexity of a protein system compared to a simple host-guest test case. Phenol, which is thought to predominantly bind to the unfolded state of T4L99A, was included as a negative control with GCNCMC predicting a slightly negative binding free energy ($K_D = 0.21M$). It is possible that this weak binding is masked in the thermal shift assay by preferential binding to the unfolded protein.[252] Crucially, the titration results are in good agreement with those obtained using FEP which implies that the methods are consistent. Indene stands out as a significant outlier with a $\Delta G_{exp} = -5.13$ kcal mol^{-1}. However, the reason remains unclear with this finding being reported in other publications.[1]

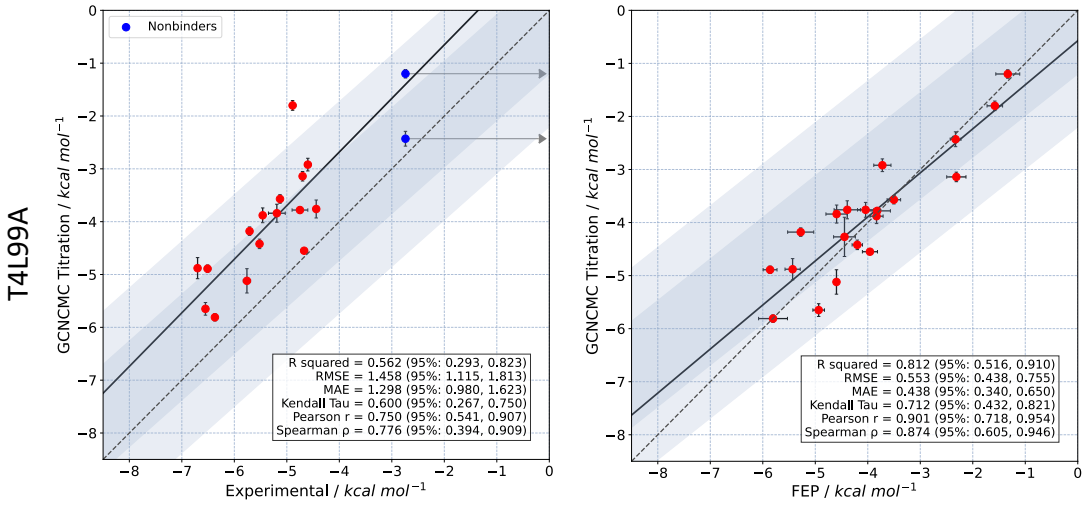


FIGURE 6.6: Calculated binding free energies for T4L99A from titration calculations vs. experiment and absolute FEP results, using Boresch restraints. Titration free energies are derived from the mean K_D values of four simulation repeats, each fitted to a sigmoid curve, and are reported in units of kcal mol^{-1}. The error is the standard error of the mean of the four K_D obtained from these fits. ABFE calculations used appropriately weighted binding free energies derived from independent simulations of all populated binding geometries with a greater than 10% observed occupancy in GCNCMC titrations. The error on the ABFE results are the standard error of the mean of 3 individual repeats. For the comparison with experimental ligand binding free energies, data are only presented for compounds with experimental ITC data.[1] Phenol and 2-Fluorobenzaldehyde are shown in blue and have a minimum experimental binding free energy of -2.74 kcal mol^{-1}. These compounds are not included in the reported statistics and line of best fit data.

6.3.3.3 Binding Modes in Titration Calculations

Although we have shown above that GCNCMC simulations sample the different binding modes of toluene, those simulations were performed at a fixed B value selected to maximise the number of accepted insertion and deletion moves thus

maximising the sampling of multiple binding modes. Here, we want to validate that the binding modes are also sampled in our titration calculations.

To ensure, and further prove, that binding modes are naturally sampled throughout the GCNCMC titration protocol we perform the same clustering analysis described for toluene. We take ligand-bound frames from simulations at B values which give an average occupancy of between 0.4 and 0.6. This is because low B values, where the average occupancy is low, will only sample the most favourable binding modes, while at high B values, all binding modes are sampled indiscriminately. Figure 6.7 shows the binding modes for benzene and toluene along with the occupancy of the clusters. In the most simple case, benzene, which has 12 energetically equivalent binding poses, we see 12 clusters all with similar populations further highlighting that symmetrically equivalent modes are well sampled. For toluene, we again see similar populations to that of the histogram analysis in Figure 6.3. Note that in this analysis we are using simulation frames from B values ranging between 0.4 and 0.6 rather than exactly 0.5, this is likely the source of the noisy 5th cluster (7 %) that is observed.

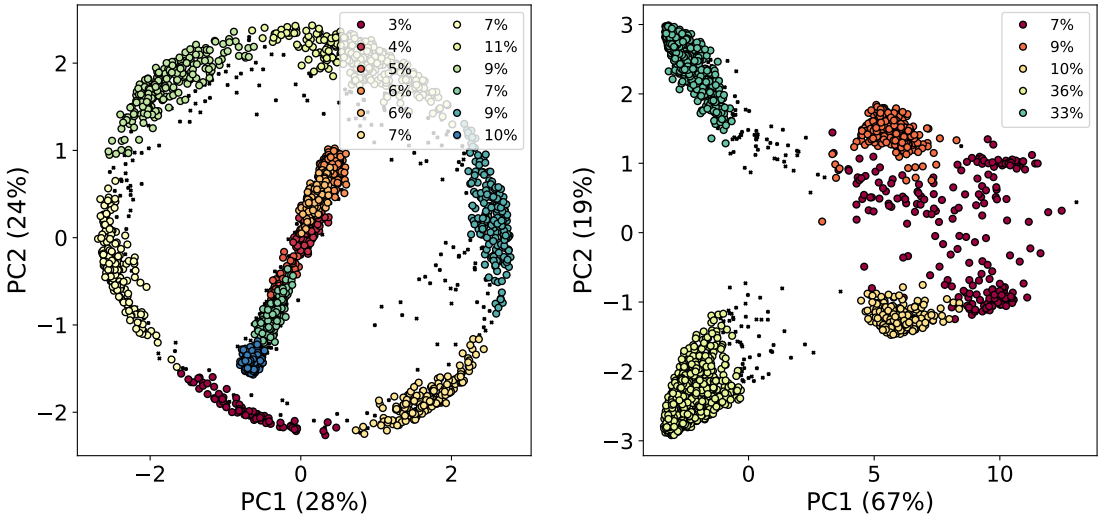


FIGURE 6.7: PCA analysis of binding modes of benzene (left) and toluene (right) to T4L99A from titration calculations. The plot is colored by cluster, which was assigned using CLoNe.

Indole and indene are two other ligands that are known to have multiple binding modes within 1 kT of each other and contribute significantly to the overall binding free energy.[1] By clustering the ligand poses based on the pairwise RMSD, we find that for indole we sample the crystal pose for 62% of our bound frames, a secondary pose for 25% of our frames, and two other minor binding modes. Finally, indene is a perfect example of a ligand that binds in four unique ways to T4L99A. From the populations, it is clear that all four modes are close in energy to each other and are likely to play a significant role in the overall free energy of the ligand binding. Unless explicitly sampled in traditional free energy calculations, this information is lost. Interestingly,

in this case, the most populated pose, 32%, does not correspond to the binding mode seen in the crystal structure, instead cluster 2 with 20% population is the experimentally reported binding mode.

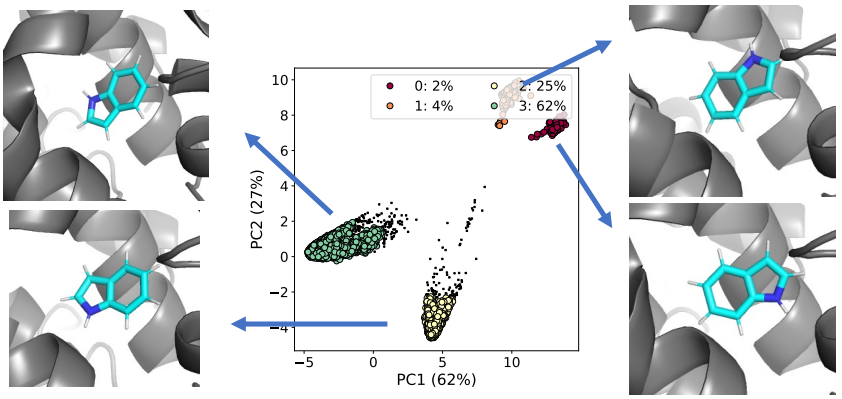


FIGURE 6.8: Four binding modes of indole sampled within GCNCMC simulations

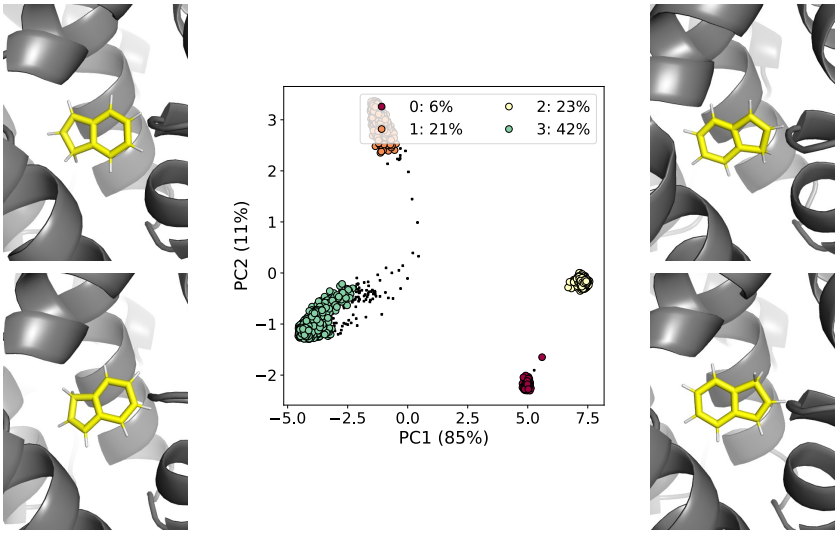


FIGURE 6.9: Four binding modes of indene sampled within GCNCMC simulations

The first point to note is that as distinct binding modes and symmetrically equivalent binding modes are sampled naturally, there is no need to perform more than one set of GCNCMC calculations or to apply a symmetry correction. Secondly, this method of calculating free energies does not require the use of artificial restraints which again, can become problematic and often require a degree of user input.

6.3.3.4 Free Energies using GAR

As with the host guest system, we use the works measured in titration simulations to calculate the binding affinities using GAR and compare them to the titration calculations. Figure 6.10 shows these results and, as before using all the collected work

measurements results in a strong correlation to the titrations. Interestingly, using only the accepted works actually results in a slight offset. A potential reason for this observation is that these work values have already been filtered once by the acceptance criteria according to different concentrations, and as such the input to GAR may become biased. In other words, the resulting work distribution are no longer independent of concentration and crucial data may be lost, particularly at the tails of the distribution.

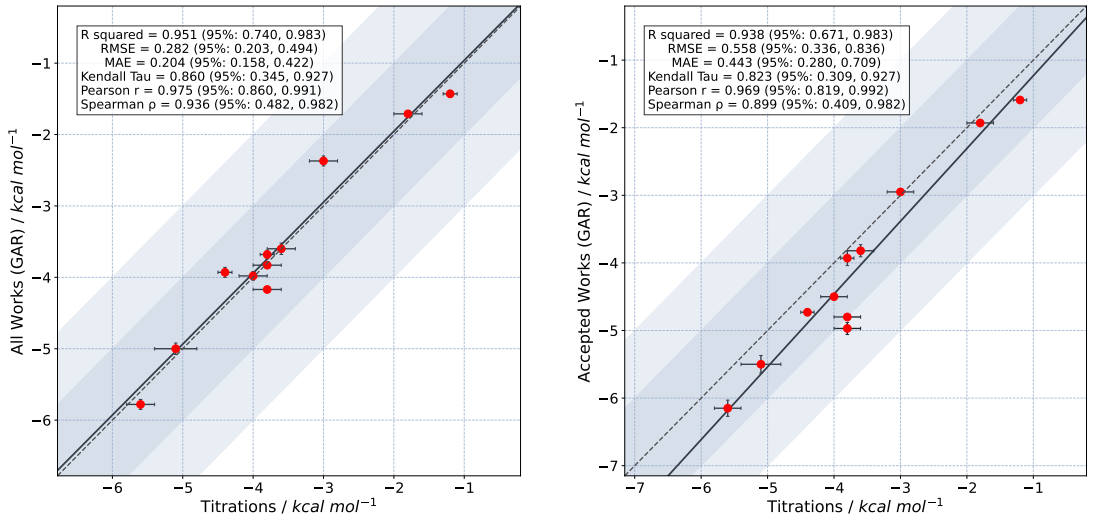


FIGURE 6.10: T4L99A binding affinities calculated using the "GCMC Acceptance Criteria" versus GCNCMC titrations. Left: Using all the collected work values. Right: Using only work values for accepted moves.

6.4 Major Urinary Protein-1

6.4.1 Binding Site Identification

6.4.1.1 Specific Simulation Details

To find the occluded binding site of MUP1, GCNCMC simulations (5 repeats) with systems containing 0.5 M of ligands **07**, **08**, and **14** were run as these are the smallest ligands in their series (Fig. 6.12). Simulations were run for 25 ns with 500 GCNCMC moves interspersed every 40 ps for a maximum simulation time of 50 ns. The GCMC region was designed to cover the whole protein anchored to the CA atom of Gly136 with a radius of 22 Å. The simulations were compared to 50 ns conventional MD simulations of the same systems. Insertions and deletions for these simulations were performed with a switching time of 50 ps.

6.4.1.2 Results

Like T4L99A, we would like to know where our set of ligands bind. In Figure 6.11 the MD and GCNCMC occupancy grids are presented. Immediately, for all three ligands, it is clear that basic MD fails to sample the binding pocket, and even setting the contour level of the MD occupancy grids to 1% still shows no binding. As with T4L99A, this lack of binding can be attributed to two factors: firstly, the binding site is completely occluded from the bulk solvent and, secondly, at 0.5 M, all three fragments aggregated, severely impacting the level of sampling that is achievable in the simulation.

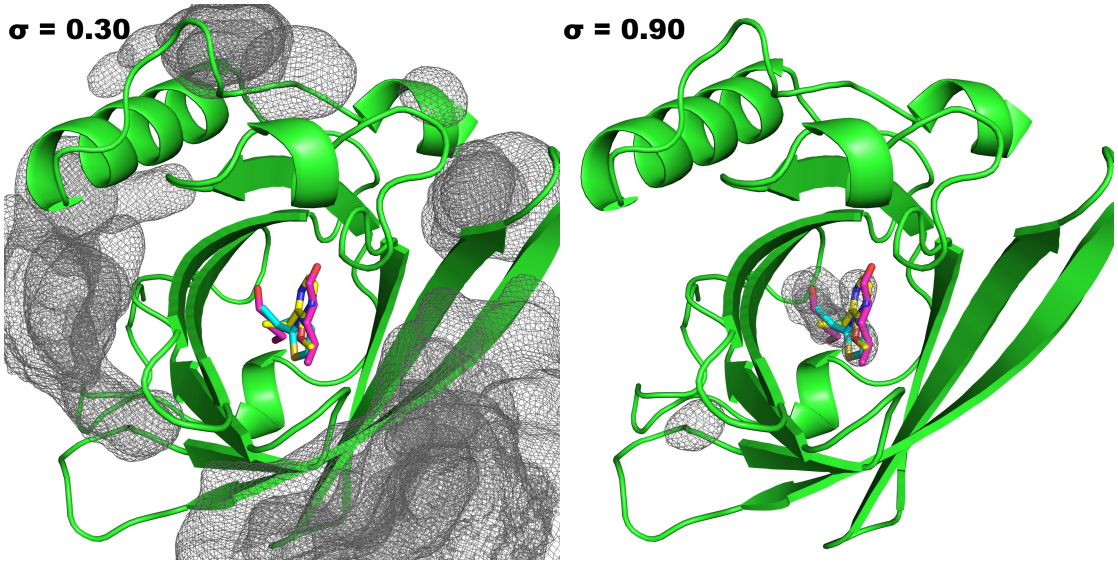


FIGURE 6.11: Occupancy grids of MD (left) and GCNCMC simulations (right) of MUP1 contoured at a value of 0.30 and 0.90 respectively. Grids represent a minimum of 30% and 90% of the frames for which a grid point was occupied by a ligand. The grids for all three MUP1 ligands (**07**, **08**, and **14**) are shown together. Representative crystal structures for each ligand are shown in cyan (PDB: 1i06), magenta (1znd), and yellow (1qy2).

6.4.2 Titrations

6.4.2.1 Specific Simulation Details

Titration calculations were performed for 14 structurally diverse small molecules binding to MUP1 and are shown in Fig. 6.12. Titrations were performed over 17 B values between -25 and -12, loosely corresponding to a concentration range of nanomolar to millimolar. In each cycle, a GCNCMC move was attempted for every 1 ps of MD. The GCMC sphere for titration calculations was defined between Phe74 and Leu123 with a radius of 5.5 Å to cover the binding site. In the simulations of MUP1, to avoid wasting computational time at the high and low concentrations, where moves

are rarely accepted, the simulations were terminated after 200 consecutive rejected moves. Equilibrium absolute FEP calculations were performed identically to T4L99A.

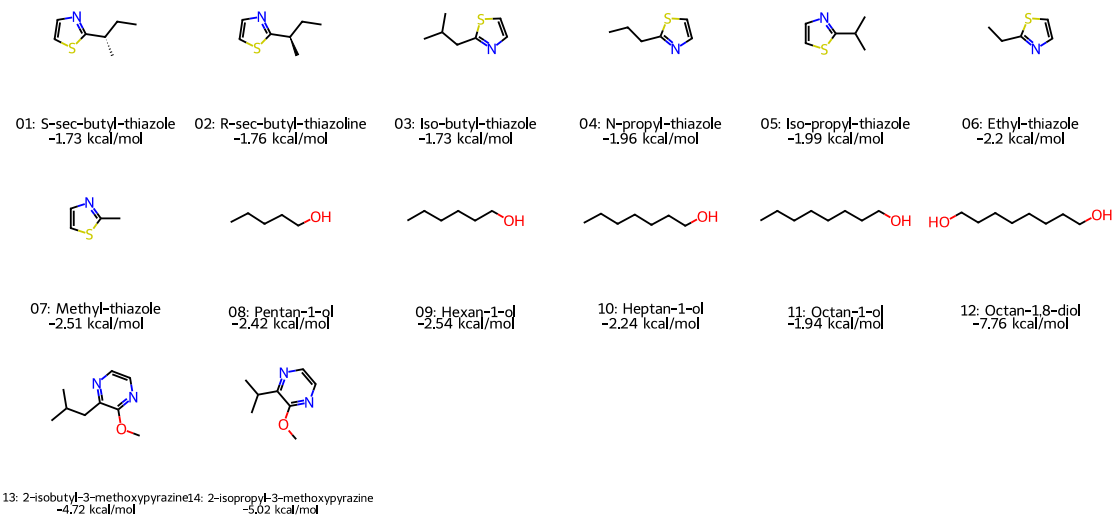


FIGURE 6.12: MUP1 ligands and their calculated values of excess chemical potential, μ'_{sol}.

6.4.2.2 Results

MUP1 titration curves are shown in Figure 6.13 with the derived free energies plotted versus experimental and FEP in Figure 6.14. Like T4L99A, the MUP1 titrations give a good correlation to both experimental data and FEP calculations.

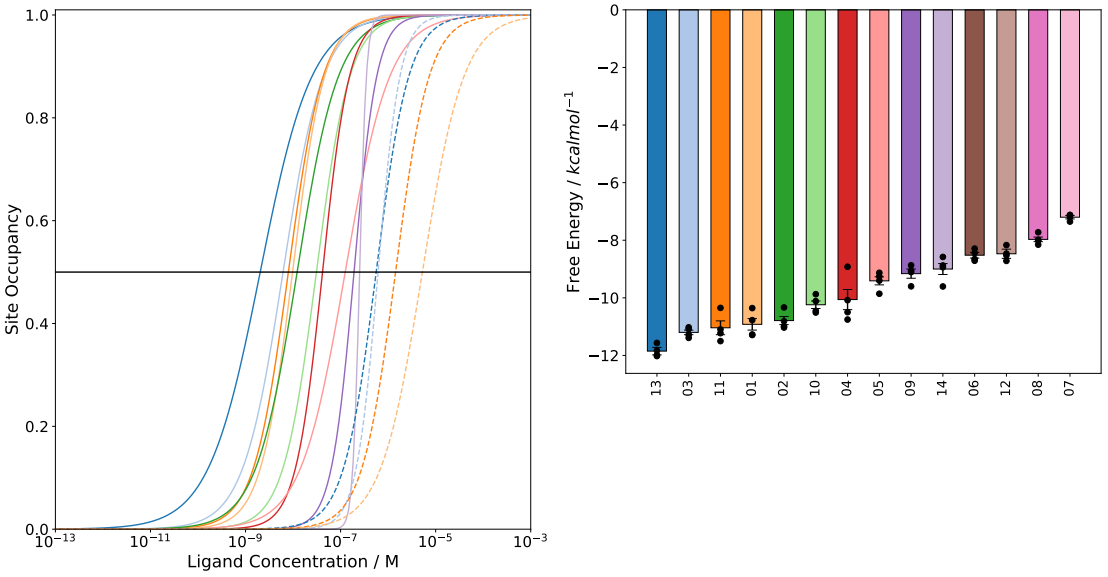


FIGURE 6.13: Titration curves for ligands binding to MUP1. Values given in the legend are the final calculated free energy, derived from the ligand concentration which gives 50% bound occupancy (K_D), and Kendall Tau values detailing the quality of the fit. Reported errors represent one standard deviation. Raw data can be found in the Appendix A.3.

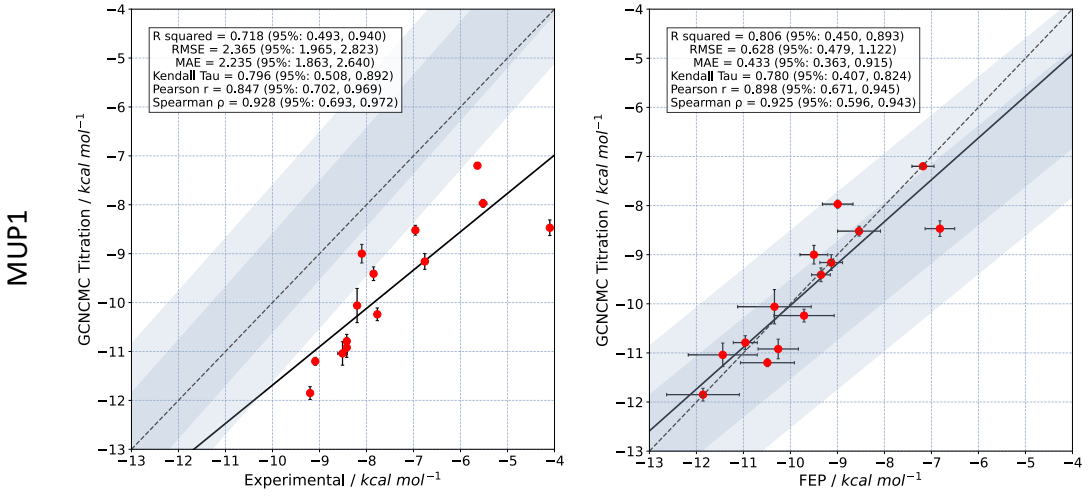


FIGURE 6.14: Calculated binding free energies for MUP1 from titration calculations vs. experiment and FEP results, using Boresch restraints. ABFE calculations were performed for the most populated binding modes from the GCNCMC titrations, which were determined by clustering.

6.4.2.3 MUP1 Binding Modes

As mentioned, some MUP1 ligands have multiple potential binding modes. Using the same clustering method as in T4L99A, the extent to which these modes are sampled is investigated. Figure 6.15 shows the binding modes for a ligand from each of the three

sub-series studied (Fig 6.12) alongside ClonE clustered PCA analysis based on pairwise RMSD of the poses. In each case, the most populated cluster overlaps well with the crystal pose with other minor poses seeming plausible (visual inspection). In particular, the two predicted binding modes of pentanol align well with the two modes in the crystal structure and are also consistant with longer-chain alcohols (1zne and 1zng).

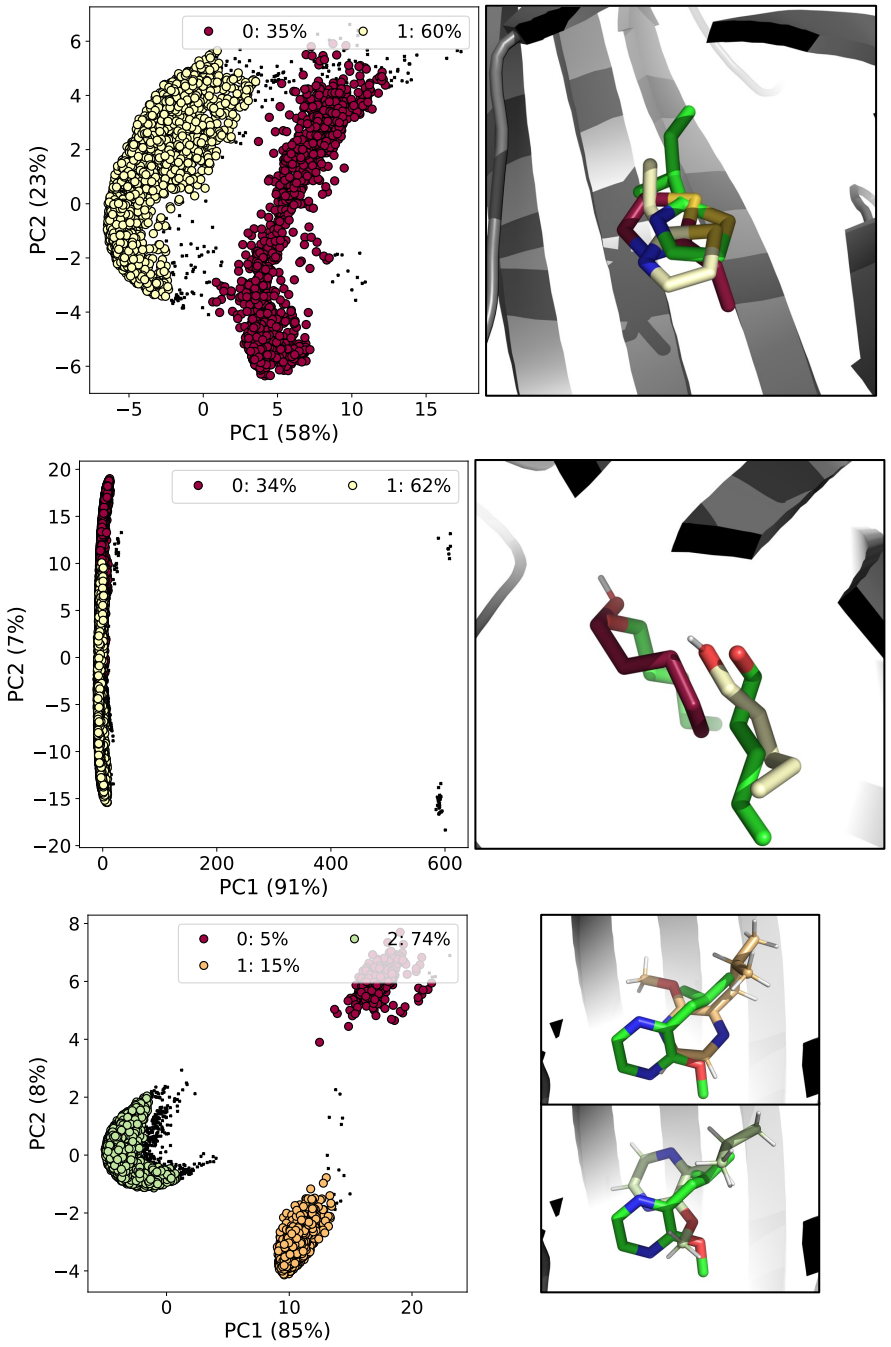


FIGURE 6.15: Binding modes of MUP1 ligands 06 (top, 1i06), 08 (middle, 1znd) and 13 (bottom, 1qy1). Crystal poses are colored in green and the binding poses are coloured to match the PCA analysis.

6.5 Summary

This chapter demonstrates the effectiveness of GCNCMC simulations in addressing key challenges in SBDD. By applying GCNCMC to two test systems T4 Lysozyme and Major Urinary Protein-1, this chapter showcases GCNCMC's ability to improve sampling efficiency and accuracy in determining binding sites and free energies compared to traditional molecular dynamics simulations.

In T4L99A, GCNCMC successfully identifies the buried binding site and captures the multiple binding modes of ligands such as benzene and toluene with populations consistent with absolute FEP and previously published results. Titration calculations estimate ligand binding affinities in line with the more established FEP method. Crucially, GCNCMC achieves this without requiring prior knowledge of binding modes. Equally, restraints and symmetry corrections are not necessary, and therefore GCNCMC can reduce computational complexity.

For MUP1, GCNCMC overcomes the limitations of traditional MD in identifying small molecule binding sites within an occluded pocket. It accurately resolves the binding modes of diverse ligands, aligning well with crystallographic poses. This further demonstrates GCNCMC's versatility in handling challenging protein-ligand systems. Like T4L99A, titration calculations were performed to good effect matching that of FEP.

Overall, this chapter demonstrates that GCNCMC has the potential to be a powerful tool in SBDD for enhancing sampling, identifying binding sites, and calculating accurate free energies, particularly for systems with occluded pockets or multiple binding modes. Given the success in locating the occluded binding pockets of T4L99A and MUP1, the application of GCNCMC to MSMD simulations is explored further in the next chapter.

Chapter 7

Enhancing Mixed-Solvent Molecular Dynamics with Grand Canonical Nonequilibrium Candidate Monte Carlo

The work presented in this chapter includes simulations and work performed by then masters project student - Victoria Nathan-Maister (VNM). Together, WP and VNM developed the theoretical formalisms and shaped the project throughout. VNM set up, ran and analyzed some of the simulations under the direct supervision of WP. The results presented are a mix of both WP and VNM. Some text in this chapter is adapted from a publication in preparation by VNM and WP. WP would like to thank VNM for her hard work on this project.

7.1 Introduction

The success in finding the occluded T4L99A and MUP1 binding sites prompted a deeper exploration into this simple but effective use case of GCNCMC. As mentioned, mixed solvent MD (MSMD) simulations have had continued success at identifying putative binding sites on protein surfaces, but have rarely been used effectively in identifying buried, occluded or cryptic pockets.[95–97,253,254]

The philosophy of MSMD[89] stems from the experimental Multiple Solvents Crystal Structure, MSCS,[47] method and involves solvating a target in a solution of water and other organic 'cosolvent' probes. As protein flexibility and dynamics are included, the opening of binding sites and competition with water molecules are, in principle, explicitly represented, though in practice sometimes poorly sampled. Some methods generally use small, water-miscible probes such as acetonitrile, isopropanol,

pyrimidine, or resorcinol,[219,226] while other methods use more hydrophobic probes, such as benzene.[95,254] In the latter cases, an artificial repulsive potential is sometimes applied between hydrophobic probes to prevent aggregation.[91,97] Methods also differ in system setup; some will opt for very high concentrations of probe molecules to encourage greater sampling but at the cost of compromising the physiological relevance of the results. Others use lower concentrations and perform longer simulations. MixMD[226] for example, solvates the protein in a spherical layer of cosolvent probes with water molecules making up the rest of the box. While this has the advantage of improving binding site identification, it biases the results as the probes are placed in the vicinity of the protein to begin with. The high local concentration of probes also makes the simulation environment less physiologically relevant.

Occluded, or cryptic, pockets can loosely be defined as pockets which require a significant degree of conformational change for either access into the pocket, or for the pocket to form. For this study, we define occluded pockets as pockets which are pre-formed but inaccessible to the solvent, and cryptic pockets as pockets which are not obvious without conformational rearrangement such as a side chain rotamer or the movement of a helix.[255] More recent applications of MSMD have studied its applicability to such systems.[93,95–97,253,256] Some success has been reported in sampling the opening of cryptic pockets which are located on the protein surface, but in the case of occluded sites, MSMD fails to map the pockets in any meaningful way. The use of accelerated, or Gaussian accelerated MD (aMD, GaMD), has been proposed as a means of improving the diffusion times by making the protein more flexible.[95,96,245] Accelerated MD methods increase sampling by applying a bias potential to the overall potential energy of the system; the resulting simulation ensemble is then reweighted to ensure a final ground state ensemble.[257] In many cases, it was shown that the addition of aMD resulted in much greater sampling of occluded and cryptic binding pockets highlighting it as a promising, system agnostic, approach. However, in some of the aMD applications to MSMD, it is unclear if ensemble reweighting was performed.

In this chapter, we propose a GCNCMC-enhanced MSMD (GC-MSMD) approach and apply it to protein systems with occluded or cryptic pockets. Taking inspiration from other studies, we use four basic cosolvent probes isopropanol, acetonitrile, n-methylacetamide and pyrimidine, which represent a broad range of the chemical space and are soluble at high concentrations thereby avoiding aggregation issues (Figure 7.1).[258] We opted for a concentration of 0.5 M to be broadly aligned with other studies, however, this decision was arbitrary. To sample all the potential binding sites while assuming no prior knowledge of the system, we use large GCMC spheres that encompass the whole of the protein being studied.

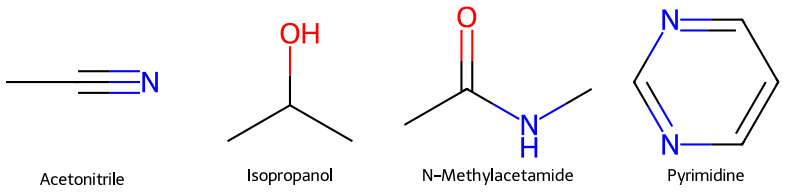


FIGURE 7.1: The four cosolvent probes used in this study.

7.2 Grid Based Analysis Methods

7.2.1 Basic Occupancy Grids

In almost all publications of MSMD methods, the analysis has generally focused on the use of 'grids', whereby, after trajectory alignment, a fictitious grid is defined around the protein and then the positions of probe molecules at each frame of the trajectory are binned onto this grid to build up an occupancy value at each grid point.[89] In other words, for how many frames was a probe present at a particular grid point? This value can then be averaged by the total number of frames to give an occupancy percentage:

$$< O >_{\mathrm{x,y,z}} = \frac{\sum_i^{N_{\mathrm{frames}}} O^i_{\mathrm{x,y,z}}}{N_{\mathrm{frames}}} \tag{7.1}$$

where $< O >_{\mathrm{x,y,z}}$ is the average occupancy of a voxel at positions x, y and z. $O^i_{\mathrm{x,y,z}}$ is the occupancy of a voxel at frame i and N_{frames} is the total number of frames in the simulation. Figure 7.2 shows this graphically. Note, there are different ways in which a probe can be assigned to a particular grid point. One way is to use an arbitrary distance such that all points within that distance to the probe are assigned while another way is to assign occupancy to all grid points which overlap with the molecules Van der Waals radii.

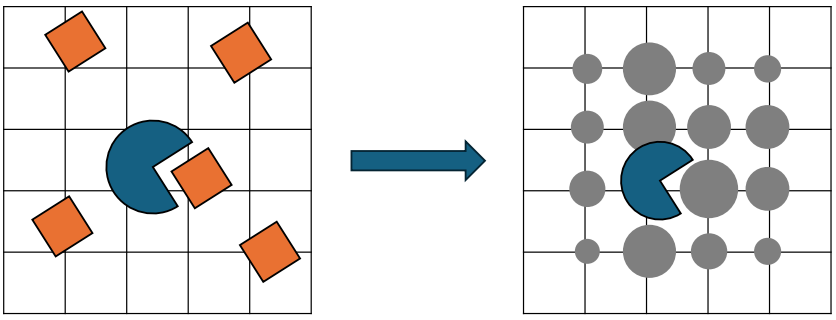


FIGURE 7.2: Graphical representation of the basic occupancy grids. Probe (orange diamonds) positions are binned onto a grid (black lines) at each frame. The total occupancy is averaged by the total number of frames such that each grid point has an associated occupancy (grey circles). In this representation, there is larger occupancy in and around the protein (blue wedge) binding site.

In this basic framework, regions of high occupancy are deemed to be potential binding sites. Grids are typically calculated for the whole probe or just particular atoms. The former gives a general idea of where a certain probe may like to reside from which one can infer the potential interactions. In the latter case, grids that are evaluated for a single atom may give a slightly finer resolution, in that grids can be built to investigate certain interaction types. For example, a grid of just the oxygen atom of isopropanol will highlight potential hydrogen bonding regions. That said, at times the position of the oxygen may be driven by favourable interactions with other parts of the probe and therefore spurious interaction maps may arise. In this study, we evaluate separate maps for the following atoms: Acetonitrile - C1 and N1, Isopropanol C2 and O1, N-methylacetamide - N1 and O1 and Pyrimidine - entire molecule.

7.2.2 MixMD Grids

In MixMD, grids are calculated slightly differently. The authors argue that looking at raw occupancies can be noisy and contain spurious minima. Instead, MixMD normalises the grids by converting the raw occupancy to a 'Z-score':

$$z_{xyz} = \frac{O_{xyz} - \mu}{\sigma} \tag{7.2}$$

where O_{xyz} is the total grid occupancy across all the frames, μ is the average occupancy over all grid points and σ is the standard deviation. This Z-score then represents the number of standard deviations between the raw grid occupancy and the mean occupancy thereby effectively down-weighting the noise to improve the signal to noise ratio.

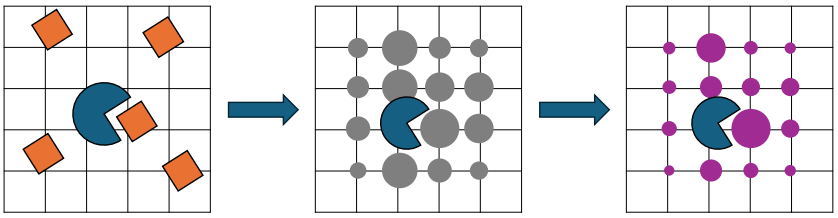


FIGURE 7.3: Graphical representation of the grids used in the MixMD protocol. Probe occupancies are normalised based on the mean and standard deviation of the overall grid. This has the effect of enhancing the more occupied sites (purple).

7.2.3 Occupancy Based Free Energy Grids

MDMix and SILCS use a different approach. For these protocols, the average occupancy of the grid is calculated in the presence and absence of the protein. In practice, this involves running a short simulation of just a probe and water mixture. By comparing the two occupancies, it is possible to discern whether a given occupancy in the protein system is greater, or lower than in bulk solvent indicating whether a particular grid point is more, or less, favourable than in bulk water. This can used to estimate the free energy of a given grid point:

$$\Delta G^{i}_{\mathrm{xyz}} = -RT \ln \left(\frac{O^{i}_{\mathrm{xyz}}}{\langle O_0 \rangle} \right), \tag{7.3}$$

where $\Delta G^{i}_{\mathrm{xyz}}$ is the free energy at grid point i with positions xyz, R is the ideal gas constant, T is temperature, O^{i}_{xyz} is the occupancy of grid point i and $\langle O_0 \rangle$ is the average probe occupancy in bulk solvent. Note, that in this case, the occupancy in both systems needs to be normalised by the same number of frames.

These grids give very rough estimates of free energy and can often be inaccurate as they rely on good sampling of the binding pocket. Intuitively, if a binding pocket is not sampled by MSMD simulations, then the free energy estimate will be infinitely high. That said, when there is sufficient sampling, these grids allow the practitioner to identify regions of high and low free energy in 3D space making it easy to target regions where binding may be advantageous. As an extension, SILCS estimates the free energy of small molecule ligands by overlaying, or docking, a molecule into the binding site and summing the free energies of the grid points with which the ligand overlaps.

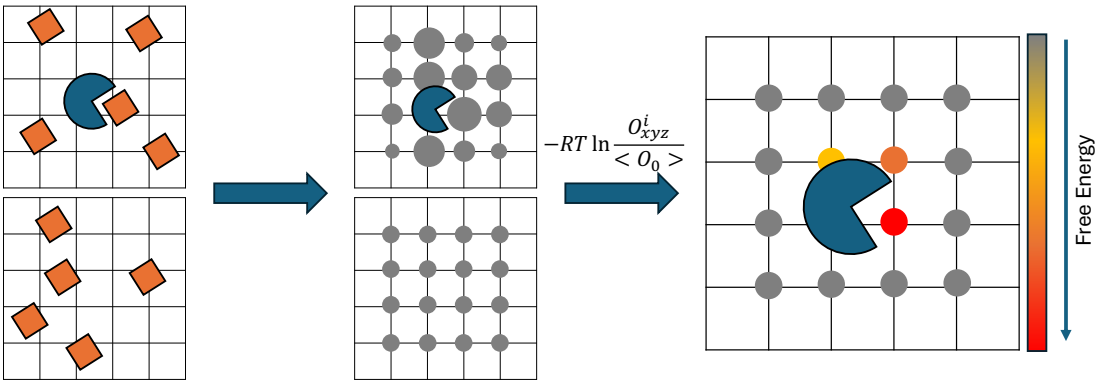


FIGURE 7.4: Graphical representation of the grids used in the MDMix and SILCS protocol. Probe occupancies in the complex system are compared to occupancies in the bulk solvent and are used to estimate the free energy of a particular grid point.

7.2.4 GCNCMC Based Free Energy Grids

Using GCNCMC simulations we present a novel alternative method for calculating the free energy of grid points. It was shown in Chapter 5 that accurate binding affinities can be calculated using the nonequilibrium works obtained during a GCNCMC simulation. Here, we propose that if we track these nonequilibrium works with respect to the ligand coordinates, then these works can be assigned to the 3D occupancy grid and use the Bennett Acceptance Ratio to estimate an equilibrium free energy for that grid point.

To do this in practice, for an insertion move, we record the C-alpha positions of the protein and the centre of geometry of the ligand being inserted, at the end of the move. We record the former so that we can align the geometry centres to a single reference frame in post-processing. We record this data at the end of an insertion move and at the start of a deletion move to give us a set of forward and reverse works. We then, as before, superimpose a grid onto our system and bin the works to grid points based on the associated coordinates; the result is two grids, or a single grid with two data arrays, containing insertion and deletion works. At each grid point, if there is sufficient data, the work arrays are used to calculate free energy using the Bennett Acceptance Ratio. Figure 7.5 shows this graphically.

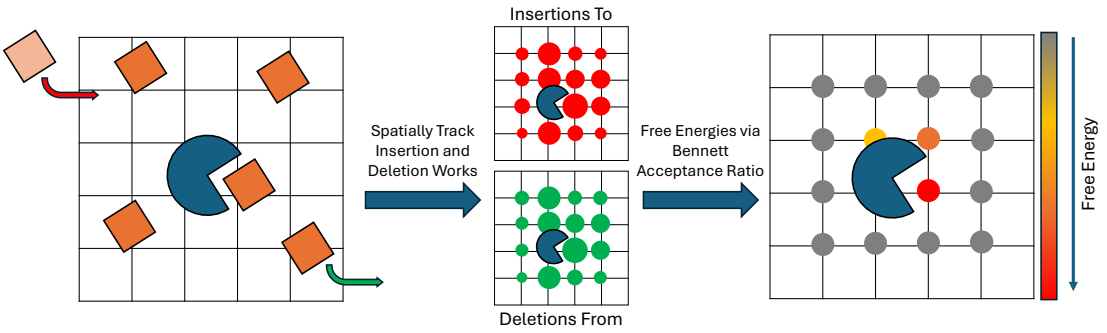


FIGURE 7.5: Graphical representation of free energy grids derived from GCNCMC simulations. Insertion and deletion works are binned onto the grid and evaluated using BAR to calculate a final free energy estimate.

7.3 Simulation Details

7.3.1 System Selection

Twelve protein targets were selected to investigate: Hen Egg White Lysozyme (HEWL, PDB: 2LYO), T4 Lysozyme L99A (T4L99A, PDB: 4W51), Major Urinary Protein 1 (MUP1, PDB: 2OZQ), Androgen Receptor (AR, PDB: 2AM9), Estrogen Receptor α (ERα, PDB: 3ERT), Peroxisome Proliferator-Activated Receptor γ (PPARγ, PDB: 3U9Q), Phosphoinositide-Dependent Kinase-1 (PDK1, PDB: 2Q8F), Protein Tyrosine Phosphatase 1B (PTP1b, PDB: 1SUG), Heat Shock Protein 90 (HSP90, PDB: 1YER), Tumor Protein 53 y220C (p53, PDB: 2J1X), Glycogen Phosphorylase B (GPB, PDB: 1P4G) and Extracellular Signal-regulated Kinase 2 (ERK2, PDB: 4GSB). Protein structures were obtained from X-Ray crystallographic data available in the Protein Data Bank.[56] All systems were selected from various MSMD studies, and their respective PDB codes, with a wide range of binding pocket characteristics. T4L99A, MUP1 and ERK2 have all appeared before in this thesis and are included here for comparison. Twelve protein targets were selected to investigate: Hen Egg White Lysozyme (HEWL, PDB: 2LYO), T4 Lysozyme L99A (T4L99A, PDB: 4W51), Major Urinary Protein 1 (MUP1, PDB: 2OZQ), Androgen Receptor (AR, PDB: 2AM9), Estrogen Receptor α (ERα, PDB: 3ERT), Peroxisome Proliferator-Activated Receptor γ (PPARγ, PDB: 3U9Q), Phosphoinositide-Dependent Kinase-1 (PDK1, PDB: 2Q8F), Protein Tyrosine Phosphatase 1B (PTP1b, PDB: 1SUG), Heat Shock Protein 90 (HSP90, PDB: 1YER), Tumor Protein 53 y220C (p53, PDB: 2J1X), Glycogen Phosphorylase B (GPB, PDB: 1P4G) and Extracellular Signal-regulated Kinase 2 (ERK2, PDB: 4GSB). Protein structures were obtained from X-Ray crystallographic data available in the Protein Data Bank.[56] All systems, and their respective PDB codes, were selected from various MSMD studies with a wide range of binding pocket characteristics. T4L99A,

MUP1, and ERK2 have all appeared before in this thesis and are included here for comparison.

As mentioned, the probe set includes isopropanol, acetonitrile, n-methylacetamide and pyrimidine as in MixMD.[258]

7.3.2 System Setups

This study aims to investigate the performance of the GC-MSMD method for locating occluded sites, and as such, the *apo* structure of each system was used, where available. This ensures that the simulations do not start in an 'open' configuration owing to the presence of a bound ligand. In the case of HEWL, AR, and ERα, the simulations began from *holo* conformations as no suitable *apo* structure was found.

For each protein system, the crystal waters were removed, and missing loops and heavy atoms were added using PDBFixer.[247] Each system was protonated according to a pH of 7.0 using the PDB2PQR[259] webserver which uses Propka 3[260] to determine the protonation states of residues. Each protein system was first solvated in a box of just TIP3P[221] water, neutralizing ions and 0.15 M of NaCl, with a 12 Å buffer to prevent self-interactions. Energy minimization was performed on this structure, followed by 10 ns of equilibration with a 2 fs timestep. The equilibration featured 0.5 ns of NVT equilibration while annealing the system to 298 K, 1.5 ns of further NVT equilibration and 8 ns of NPT equilibration using the Monte Carlo barostat at 1 bar.

This preliminary equilibration was intended to eliminate any high-energy protein conformations such that a shorter equilibration could be performed for each of the production simulations with each probe. Proteins were then removed from the water box and resolvated in a box of TIP3P water,[221] neutralizing ions, 0.15 M NaCl and 0.5 M of cosolvent. For each probe/protein combination, ten repeats were performed, each with a uniquely generated solvent box. A shorter equilibration was performed for 4 ns on each repeat, (0.5 ns of NVT equilibration with heating to 298 K, 0.5 ns of further NVT and 3 ns of NPT equilibration) before running production MSMD and GC-MSMD simulations (detailed below).

A Python module, named SetCo, was developed to build these mixed solvent boxes using only a few lines of code (`https://github.com/WillGPoole/SetCo`). In short, the module automatically generates input files and executes PackMol[261] for the box generation. Probe parameters are also automatically generated using Ambertools.[262] Development of this module is still ongoing.

7.3.3 MD Production

Classical mixed-solvent MD simulations were performed using OpenMM 8.0[161] with the Amber ff14SB forcefield.[220] The acetonitrile, isopropanol, N-methylacetamide and pyrimidine probes were parameterised using the GAFF forcefield with AM1-BCC charges.[222,223] All simulations were performed at 298 K using a Langevin[168] BAOAB integrator with a friction coefficient of 1 ps^{-1} and a time step of 4 fs (hydrogen mass = 4 amu). The cut-off for nonbonded interactions was 12 Å with a switching function for the Lennard-Jones interactions applied at 10 Å. Particle mesh Ewald[183] was used to calculate the effect of long-range electrostatics. Classic MSMD simulations were performed in the NPT ensemble at a pressure of 1 bar using a Monte Carlo barostat. For each combination of target and probe, 50 ns production MD simulations were run for each of the 10 unique solvent boxes. The final 10 ns of each repeat were compiled into one large trajectory for analysis.

7.3.4 GCNCMC Production

GCNCMC/MD simulations were performed using the *grandlig* python module using the same settings as the classical MSMD simulations described above. After the 4 ns equilibration from MSMD, the system was further equilibrated in the μVT ensemble for 1.5 ns with a GCNCMC move attempted every 10 ps for a total of 150 moves.

Production simulations ran for 25 ns with a GCNCMC move attempted every 0.05 ns, for a total of 500 proposed moves. Each move included 500 perturbation steps and 25 propagation steps, totalling 0.05 ns, such that the total simulation time would sum to 50 ns (25 ns + (500 moves x 0.05 ns/move)), as in the MSMD simulations. Each trial move is accepted or rejected based on the acceptance criteria in equations 2.172 and 2.173. If a move is accepted, the simulation proceeds with the new state, and if a move is rejected, the simulation proceeds from the state prior to the move.

For each target, a spherical GCMC region encapsulating the whole protein was defined by anchoring the sphere centre to the closest C-alpha atom to the centre of geometry of the protein. The sphere is given a fixed radius to cover as much of the protein as possible. This reduces the number of insertions and deletions occurring in bulk solvent, while still accounting for the fact that the protein may change orientation during the simulation. The atoms used to anchor the sphere and the radii are detailed in Table 7.1. The excess chemical potential of the acetonitrile, isopropanol, n-methylacetamide and pyrimidine were taken to be -2.69, -3.17, -8.29 and -4.54 kcal mol^{-1} respectively, and were calculated as described previously (Sec. 4.3.1).

TABLE 7.1: Protein atoms used to anchor the GCMC sphere and the sphere radius.

Protein	Protein Atom	Radius / Å
Hen Egg White Lysozyme	Leu56-CA	20
T4 lysozyme L99A	Gln105-CA	26
Major Urinary Protein 1	Gly118-CA	24
Androgen Receptor	Leu744-CA	30
Estrogen Receptor	Leu453-CA	30
Peroxisome Proliferator-Activated Receptor	Ile325-CA	32
Phosphoinositide-Dependent Kinase-1	Met186-CA	33
Protein Tyrosine Phosphatase 1B	Val213-CA	29
Heat Shock Protein 90	Leu48-CA	32
Tumor Protein 53 Y220C	Thr253-CA	26
Extracellular signal-regulated kinase 2 (large)	Leu148-CA	36
Extracellular signal-regulated kinase 2 (small)	Leu114, Pro91	12
Glycogen phosphorylase B	Arg138-CA	50

7.3.5 Data Analysis

7.3.5.1 Occupancy Analysis

For both MD and GCNCMC, the final 10 ns of each simulation repeat were concatenated together to give 100 ns of equilibrium frames compiled from the end of each of the 10 repeats. Using VMD's[263] volmap feature, a 3D grid with voxels every 0.5 Å was generated. At each frame, a grid point is assigned a value of 1 or 0 to indicate if that grid point is occupied in that frame. The atom, or atoms, of interest are treated as spheres with their atomic radii, and a grid point is said to be occupied if it resides within this sphere. The grid occupancies are then summed and averaged by the total number of frames given to the analysis. The final result is a grid where each point has a value between 0 and 1, with 1 indicating that a grid point was occupied in 100% of the analysed frames.

7.3.5.2 GCNCMC Based Free Energy Grids

To generate free energy maps derived from GC-MSMD simulations, work measurements must be recorded throughout the simulation, along with the 3D coordinates associated with the work value and the 3D coordinates of some reference to use for alignment. In this case, we use the protein C-alpha atoms. For insertion moves, we record the centre of geometry for the final coordinates of the molecule being inserted, and for a deletion move, the initial coordinates. The C-alpha positions

are recorded at the same time, and can then be used to define a transformation for each centre of geometry measurement to align all the coordinates to a single reference structure - e.g. the first frame.

Using the same grid as the occupancy maps as a reference, we use the aligned centre of geometries of the moves (initial for deletion, final for insertion) to assign work measurements to any voxel within 3 Å of the move; though using the Van der Waals radii of the molecule may be a better approach in the future. The binned insertion and deletion works at each point are then used in BAR to calculate a transfer-free energy estimate for that grid point. The excess chemical potential of the probe molecule is then subtracted to give a final free energy estimate. These grids can now be viewed in PyMol and contoured based on their free energy. An error estimate from the *pymbar* BAR function is also assigned to each grid point and is based upon the uncertainty measurement presented by Bennett.[206]

7.4 Results and Discussion

7.4.1 Hen Egg White Lysozyme

Simulations and Analysis by VNM. Text by WP.

Hen Egg White Lysozyme (HEWL) is a very simple test system for MSMD simulations and was one of the first systems tested by the MSCS method.[264] It was also one of the test systems for the early development of MixMD.[92] Although the binding site of this protein is not occluded, it provides a good baseline as it binds one of our probes, acetonitrile, making comparison to the crystal structure straightforward. This also allows us to ensure that the benefits of vanilla MSMD are not lost owing to the addition of the GCNCMC component, as binding is expected with both methods. Figure 7.6 shows the resulting occupancy analysis from MSMD and GC-MSMD simulations. In both cases, we see a clear signal for acetonitrile overlapping with the crystal pose. The grids in the figure are contoured at an occupancy level of 40% and show two other possible regions which favourably attract an acetonitrile probe. Interestingly, contouring the MSMD grid at 50% or higher results in all three sites disappearing, giving the illusion they are equally stable. Conversely, in GC-MSMD increasing the contour level to 80% results in the loss of the spurious minima leaving only the crystal pose. The reason for these two extra sites is unknown as there are no acetonitrile molecules modelled in the crystal structure at these locations. These regions may correspond to wrongly modelled water molecules, crystal packing interactions, or forcefield limitations.

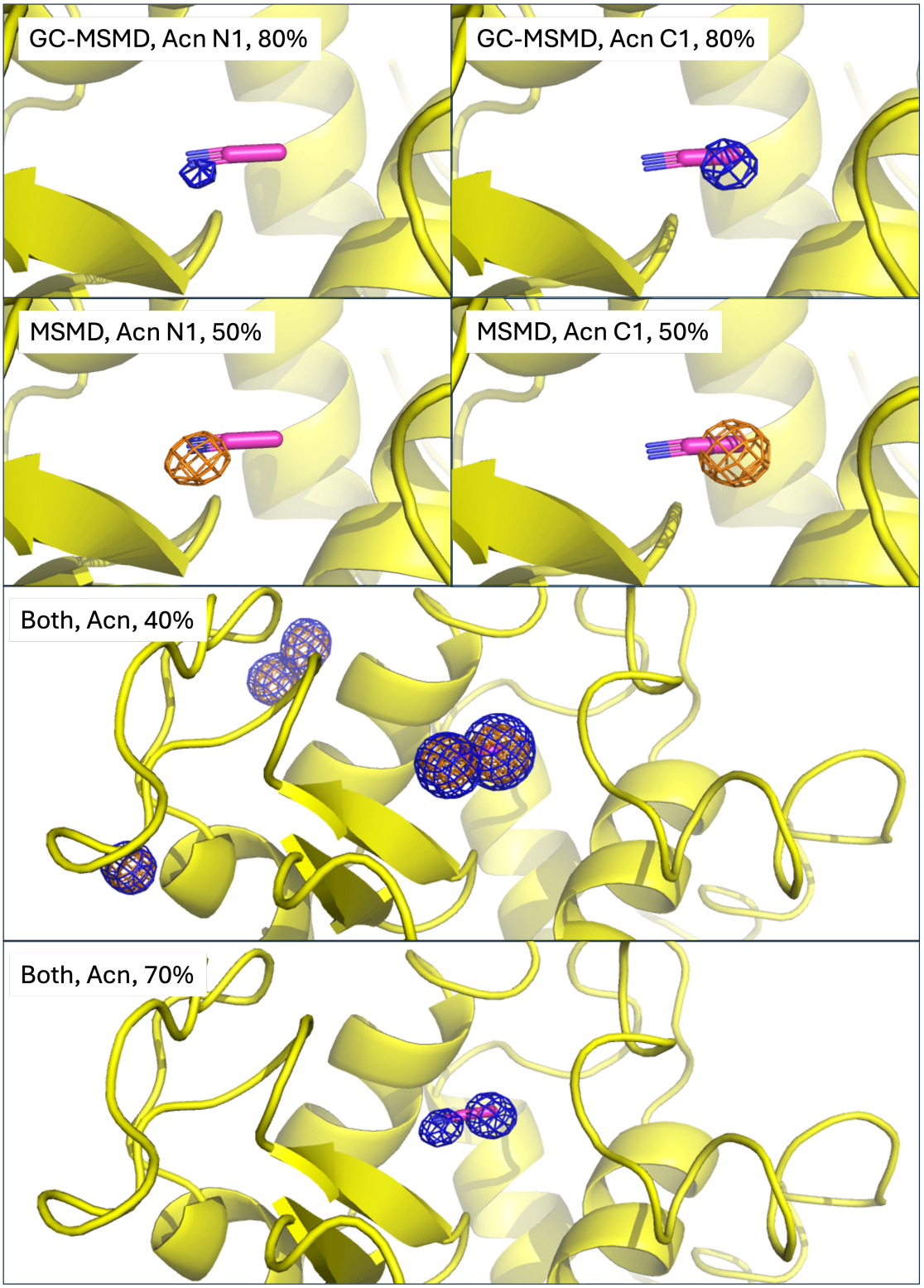


FIGURE 7.6: From top to bottom: GC-MSMD results for ACN probes as occupancy meshes (blue), illustrating that GC-MSMD captures the position of the bound acetonitrile at high occupancy. MSMD (Orange) captures the same pose, but at lower occupancy. Below, the results of both simulation types at 40% and 70% occupancy are overlaid, showing that while both GC-MSMD and MSMD identify other potential binding sites, for GC-MSMD these sites disappear at high occupancy leaving only the known site.

It is thought that GC-MSMD gives a higher occupancy owing to insertion moves quickly replacing probe molecules which may repeatedly bind and unbind during vanilla MD portions of the simulation. This is indicated in Figure 7.7 which shows that in GC-MSMD simulations, should the acetonitrile unbind, it is quickly replaced by insertion moves, whereas MSMD simulations must rely on dynamics alone for binding. Furthermore, probes which may become kinetically trapped in less favourable binding sites can be removed by GCNCMC reducing the occupancy of these sites compared to MSMD.

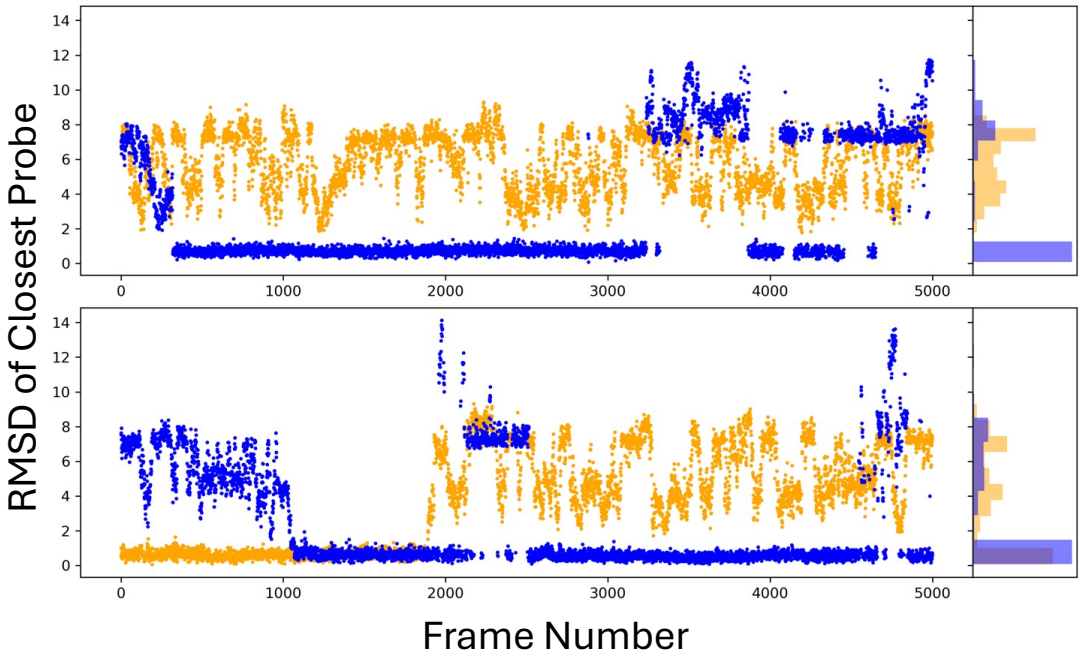


FIGURE 7.7: Acetonitile probe RMSD, measured in Angstroms, to the crystal pose as a function of simulation time. The RMSD is reported for the probe which is closest to the crystal pose in a particular frame and therefore should not be interpreted as continuous. Orange represents the MSMD simulation where it is clear that an acetonitrile molecule is not bound for significant proteins of the simulation. The blue dots indicate GC-MSMD simulations. Histograms on the right represent the distribution of the data in the time series plot. These plots represent two of the ten repeats and are chosen to best illustrate the point.

7.4.2 Major Urinary Protein-1

Simulations and Analysis by VNM. Text and Figures by WP.

The success of GC-MSMD simulations with MUP1 has already been demonstrated in the previous chapter. However, in that case, the molecules used as probes were relatively large. Here, we use the set of smaller probes, which may find it easier to diffuse into the MUP1 pocket.

As predicted, some binding of the small probes in classical MSMD simulations was observed. This is likely due to the smaller probe size. However, this binding was limited to, at maximum, 50% of analysis frames for the nitrogen atom of acetonitrile, whereas the pyrimidine in GC-MSMD simulations achieved a maximum occupancy of 90%.

Visual inspection of how well the occupancy maps overlay with the known binders studied previously shows that both MSMD and GC-MSMD managed to replicate many of the interactions, albeit at lower occupancies for the former. Figure 7.8 shows a selected subset of these maps overlaid with crystal structures with PDB codes 1i06, 1znd and 1qy1. For the first, 1i06, pyrimidine was the only probe in MSMD to overlay with the aromatic core of the ligand with an occupancy of 20%. In GC-MSMD, pyrimidine overlays well for 90% of frames as well as being well placed in two other positions corresponding to the ligand alkyl chain. The middle row, 1znd, shows the two binding modes of pentanol indicating two positions where a hydrogen bonding or alcohol group may like to reside. In GC-MSMD the alcohol-containing probe, isopropanol, reproduces one of these two sites with 50% occupancy, and a second, lower occupancy site adjacent, corresponding to the slightly different binding pose of hexanol (1zne, not shown). Interestingly, in MSMD the interaction of the alcohol was only picked up by acetonitrile, potentially indicating that acetonitrile may be better at diffusing into occluded pockets compared to isopropanol. In GC-MSMD, acetonitrile was also found in these positions but at a lower percentage occupancy than isopropanol. Finally, the ligand 1qy1 is bound more centrally in the pocket and has interactions common to both 1i06 and 1znd. As before, pyrimidine and acetonitrile (not shown) seem to give the best overlay with the ligand in MSMD for 20% and 40% occupancy respectively. These two probes map the same interactions GC-MSMD simulations but at higher occupancies, with acetonitrile mapping both the hydrophobic and hydrophilic interactions at an occupancy of 60%.

Overall, as expected for this occluded site, GC-MSMD outperforms vanilla MSMD in terms of overall occupancy, as well as giving a more complete mapping of the site. This implementation of MSMD, in contrast to the implementation using larger molecules in the previous chapter, managed to map some of the pockets, indicating a probe size limitation in MSMD simulations. Owing to the high occupancies in the pocket observed using GC-MSMD there are no false positives as all other sites are observed at a lower occupancy whereas, in MSMD, some of the lower occupancy sites may be lost in the noise.

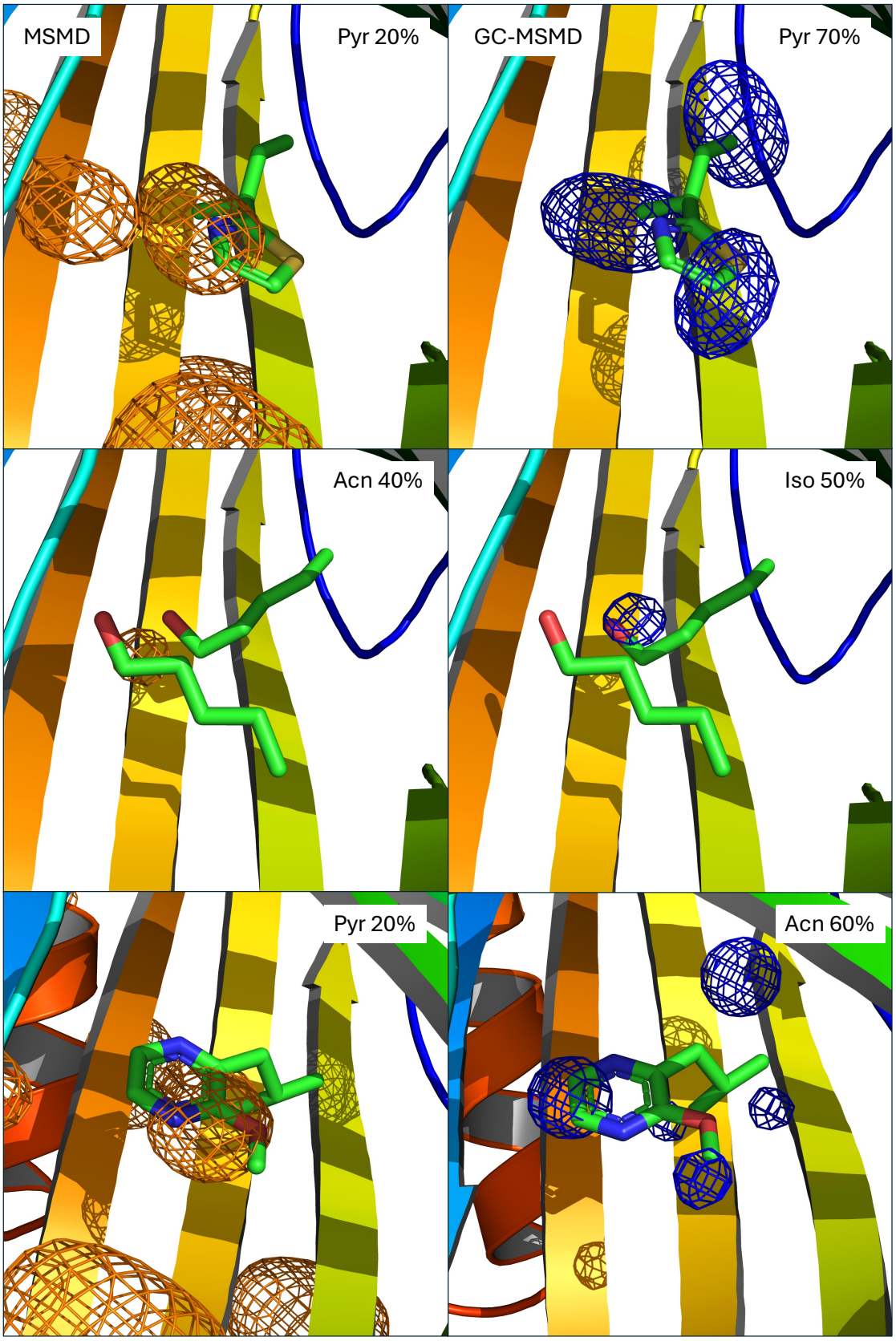


FIGURE 7.8: Occupancy analysis from MSMD (left column) and GC-MSMD simulations (right coulmn) of MUP1 with the indicated probes and occupancies. The maps are overlaid onto structures of known binders, and the PDB codes for the top, middle and bottom rows are 1i06, 1znd and 1qy1, respectively. Occupancy percentages represent the maximum occupancy observed for that protocol-probe combination.

7.4.3 Androgen Receptor

Simulations and Analysis by VNM. Text and Figures by WP.

Androgen Receptor (AR) is a nuclear transcription factor which mediates the effects of androgens such as testosterone or dihydrotestosterone. It contains two significant binding regions: A DNA-binding domain and a ligand-binding domain, the latter of which binds androgenic hormones and is largely occluded. Several cosolvent MD methods have previously investigated this occluded site.[96,254] Notably, it has been mapped by MixMD in the absence of any enhanced sampling technique.[265] However, in the present study, MSMD simulations completely failed to map the occluded ligand-binding site with any of four probes (Figure 7.9). It is possible that this discrepancy can be attributed to the setup protocols used by MixMD,[265] in which probe molecules are placed in a shell around the protein rather than dispersed through the solvent, possibly increasing the likelihood that probes will access challenging sites.

Unlike in our classic MSMD simulations, all four probes sampled the binding site to some degree in GC-MSMD simulations. With n-methylacetamide, acetonitrile, pyrimidine and isopropanol visiting the site for 30%, 60%, 70% and 70% of frames respectively. Figure 7.9 shows the mapping of each probe in comparison to a bound ligand in the PDB (2axa). The nitrogen of ACN maps areas associated with hydrogen bond acceptance, the carbonyl oxygen of NME lines up with the carbonyl oxygen of the ligand, the carbon atoms of ACN and ISO map the alkyl side chain, and PYR generally maps most of the ligand at lower occupancy.

In all cases, however, there is a region deeper in the pocket, occupied by a fluorophenyl ring, which is not mapped by any of the probes. Without the presence of a ligand, this sub pocket becomes occupied by a tryptophan residue (TRP741), requiring a conformational change to allow binding (Figure 7.10, left). This movement is unfortunately not captured by GCNCMC insertion moves, highlighting a possible sampling issue which will need to be addressed in the future. Finally, two allosteric sites (2piq and 2pix) are well reproduced by both methods as both are relatively solvent exposed (Fig. 7.10, right).

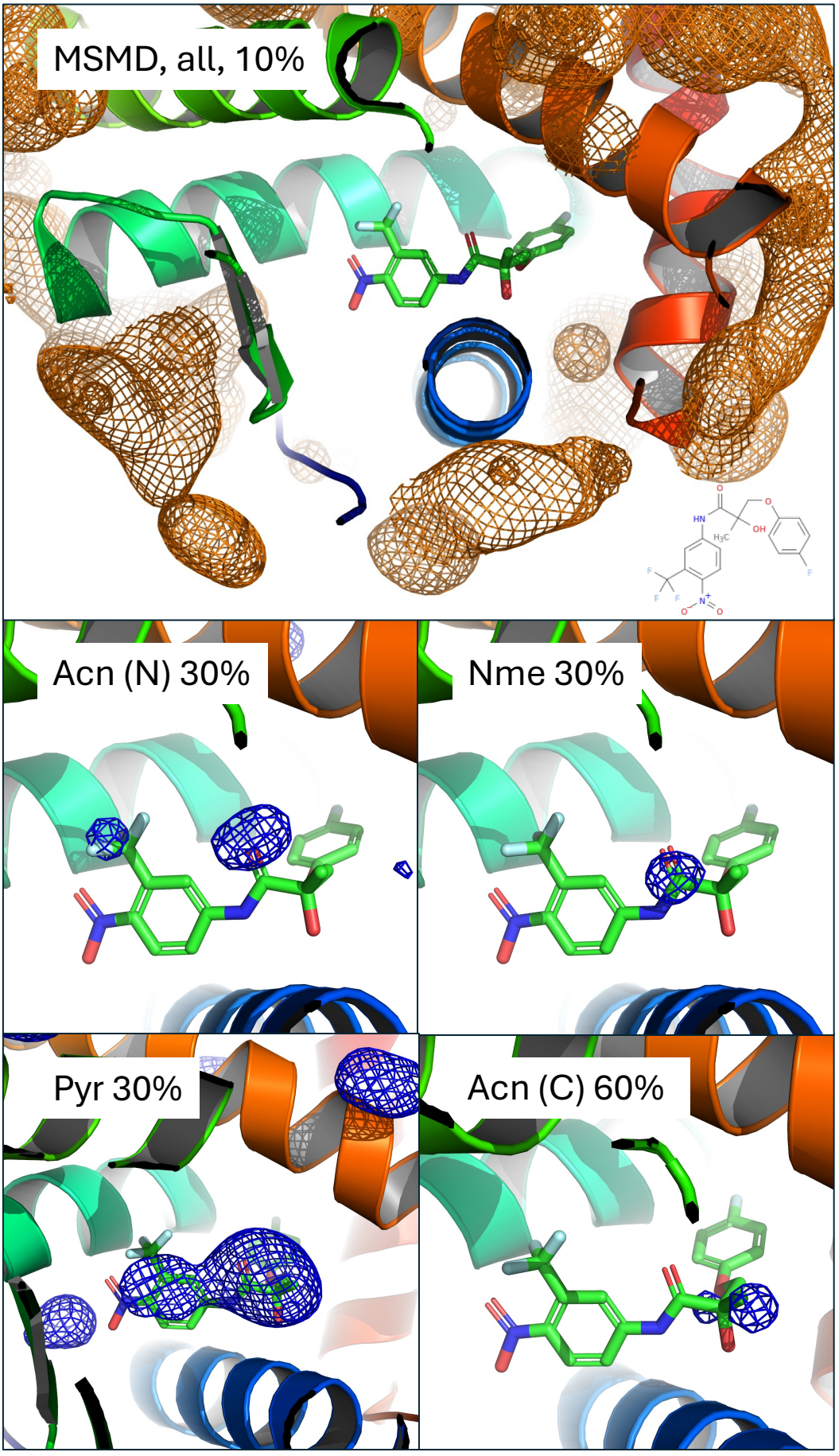


FIGURE 7.9: Occupancy analysis from MSMD (top) and GC-MSMD simulations (middle and bottom) of AR with the indicated probes and maximum observed occupancies. The maps are overlaid onto PDB 2axa.

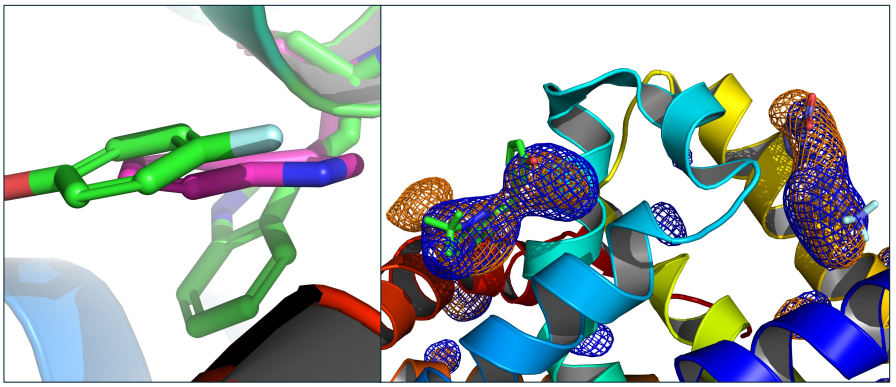


FIGURE 7.10: Left: Fluorophenyl ring of the 2axa ligand (green) overlayed with the tryptophan conformation seen in the starting structure (magenta). Upon ligand binding, the tryptophan is displaced (green). Right: Mapping of the two allosteric pockets 2pix (leftmost ligand, green) and 2piq (rightmost ligand, purple) by MSMD (orange) and GC-MSMD (blue) at an occupancy of 60%

7.4.4 p53-Y220C

Simulations and Analysis by VNM. Text and Figures by WP.

The Y220C mutant of the tumour suppressor protein p53 is implicated in approximately 125,000 cancer cases per annum. The WT protein plays a crucial role in the regulation of critical gene networks with impaired p53 signalling being the hallmark of an estimated 20 million cases yearly.[266] This particular mutant destabilises the, already fragile,structure of p53, causing a severe reduction in its melting temperature from 45 °C to 36 °C thus causing it to rapidly unfold at body temperature resulting in a loss of function. Interestingly, it has been shown that structurally unstable mutants which display WT-like conformations still retain some transcriptional activity at low temperatures, indicating that stabilising these mutants with a small molecule drug may reactivate their activity. p53-Y200C has been the subject of many drug design campaigns without much avail.[267] Crucially, this particular mutant has a narrow, mutationally induced pocket at the surface.[266,268]

Given the exposed nature of this binding pocket, there was little difference observed between MSMD and GC-MSMD simulations (Figure 7.11). Encouragingly, a clear mapping of certain interactions was seen for both methods. GC-MSMD occupancy maps generally have a slightly higher occupancy but in this case, the discrepancy is negligible and does not improve the interpretation of the results.

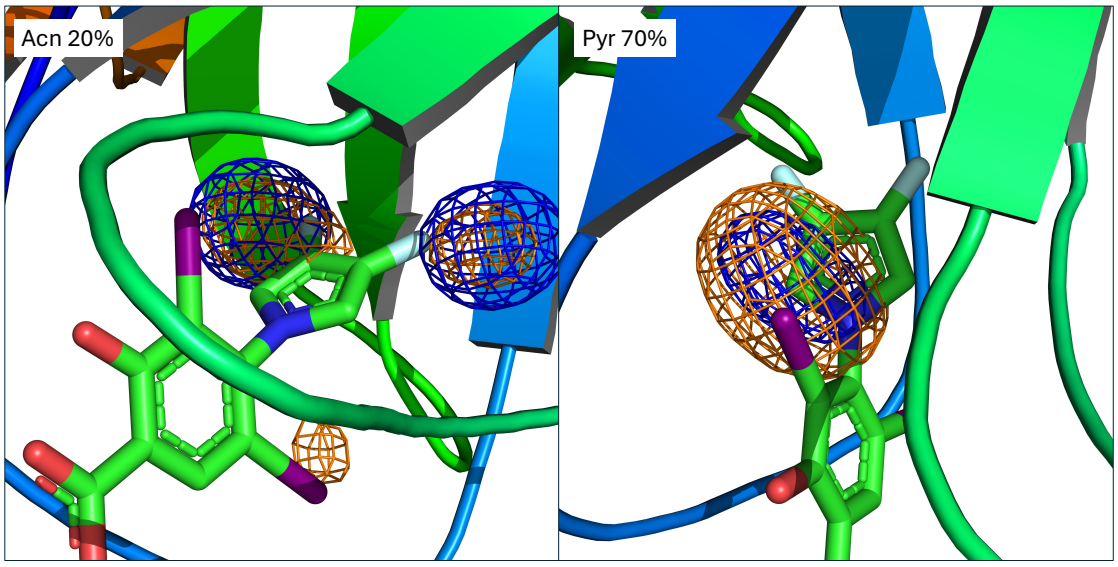


FIGURE 7.11: Left: Acetonitrile MSMD (orange) and GC-MSMD (blue) occupancy maps both contoured at 20%. Right: Acetonitrile MSMD (orange) and GC-MSMD (blue) occupancy maps both contoured at 70%. Overlaid structure 8a32. MSMD and GC-MSMD map the pocket well with the same maximum occupancies.

7.4.5 Heat Shock Protein 90

Simulations and Analysis by VNM. Text and Figures by WP.

Heat Shock Protein 90 (HSP90) is another highly abundant protein which plays an essential role in many cellular processes. HSP90 acts as a molecular chaperone to many proteins, including those associated with cancer cells, leading to enhanced cancer growth and survival. it is therefore one of the hottest targets in the pharmaceutical industry.[269] Crucially, HSP90 has been studied extensively *in silico* owing to its various difficulties including large backbone rearrangements upon ligand binding and highly preserved water networks.[95,96,123,148,150]

Figure 7.12 shows the occupancy results of MSMD and GC-MSMD simulations. Both methods generally do well at mapping the more exposed region of the binding site. The aforementioned occluding backbone is shown in orange for PDB code 2xdl. The binding of larger ligands, such as that in 3ft8, causes a large rearrangement of the backbone, shown in purple. Unfortunately, neither GC-MSMD moves nor regular MSMD can capture this rearrangement. These results are in line with previously published MSMD results with only accelerated methods achieving partial sampling of the occluded region.[95,96]

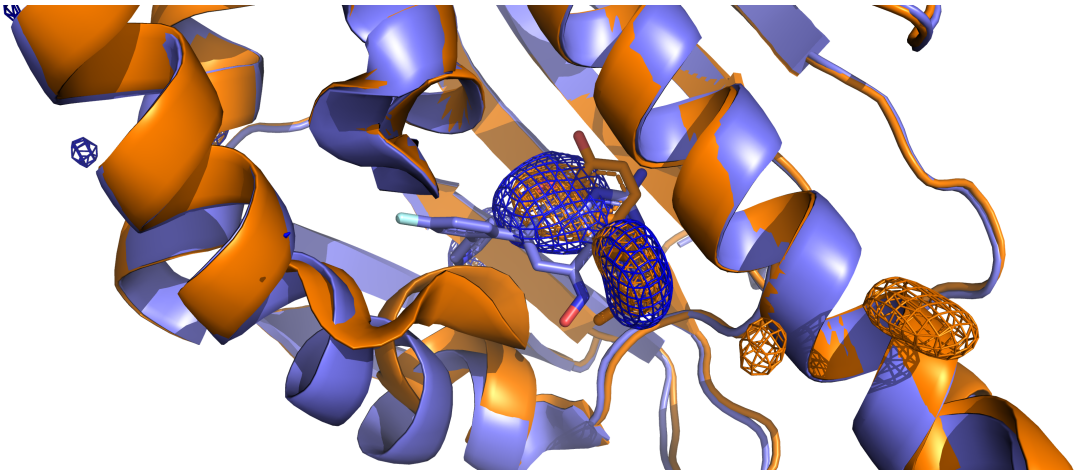


FIGURE 7.12: Pyrimidine MSMD (orange) and GC-MSMD (blue) maximum occupancy maps both contoured at 40%. Overlaid structures are 3ft8 (purple) and 2xdl (orange). MSMD and GC-MSMD map the more exposed region of the pocket well. The occluding backbone, as in 2xdl, prevents further exploration of the pocket.

7.4.6 Protein Tyrosine Phosphatase 1B

Simulations and Analysis by VNM. Text and Figures by WP.

Protein Tyrosine Phosphatase 1B (PTP1B) is a negative regulator of the insulin receptor and holds promise as a target for the treatment of type II diabetes. The pocket of interest is fully occluded by the C-terminal helix and has failed to be mapped by any published MSMD simulations (Figure 7.13). Upon ligand binding, this helix becomes fully disordered and is often missing in many ligand-bound crystal structures. As expected, this pocket was not mapped by either of our MSMD or GC-MSMD simulations. Again, some partial mapping of this site has been seen in accelerated MD approaches, albeit without any reweighting.[95]

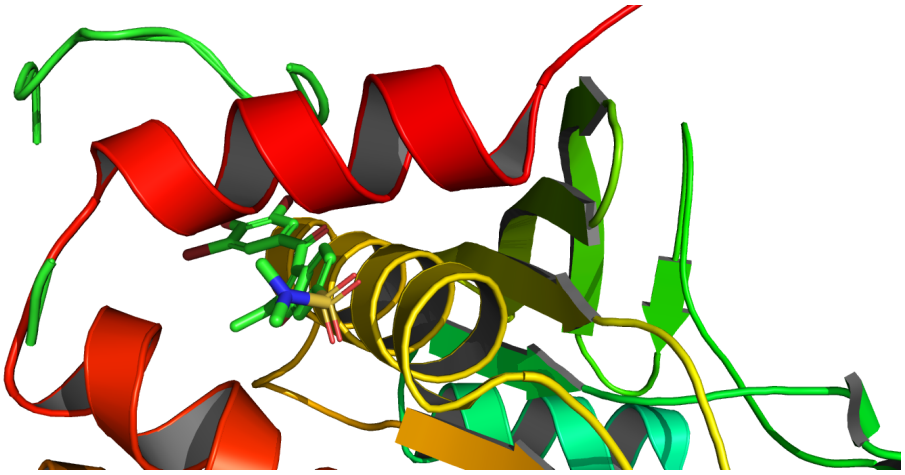


FIGURE 7.13: Overlaid crystal structures of PTP1B. The bound ligand (green, 1t48) is occluded by the red alpha helix in the *apo* structure (1sug). Upon ligand binding the alpha helix becomes disordered as shown by the green cartoon with no secondary structure.

7.4.7 ERK2 Revisited

Simulations, Analysis, Text and Figures by WP.

Finally, we return to the MiniFrag system tested in Chapter 3. Vanilla MSMD of the MiniFrag probes with ERK2 achieved partial mapping of the three subsites within the kinase active site. Sites 1a and 1c were mapped well by two of the eight MiniFrags while site 1b saw minimal exploration owing to the occluding lysine residue discussed in Chapter 3. Here, we apply GC-MSMD simulations to ERK2 using the same MiniFrag probes. We also perform separate simulations using a large, full protein, sphere and a smaller sphere encompassing just the binding site. As a reminder, the MiniFrags are shown in Figure 7.14 with their calculated excess chemical potentials.

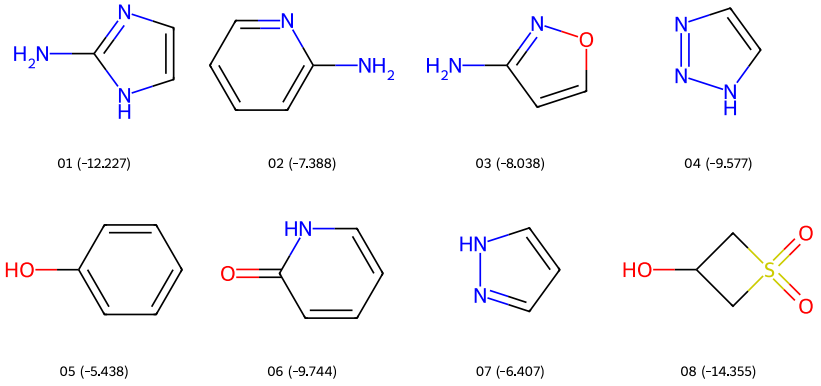


FIGURE 7.14: MiniFrags which bind to ERK2 and their calculated excess chemical potentials in water. MiniFrags are numbered with μ' in parenthesis in units of kcal mol^{-1}.

Figures 7.15 and 7.16 show the occupancy maps from the large and small sphere GC-MSMD simulations respectively and Table 7.2 gives the overall results for each probe. In general, little to no improvement over regular MSMD was observed except in the rare case where GC-MSMD returned slightly higher occupancies. However, for some probes, GC-MSMD resulted in lower occupancies. This is likely owing to the exposed nature of the site and the difficulties associated with moving the lysine side chain residue and displacing bound water molecules which may occupy the ligand binding sites. These quirks mean that GCNCMC moves are often rejected and it is likely that regular MSMD outperformed this system simply because of the longer run time, though this will need to be explored further in the future.

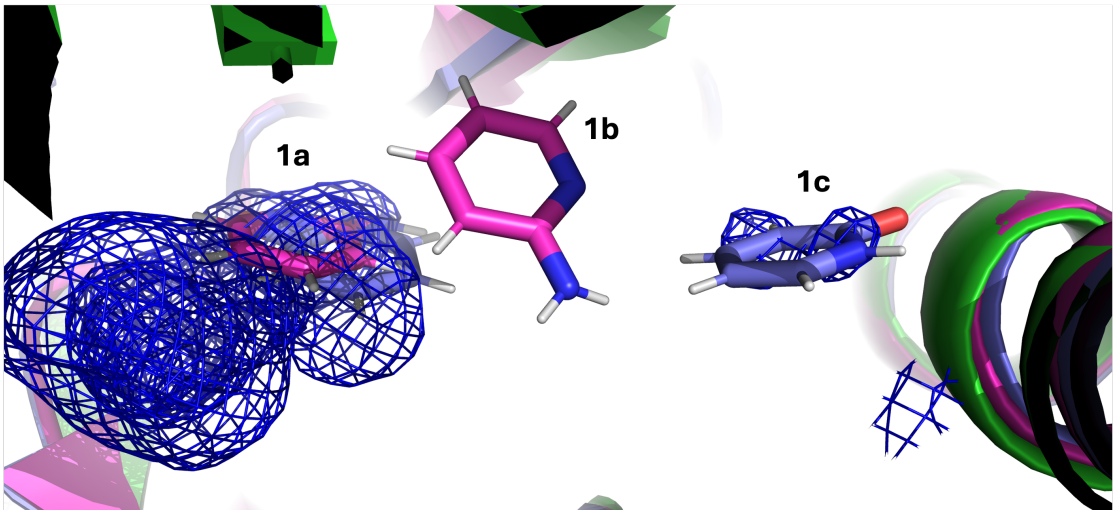


FIGURE 7.15: GC-MSMD occupancy grids for all eight MiniFrags contoured at 40%. Representative crystal ligands in sites 1a, 1b and 1c are shown. These simulations are performed using a **large** GCMC sphere which encapsulates the entire protein.

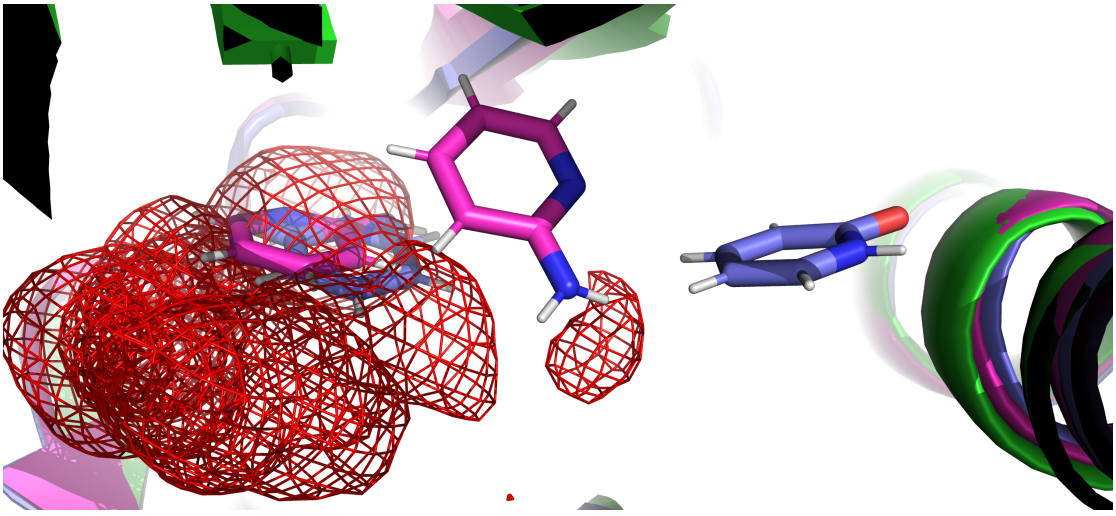


FIGURE 7.16: GC-MSMD occupancy grids for all eight MiniFrags contoured at 40%. Representative crystal ligands in sites 1a, 1b and 1c are shown. These simulations are performed using a **small** GCMC sphere which encapsulates only the binding site.

TABLE 7.2: Maximal occupancies from MSMD and GC-MSMD simulations of ERK2. The columns represent the 8 MiniFrags and 4 MSMD probes. Rows represent the 3 subsites for each method.

		01	02	03	04	05	06	07	08	pyr	iso	nme	acn
MSMD	1a	10	20	50	35	30	30	20	00	40	30	00	20
	1b	00	10	00	00	00	00	00	00	00	00	00	00
	1c	00	00	00	20	00	00	50	00	00	10	00	20
GC-MSMD (big)	1a	20	10	50	50	30	30	30	00	-	-	-	-
	1b	00	00	00	00	00	00	00	00	-	-	-	-
	1c	00	00	00	00	00	00	40	00	-	-	-	-
GC-MSMD (small)	1a	20	30	20	30	20	40	60	00	-	-	-	-
	1b	00	20	00	00	00	00	10	00	-	-	-	-
	1c	00	00	00	00	00	00	10	00	-	-	-	-

7.4.8 Overall Results

Only a subset of the systems simulated have been discussed in detail. Table 7.3 gives the overall results for all the systems studied. To do this, we have employed a ranking system akin to university degree classifications: 1, 2.1, 2.2 and 3, where 1 refers to the full mapping of the binding site, 2.1 is a high partial mapping, 2.2 is low partial mapping and 3 is where the pocket is not mapped at all. Scores are assigned subjectively based on how well crystal ligands are mapped. As an example, MUP1 would receive a 1 in GC-MSMD and a 2.2 for vanilla MSMD. The method which gives a higher occupancy percentage is also indicated, however, it should be mentioned that higher occupancies do not necessarily mean greater accuracy and it may be a product of kinetic trapping. Further studies should focus on assessing the convergence of these occupancies.

TABLE 7.3: Summary of the results for all twelve systems studied in MSMD and GC-MSMD simulations. A '-' indicates that both methods mapped the site with equal occupancy.

Protein	MSMD Score	GC-MSMD Score	Highest Occupied
Hen Egg White Lysozyme	1	1	GC-MSMD
T4 lysozyme L99A	3	1	GC-MSMD
Major Urinary Protein 1	2.2	1	GC-MSMD
Androgen Receptor	3	1	GC-MSMD
Estrogen Receptor α	2.2	2.1	-
Peroxisome Proliferator-Activated Receptor γ	3	2.2	GC-MSMD
Phosphoinositide-Dependent Kinase-1	3	3	-
Protein Tyrosine Phosphatase 1B	3	3	-
Heat Shock Protein 90	2.2	2.2	-
Tumor Protein 53 y220C	2.1	2.1	GC-MSMD
Extracellular signal-regulated kinase 2	2.2	2.2	GC-MSMD
Glycogen phosphorylase B	3	2.1	GC-MSMD

Table 7.3 indicates that vanilla MSMD simulations failed to map, even partially, 6 of the 12 systems studied, whereas GC-MSMD only failed on two. In four cases (T4L99A, AR, MUP1, GPB) GC-MSMD majorly outperforms MSMD simulations with more subtle improvements seen in another two systems (ERα, PPARγ) where ligands were better mapped in GC-MSMD but at the same occupancy as MSMD. For the three systems where the mapping was the same, GC-MSMD returned marginally higher occupancies and in some cases fewer false positives (HEWL, p53, ERK2). The final three systems (HSP90, PDK1, PTP1B) showed little difference between the methods.

7.4.9 Free Energy Grids

7.4.9.1 T4L99A

Simulations and Analysis by VNM. Method, Text and Figures by WP.

To assess the performance of GCNCMC-based free energy grids, we performed GC-MSMD simulations for the same set of T4L99A molecules as in the previous chapter. The resulting free energy grids can be visualised in the same way as the occupancy grids by contouring at a certain free energy, such that the grid points that are shown will all have a free energy value equal to, or less than, the contour value. Alternatively, the grid can be viewed as a surface and coloured by free energy, as in Figure 7.17, clearly showing a region with high affinity in the T4L99A binding site.

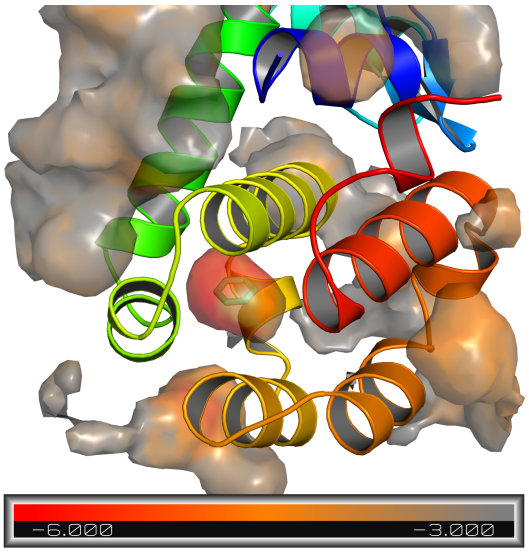


FIGURE 7.17: Free energy surface of T4L99A derived from GC-MSMD simulations with benzene as the probe. The crystal pose for benzene is shown in green (pdb: 181l). A clear region of high affinity is shown where the surface transitions from grey to red.

We performed this analysis for all the previously studied ligands and recorded their free energies. For now, the free energies were recorded by eye as the contour value at which the grid disappears or fails to map the crystal pose significantly. A more rigorous approach will be explored in the future. Figure 7.18 details the free energy estimates in comparison to the more expensive GCNCMC titrations and FEP. Given the approximate nature of this method, a remarkable agreement between the methods is observed. This serves as a good proof of concept and should be explored further in the future.

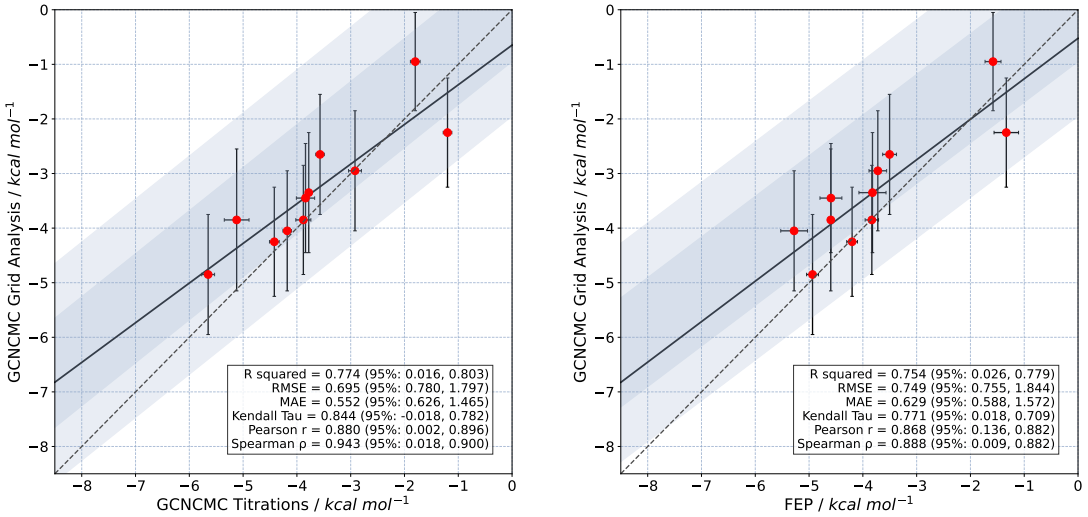


FIGURE 7.18: Calculated binding free energies for T4L99A from GC-MSMD simulations vs. GCNCMC titrations with a small GCMC sphere (Chapter 6), and FEP results.

7.4.9.2 MUP1

Simulations and Analysis by VNM. Text and Figures by WP.

The GC-MSMD free energy grids were also tested on MUP1 but in a slightly different fashion. For T4L99A, the grids were tested using actual fragments such that they could be directly compared to free energy estimates derived from GCNCMC titrations and FEP. For MUP1, we look at the four basic cosolvent probes to see if they can achieve similar mapping to the occupancy maps, while also identifying regions of high affinity. In this case, there is no direct free energy data to compare to.

Figure 7.19 shows the results for isopropanol and pyrimidine. As expected, the mapping is similar to the raw occupancy maps but now with information on how tightly a probe binds at a particular grid point. In an SBDD context, this visualisation could help medicinal chemists understand, and potentially exploit, regions of high affinity.

In principle, as demonstrated by the SILCS methodology,[91] a somewhat effective method to calculate the free energy of molecules which may bind in a given pocket is to sum the free energies of the overlapping probe grid points. This "Ligand Grid Free Energy" method requires grids of probes with different chemical moieties. As an example, in the case of pentanol, one could add up the free energy of a carbon based grid for the 5 carbon atoms and a hydrogen bond donor grid for the alcohol group to estimate the total free energy of pentanol. In the current implementation, the GC-MSMD free energy grids are for whole probes and cannot be broken down further into specific atoms. Despite this, the MUP1-isopropanol combination makes for a good example to estimate the free energies of the five alcohols tested previously (Chapter 6). Using the example in Figure 7.19, we can estimate the binding affinity of the propanol molecule in 1znd as -8.5 kcal mol^{-1} by summing the middle and third column where the probe map aligns well with the ligand. This is in good agreement with the value calculated by titration in Chapter 6 (-8.0 kcal mol^{-1}). Of course, this method is very subjective and is included as a proof of concept, SILCS uses a more rigorous protocol and should be explored further in the context of GCNCMC free energy grids.

For the higher order alcohols, hexanol (1zne), heptanol (1zng) and octanol (1znh) we estimate free energies of -9.2, -9.9 and -11.6 kcal mol^{-1} respectively. This is in remarkably good agreement with the titration calculations which predicted -9.1, -10.2 and -11.0 kcal mol^{-1} respectively. These results are promising but require more robust validation before these GCNCMC free energy grids can be used to predict the affinities of small molecules reliably. Again, it should be noted that these values were derived by eye and it will be worthwhile in the future to test this method on the same systems and protocols as SILCS to calculate the free energy of larger, drug-like molecules from grids.

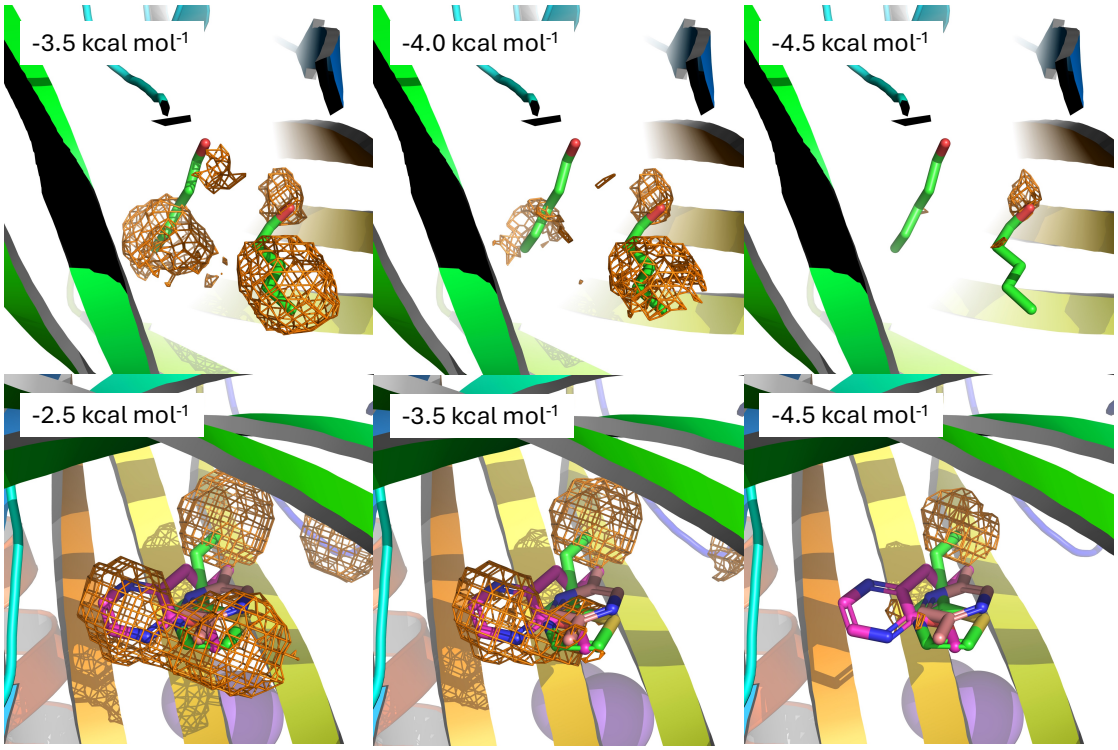


FIGURE 7.19: GC-MSMD free energy grids for MUP1. Top: Isopropanol grids with PDB 1znd overlaid. Bottom: Pyrimidine grids with PDBs 1i06, 1qy1 and 3kfi overlaid. In each case, good mapping of the relevant interactions is observed and by decreasing the free energy contour level, the strongest interactions remain.

7.5 Summary

In this chapter, we have applied GCNCMC in the context of mixed solvent MD. MSMD in itself is a sampling protocol which exploits high concentrations of organic probes to map potential binding sites on biological targets. A longstanding issue with such simulations is the ability to sample occluded and cryptic pockets owing to the timescales required to either diffuse into occluded pockets or induce a conformational change to open cryptic sites. Given the success in sampling the occluded binding sites of T4L99A and MUP1 in the previous chapter, it was believed that GCNCMC may enhance MSMD simulations in systems with occluded binding sites. Indeed, the study demonstrated that with the addition of GCNCMC, through its enhanced sampling capabilities, improved pocket mapping was observed in almost all systems, particularly those with occluded pockets.

A novel grid-based free energy analysis was also presented combining sampling from GCNCMC with the Bennett acceptance ratio to estimate the affinity of grid points using nonequilibrium work measurements. We tested the approach on the T4L99A system from a previous chapter by using known ligands with free energy data as our cosolvent probes. This allowed for a direct comparison of the estimates derived from the free energy grids to FEP and GCNCMC titration calculations. We then studied the

free energy grids of the MUP1-isopropanol combination and employed a method similar to SILCS which sums up the free energies of grid points that overlap with larger, bound ligands. In this case, a remarkable agreement with our titration calculations was observed. However, this will need a much more robust validation and comparison to other methods. Of particular interest is seeing how our results compare to the results published by SILCS if the same protocols were followed. Overall, this part of the study has provided a proof of concept which should be explored in the future. Taking inspiration from other grid-based analysis methods such as the Grid Inhomogeneous Solvation Theory and SILCS with all be beneficial.[254,270,271]

Cryptic binding sites remain challenging with both methods failing to capture the large conformational changes required for binding (e.g. PTB1B, HSP90, PDK1, ERK2). This highlights the need for complementary techniques; one interesting avenue for exploration is to combine the use of accelerated, or Gaussian-accelerated, MD with GCNCMC.[96,257,272,273] It is thought that aMD could induce the conformational changes required to sample these cryptic sites, subsequently improving the GCNCMC sampling in these regions, and allowing changes to be quickly stabilised. Developing a reweighting protocol will further set this method apart from other studies.

A final consideration is the fact that these simulations use GCMC spheres which encompass the whole protein. As this particular application of GCNCMC is designed to find unknown binding sites, we wanted to assume no prior knowledge of the protein in question. The result however is relatively large spheres which require a high number of GCNCMC moves to sample extensively. In a structure-based campaign, potential binding regions may already be known, meaning that the size of the sphere can be adjusted to cover only these regions and greatly focus the sampling on that particular region, improving the quality of the mapping.

Overall the integration of GCNCMC into MSMD represents yet another way in which GCNCMC can be exploited to enhance a more traditional structure-based method, enabling better identification and characterisation of putative binding sites.

Chapter 8

Fragment Screening using Grand Canonical Nonequilibrium Candidate Monte Carlo

8.1 Introduction

A final application of GCNCMC has been explored in the context of fragment screening and hit identification. It has already been shown how GCNCMC can identify potential binding sites and calculate ligand binding affinities in a particular site. The latter application can be expensive if many B values are simulated and is useful when the user wants a complete understanding of the ligand binding. In this chapter, we present a cheaper alternative which can used to discriminate between binders and non-binders at a user defined concentration. As an extension, despite being cheaper than titrations, we show that accurate free energy estimates are still achievable within this protocol. For this application, we recall the equation for ligand efficiency (LE) discussed in Chapter 1.3:

$$LE = -\frac{\Delta G^{\ominus}}{HAC} \tag{8.1}$$

which can be re-written in terms of a concentration, K_D:

$$LE = -\frac{k_B T \ln K_D}{HAC} \tag{8.2}$$

where HAC is the number of heavy atoms for a particular ligand.

It is now possible to calculate the value of K_D required for a given ligand, with N heavy atoms, to achieve a desired ligand efficiency. It was also shown in Chapter 5 that K_D is equal to the ligand concentration, in solvent, required to maintain a binding

site occupancy of 0.5. As such, GCNCMC simulations can be performed at this concentration, and any ligand which converges on an average occupancy of greater than 0.5 can be deemed a hit. In other words, a ligand with an average occupancy of greater than 0.5 indicates that the ligand's true dissociation constant is lower than the value specified and thus has a stronger binding affinity. Therefore, a hit can be thought of as a ligand which has a ligand efficiency greater than that specified by the user. If desired, more sets of simulations can be run at different concentrations to find only the strongest binders. For this test, we have chosen ligand efficiency as the target parameter owing to its prevalent use in FBDD but other suggestions may include the solubility of the ligand, or concentrations lower than a desired binding affinity.

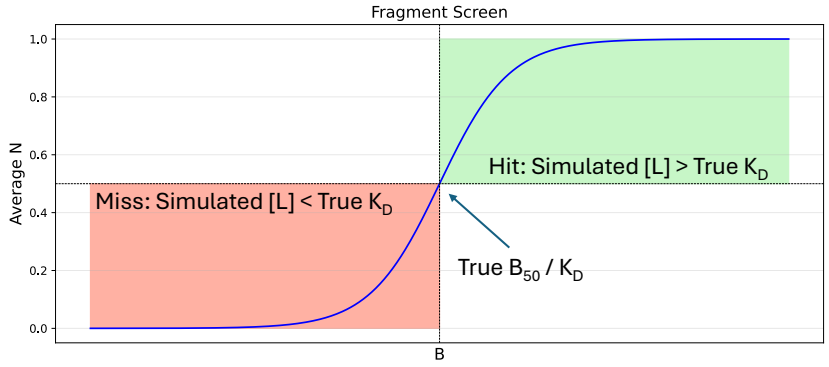


FIGURE 8.1: Overview of the GCNCMC fragment screen. If the user-defined simulation concentration is greater than the true dissociation constant then an average occupancy greater than 0.5 is expected and deemed a hit (green quadrant). When the average occupancy is less than 0.5, the simulated concentration must be lower than the true dissociation constant and can deemed a miss (red quadrant).

As a proof of concept, we use another mutant of T4-lysozyme - T4L99A/M102Q. This mutant of T4 makes the binding site more polar by introducing a glutamine residue along one edge of the site allowing for hydrogen bonding.[128,242] This mutant binds a wider variety of ligands compared to its apolar version. Two ligand efficiencies were tested - 0.3 and 0.55. The former was chosen as it is usually taken to be the minimum required to progress a fragment hit, and the latter was selected after manually inspecting published computational binding affinities of the ligands to artificially generate more misses. Note, for this study we are comparing against published FEP data from Boyce *et al.*[242] rather than experimental data. This is because some of the ligands studied were shown to not bind at all owing to the lack of sensitivity of the experimental method. Using FEP data means that each ligand has a direct value to compare against, allowing for a more direct validation of the method.

8.2 Simulation Details

8.2.1 System Setups

The apo structure of T4L99A/M102Q (PDB: 1lgu) was protonated according to a pH of 7.2 using PDBFixer.[247] Missing loops were added where appropriate, and protein termini were capped using N-methyl and acetyl caps. Each system was then solvated in a box of TIP3P[221] water with a buffer of 12 Å around the protein. Na+ and Cl- ions were added to neutralize the system and up to a salt concentration of 0.15 M. All co-crystal additives were removed, as well as crystallographic waters. The resulting system was then equilibrated for 5 ns in the NVT ensemble at 298 K and then a further 15 ns in the NPT ensemble at 298 K and a pressure of 1 Bar using a Monte Carlo barostat.

Fragment screening simulations were performed using OpenMM 8.1 with the *grandlig* python module. Simulations were performed at 298 K and all MD was performed using the Langevin BAOAB integrator with a friction coefficient of 1 ps^{-1} and a time step of 4 fs with hydrogen mass partitioning (Hydrogen mass = 2 Da). The cut-off for nonbonded interactions was 12 Å with a switching function applied at 10 Å for the Lennard-Jones interactions. Particle mesh Ewald (PME)[183] was used to calculate the effect of the long-range electrostatics. The protein was modelled using the AMBER ff14SB forcefield.[220] All simulations use TIP3P[221] waters and all ligands (Fig 8.2) are parameterized using the openff-2.2 small molecule forcefield[156] with AM1-BCC charges.[223] Ions, wherever present, were modelled with Joung-Cheatham parameters.[249]

Screening simulations were then initiated with a further 2 ns of NVT equilibration followed by 150 GCNCMC moves in the μVT ensemble. For production, 500 ligand GCNCMC moves were performed per repeat with a move performed every 500 ps. Each GCNCMC move has a switching time of 50 ps and was performed using the flat bottom restraint scheme described in Chapter 9. The total production time therefore is 275 ns with four repeats per ligand. The GCMC sphere was centred between the C-alpha atoms of Leu84 and Ala99 with a radius of 8 Å to cover just the known binding site. The average occupancies at the end of the simulations are used for analysis. The excess chemical potential of the ligands was calculated as described previously (Sec. 4.3.1). The concentrations for each ligand were determined by calculating the dissociation constant each ligand would require to achieve the desired ligand efficiency and is therefore based upon its heavy atom count (HAC) (Eq. 8.3). A full list of simulation parameters are given in Table 8.1.

$$K_D = \exp \frac{(LE)(HAC)}{-k_B T} \tag{8.3}$$

TABLE 8.1:

LE=X$refers to minimum free energy and K_D$ required for a ligand efficiency of at least X. $[L]_{LE=X}$ details the concentrations at which simulations were performed. Ligands are sorted from the highest to lowest ligand efficiencies calculated from the FEP results of Boyce *et al.*]T4L99A/M102Q Simulation parameters. $\Delta G^{\ominus}$ and μ' values are in units of kcal mol^{-1} and concentration values are in units of mM. $\Delta G^{\ominus}_{LE=X}$ and $[L]_{LE=X}$ refers to minimum free energy and K_D required for a ligand efficiency of at least X. $[L]_{LE=X}$ details the concentrations at which simulations were performed. Ligands are sorted from the highest to lowest ligand efficiencies calculated from the FEP results of Boyce *et al.*[242] Values for the excess chemical potential, μ', are calculated as described in Section 4.3.1.

ID	$\Delta G^{\ominus}_{comp}$	LE_{comp}	HAC	μ'	$\Delta G^{\ominus}_{LE=0.3}$	$\Delta G^{\ominus}_{LE=0.55}$	$[L]_{LE=0.3}$	$[L]_{LE=0.55}$
22	-5.61	0.94	6	-6.02	-1.80	-3.30	47.9	3.80
21	-6.81	0.85	8	-6.06	-2.40	-4.40	17.4	0.59
18	-6.16	0.77	8	-5.75	-2.40	-4.40	17.4	0.59
25	-5.43	0.68	8	-1.83	-2.40	-4.40	17.4	0.59
26	-5.27	0.66	8	-2.65	-2.40	-4.40	17.4	0.59
24	-5.26	0.66	8	-1.68	-2.40	-4.40	17.4	0.59
02	-4.39	0.63	7	-5.94	-2.10	-3.85	28.8	1.50
13	-4.72	0.59	8	-4.22	-2.40	-4.40	17.4	0.59
07	-5.82	0.58	10	-7.36	-3.00	-5.50	6.31	0.09
14	-3.28	0.55	6	-3.54	-1.80	-3.30	47.9	3.80
15	-4.63	0.51	9	-3.94	-2.70	-4.95	10.5	0.23
08	-4.07	0.51	8	-5.54	-2.40	-4.40	17.4	0.59
23	-4.53	0.50	9	-2.29	-2.70	-4.95	10.5	0.23
01	-3.47	0.50	7	-0.46	-2.10	-3.85	28.8	1.50
12	-3.93	0.49	8	-10.2	-2.40	-4.40	17.4	0.59
09	-3.84	0.43	9	-5.00	-2.70	-4.95	10.5	0.23
10	-3.39	0.42	8	-3.58	-2.40	-4.40	17.4	0.59
04	-4.16	0.42	10	-4.84	-3.00	-5.50	6.31	0.09
03	-3.14	0.39	8	-9.52	-2.40	-4.40	17.4	0.59
06	-3.02	0.34	9	-4.46	-2.70	-4.95	10.5	0.23
20	-2.52	0.32	8	-7.50	-2.40	-4.40	17.4	0.59
19	-1.86	0.23	8	-8.61	-2.40	-4.40	17.4	0.59
17	-1.91	0.21	9	-5.60	-2.70	-4.95	10.5	0.23
05	-1.31	0.12	11	-5.51	-3.30	-6.05	3.80	0.04
16	0.21	0.02	11	-16.3	-3.30	-6.05	3.80	0.04

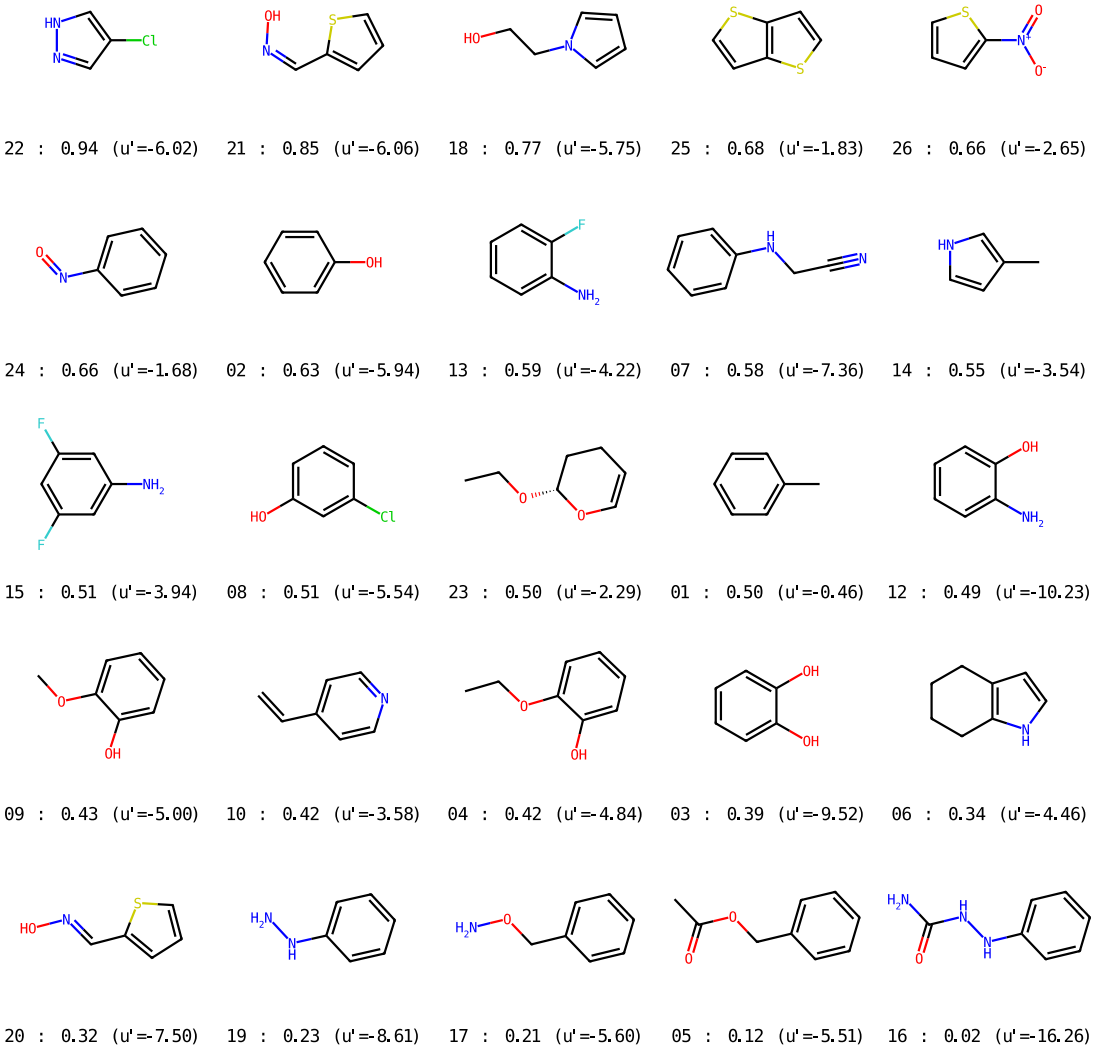


FIGURE 8.2: T4L99A/M012Q ligands used for the GCNCMC fragment screen. The fragments are ordered by their ligand efficiency calculated from computational free energies published by Boyce *et al.* Calculated excess chemical potentials are given in brackets.

8.3 Results and Discussion

Figure 8.3 shows the results of the screen. Blue bars represent the final average occupancy from simulations performed at an LE of 0.3 and orange bars represent LE=0.55. The dashed green line represents an occupancy of 50%; a ligand can be deemed a hit at a given LE if the bar is above this line. The purple dividing lines represent the two ligand efficiencies such that any ligand to the right of the line should be a miss, the left line represents LE=0.55 and the rightmost represents 0.3. Immediately it is clear that increasing the LE to 0.55 from 0.3 unsurprisingly results in

more ligands dropping out, leaving only the strongest binders. This is highlighted by the number of orange bars falling under the green line after the first dividing barrier.

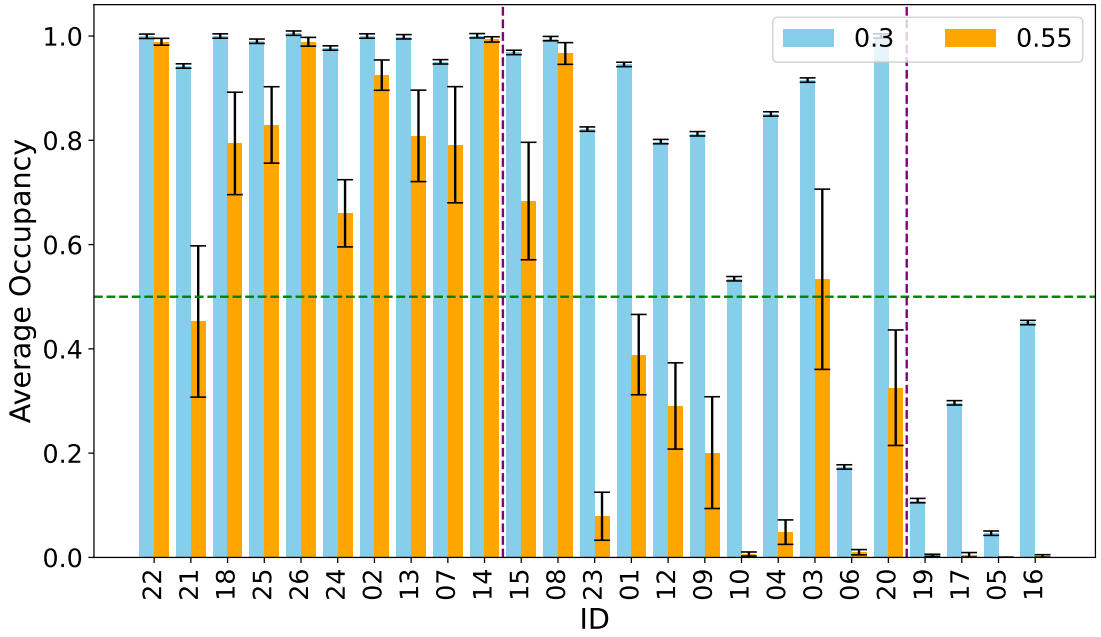


FIGURE 8.3: GCNCMC based fragment screen of small fragments to T4L99A/M102Q. Compounds are labelled as in Figure 8.2 and are ordered from left to right in decreasing affinity. The first purple line indicates where LE=0.55 sits, meaning all orange bars, to the right of this line, should be under the green dotted line (0.5). The second purple line indicates LE=0.3 where all orange and blue lines to the right should now sit below the green line.

The statistics (Table 8.2) for both ligand efficiencies show that using 0.3 returns a slightly better accuracy (total number of correct predictions over the total number of fragments) (0.96 vs 0.80) with a larger true positive rate (TPR) and lower false positive rate (FPR). This is likely due to compounds 15 and 08 being classified as false positives at LE=0.55. Both have ligand efficiencies of 0.51, which is close to the threshold value and the discrepancy may be due to forcefield inaccuracies.

TABLE 8.2: T4L99A/M102Q Fragment Screen Statistics. Sensitivity, also referred to as the true positive rate (TPR), is calculated by the number of true positives (TP) divided by the number of expected positives (P). Specificity, or true negative rate (TNR), is the number of true negatives (TN) divided by the number of expected negatives (N). FPR=1-TNR, Precision=TP/TP+FP, FNR=FN/FN+TP and Accuracy=TP+TN/P+N

LE	Sensitivity (TPR)	Specificity (TNR)	FPR	Precision	FNR	Accuracy
0.30	0.95	1.00	0.00	1.00	0.05	0.96
0.55	0.89	0.75	0.25	0.67	0.11	0.80

Finally, as each simulation performed generates nonequilibrium work values, we can combine these data to predict free energies using either BAR or GAR (Sec. 5.2.2, Eq.

5.28). The results of this are shown in Figure 8.4 and compared to FEP results published by Boyce *et al.* in Figure 8.5.[242] In general, the free energy estimates from the fragment screen are in good agreement with the published FEP results, however some large outliers do exist. As in the GC-MSMD simulations, these non-equilibrium work measurements are a byproduct of the screening simulation meaning the added benefit of calculating fairly accurate affinities quickly, and simply, is incredibly useful. In the future, it would be best to calculate these binding affinities via FEP using the *grandlig* code and with the same parameters as the screen, to prevent any potential discrepancies with the published data arising from the forcefield, sampling or other aspects of the simulation protocol. The source of outliers should also be explored further; it may be that the work distributions require more data. Figure 8.6 shows that the errors in the calculated values with respect to the FEP data are normally distributed implying random noise.

Speculatively, errors may arise owing to a lack of insertion works and biased deletion works in the tight binder regime and a lack of deletion works in the weak regime. The former arises as each deletion move is derived from similar starting poses, or in other words, each deletion move, many of which are rejected, starts from the same binding pose biasing the distribution, and if most of these moves are rejected, then there is little opportunity for further insertions. The latter arises from the fact that not many insertion moves are accepted meaning there is often no molecule to delete. The issue in both cases could be addressed by running the screen with more concentrations, although this approach closely resembles titration experiments and is, therefore, more computationally demanding and loses the advantages associated with a rapid screening methodology.

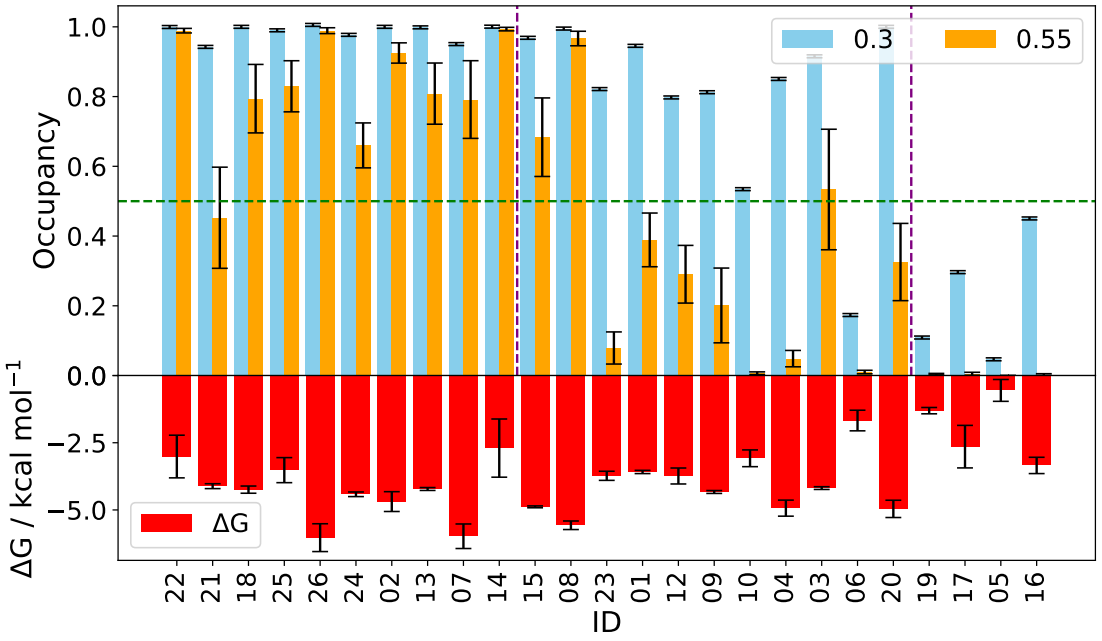


FIGURE 8.4: T4L99A/M102Q fragment screen results with calculated free energy values derived using the measured nonequilibrium works and GAR.

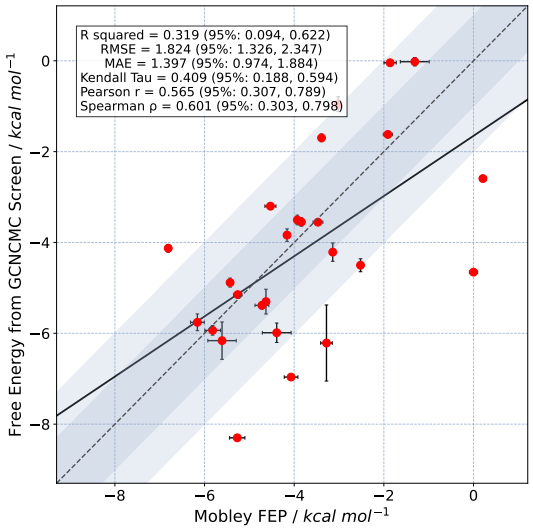


FIGURE 8.5: Calculated free energy values from non-equilibrium works using GAR compared to published FEP results by Boyce *et al.*[242]

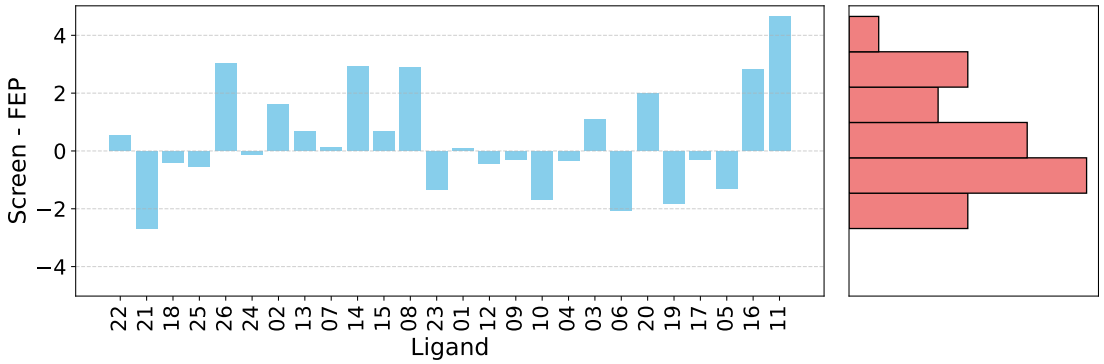


FIGURE 8.6: T4L99A/M102Q free energy errors between free energies derived from the GCNCMC screen and the published FEP results.

8.4 Summary

In this chapter, we have presented another proof of concept study describing a different application of GCNCMC in the context of fragment screening. We show, using a model system, this protocol can predict molecules as hits or misses, using a single GCNCMC simulation, based on the input ligand concentration in bulk solution. Setting this concentration to a useful metric in FBDD such as ligand efficiency, or a minimum desired concentration, provides an effective method of filtering a pool of ligands. In contrast to GCNCMC titration calculations, this protocol does not require many simulations at multiple B values. For T4L99A/M102Q at two different ligand efficiencies, we predict hits and misses with high accuracy compared with published FEP data.

We then extend the method to use the nonequilibrium works recorded throughout the simulations to estimate each ligand's binding affinity. We observe a moderate correlation against the published FEP data with many data points falling within 1-2 kcal mol^{-1} (48% and 79%) of the published values. However, as mentioned, it is pertinent that FEP calculations should be performed in-house using the same code and parameters as the screen. It is worth re-mentioning that these works are a by-product of the GCNCMC simulations and using them in this way provides an orthogonal method at no extra cost. Future studies should focus on refining this part of the protocol and finding other ways to use the work distributions to predict affinities more accurately.

Finally, this protocol should be tested with a larger pool of protein systems, particularly those with biological relevance. However, issues outlined in the following chapter should be addressed first to make this method successful in more challenging protein systems.

Chapter 9

Identification of Sampling Challenges in GCNCMC

9.1 Introduction

Throughout this thesis, several sampling and efficiency issues have been encountered. These limitations are typical of alchemical based methods and are indeed present in other free energy methods. The initial goal of expanding the applicability of GCNCMC to small molecules was to improve the sampling of binding in occluded pockets, in much the same way as for water. As noted here and elsewhere, alchemical free energy calculations are hampered by many sampling challenges.[133,148] Some of these issues - notably knowledge and treatment of multiple binding modes - are addressed by GCNCMC. However, other sampling challenges remain - in particular, when the alchemical transformation (in this case, fragment insertion/deletion) requires the concomitant binding or displacement of solvent molecules or protein conformational change. In this chapter we explain some of these sampling challenges and discuss, with preliminary simulations, potential avenues for further development to address these issues. The results presented here are preliminary and reflect proof of concept only.

9.2 Molecules Leaving the GCMC Sphere

The first issue is that molecules residing outside the GCMC sphere at the end of the move must be automatically rejected since the reverse move cannot be proposed, breaking the condition for detailed balance. This can lead to a high proportion of moves being rejected, impacting the overall efficiency and convergence of the GCNCMC simulations, as exemplified in Chapter 4 by the host-guest system. This

typically happens when the binding site is exposed to solvent or is occupied by water molecules or protein side chains. In these cases, when the ligand is weakly interacting, the lowest free energy pathway for the ligand is to simply diffuse into the solvent rather than compete to bind. This problem is exacerbated at longer switching times where the ligand is weakly interacting for longer. Each specific scenario is discussed in further depth below. A naive, but somewhat effective, approach to prevent unbinding is to apply a weak flat-bottom restraint to the GCMC sphere such that the switching molecule cannot leave. We test this approach on the same host-guest system as before.

While the use of restraints breaks the 'no restraint needed' design philosophy of GCNCMC, the use of a weak flat bottom restraint is easy to describe and should have a minimal impact on the fully bound. Furthermore, it has the added benefit of preventing the sampling of irrelevant configurational space by keeping the ligand in the desired pocket while still allowing for all potential binding modes to be sampled. In this implementation, molecules which reside outside of the sphere at the end of the move are still automatically rejected, even with a restraint. As such, the correction for the flat bottom restraint takes the same form as the standard state correction already included in the Adams value, specifically:

$$B_{eq}(c) = \beta\mu'_{sol} + \ln\left(\frac{V_{GCMC}}{V(c)}\right) \tag{9.1}$$

As in traditional free energy calculations, the effect of imposing the restraints onto a fully interacting ligand must also be accounted for, and for a GCNCMC calculation, this would have to be done by calculating the work done by turning on the restraint either at the end of an insertion move or at the start of a deletion move. However, often with flat bottom restraints, this effect is rather minimal when the ligand is bound since the bound pose is often stable and therefore does not test the walls of the restraint. Currently, we choose to not account for this in the acceptance criteria. To test if adding a restraint is a valid approach, we perform the same simulations as in Chapter 4 using a flat bottom restraint and compare the work distributions to the unrestrained simulations.

As expected, Figure 9.1 (top) shows that the average occupancy across all switching times is 0.5. This confirms that the use of a flat bottom restraint ensures that most of the attempted GCNCMC moves remain valid by preventing the ligand from leaving the GCMC region. In fact, at $\tau = 500$ ps, the proportion of moves rejected due to leaving the region was 20% (not shown) which is a big improvement on the 95% seen previously. As such, the addition of the restraint means that fewer moves are wasted through automatic rejection thus improving computational efficiency.

Applying the acceptance criteria (Fig. 9.1 middle) to the work distributions post-simulation results in a more discernible trend in the average occupancies. It is

clear that with increasing switching time, the average N in the simulation converges to one value. This trend is also roughly in line with the convergence of the mean insert and delete works (bottom left), in that as the work distributions converge, so does the value of average N.

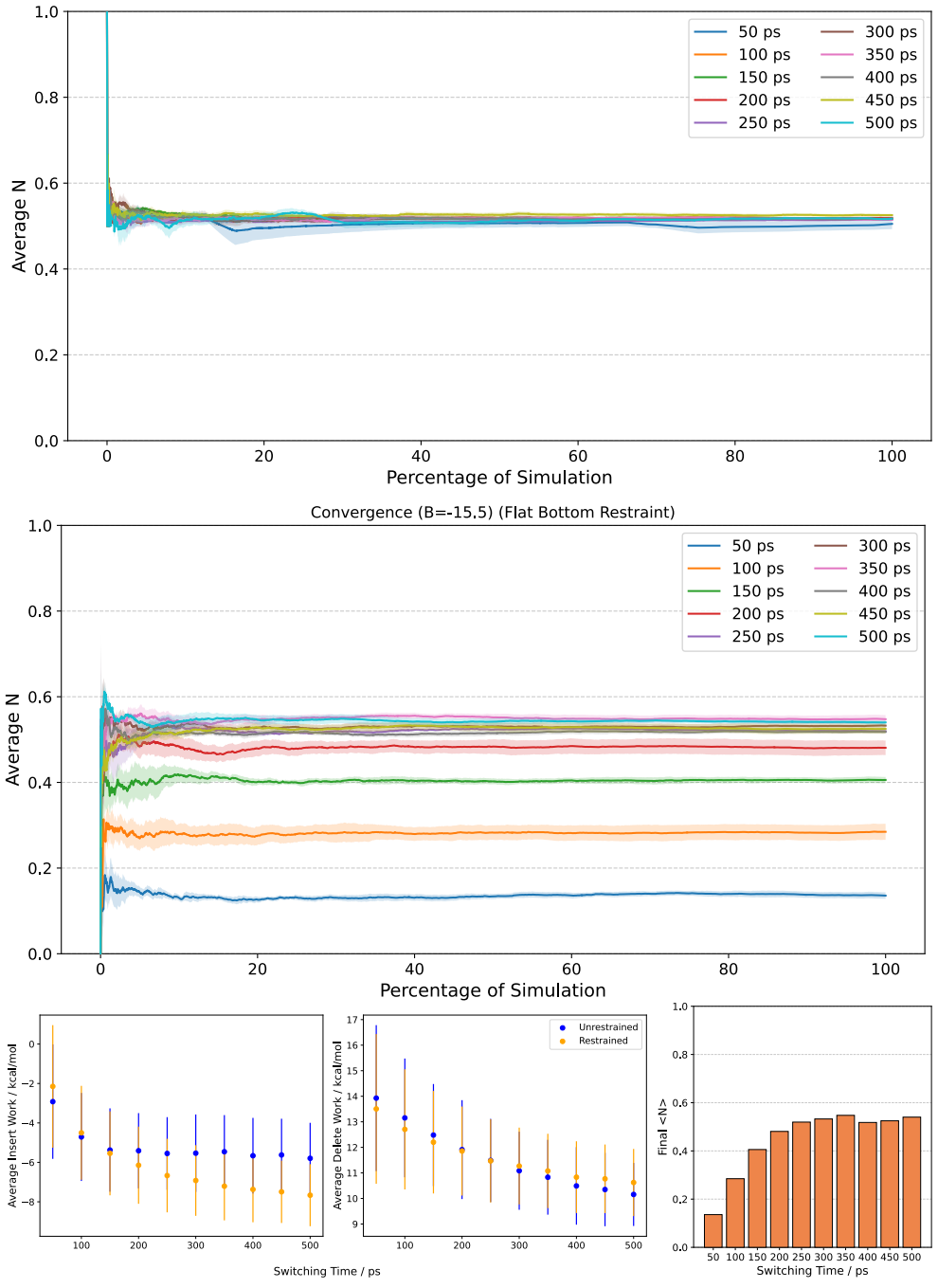


FIGURE 9.1: Top Row: Convergence of the GCNCMC simulations at different switching times using a flat bottom restraint. Second Row: Convergence of simulations after applying the acceptance criteria with a B value of -15.5. Bottom Row: Comparison of the nonequilibrium works measured between unrestrained and restrained simulations.

It should be mentioned that the use of the restraint results in a clear deviation in the mean insertion works compared to an unrestrained simulation, particularly at higher switching times. Interestingly, the values are shifted to more negative work values, implying that using a restraint leads to more favourable moves. We believe that this may be in part due to the ligand being kept bound for longer, by the restraint, allowing it to freely adopt more favourable conformations. Furthermore, the use of a restraint reduces the number of unbinding and rebinding events and restricts the dynamics of the molecule being coupled or decoupled to the relevant configurational space. In other words, the restraint ensures that only one binding pathway is sampled. The free energy of transfer as a function of switching time is shown in Figure 9.2 which shows the free energy converges at higher switching times, again, in line with the convergence of the work distributions and in the average N. This is in massive contrast to the same plot in Chapter 4 (Fig. 4.12) which seems to show no convergence likely owing to the nonequilibrium moves sampling irrelevant regions of configurational space.

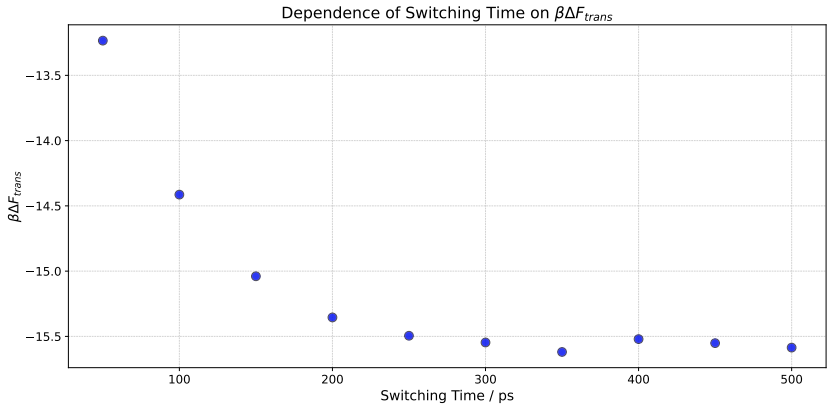


FIGURE 9.2: Effect of switching time on the free energy of transfer from gas to complex.

Lastly, it is also possible that this observation is simply a byproduct of the fact that there are many more measurements available in the restrained simulations and therefore the mean value is better converged. Without further testing, it is difficult to say if the more favourable nonequilibrium works are correct or not.

In summary, applying a flat bottom restraint to the molecule being switched over the course of insertion of deletion move results in a higher probability that the move will be valid by preventing the molecule from diffusing away. As a byproduct, it may also facilitate higher quality moves by limiting the sampling to one reaction path. Future work should investigate the effects of this restraint and evaluate if it should be used routinely in GCNCMC simulations. One other avenue for exploration is to possibly decouple, or couple, the restraint at the same rate the ligand is being switched on or off.

9.3 Water Molecules

Water molecules, which can play a vital role in ligand binding dynamics, are often displaced when a ligand binds.[146,147] In other words, most binding sites are occupied by water molecules which are exchanged by a ligand upon binding. This is problematic in simulations as these water molecules can sometimes be tightly bound, or at least the interactions between water and the protein are stronger relative to a weakly interacting ligand. This means that waters which may be displaceable in principle are unable to be displaced in practice by GCNCMC insertion moves, as the lowest free energy pathway results in the ligand unbinding rather than displacing the water. The issue can be further sub-categorized into water molecules bound in solvent exposed pockets, and those bound in occluded pockets. These cases are further discussed below.

9.3.1 Occluded Pockets

Waters which are bound in occluded pockets present a unique problem for alchemical methods, including GCNCMC moves. These water molecules, which may be displaceable by a ligand, simply do not have an escape passage, or at least the timescales of the water molecule leaving the pocket are far greater than what we simulate. Ironically, this is the exact issue water-based GCMC was designed to address.[147,150]

In the context of a ligand GCNCMC insertion move, as the ligand is turned on in the pocket, one of two things is likely to happen. In scenario one, the bound water is so tightly bound that the ligand has no choice but to diffuse into the solvent while in its low interacting states. In scenario two, the ligand and water both exist in the pocket as neither of them can escape, resulting in an extremely high work measurement as both entities' repulsion terms become large. In both cases, it is clear that a rigorous method of encouraging the removal of water is required.

A nice illustrative toy system for this scenario is a buckminsterfullerene, or buckyball (C60). C60 has a fused ring structure with a small internal cavity which can accommodate very small organic molecules such as water and methanol. However, diffusing molecules into and out of this cage structure is physically impossible without first opening the cage through a series of chemical reactions.[274]

In our simulations, we can place a water molecule into the centre of the ball and attempt to insert methanol molecules into the cage. As the water cannot escape, we expect to see high work measurements compared to inserting a molecule into the cage without the water. We then compare the insertion works of inserting into an empty and full cage. Then, in a third test we will apply water GCMC moves throughout the

ligand insertion, providing a means for the water to be rigorously removed when the repulsive potential becomes too high. For these studies we are not subjecting the ligand insertion moves to the acceptance test, rather we are attempting to devise a protocol for removing the obstructing water molecule and recovering similar work values to the insertion into an empty site. Ligand insertion moves are performed over 150 ps ($n_{pert} = 1500$, $n_{prop} = 50$) with a single water GCMC move attempted at every ligand perturbation. The water GCMC moves are accepted or rejected according to its default μ' value, -6.09 kcal mol^{-1}. The data shown is the mean average of 500 insertion moves and the system is reset to its initial coordinates after each move.

Figures 9.3 and 9.4 show the results of this test. In the first, the cumulative insertion work curves are shown as a function of the number of perturbations for the move. As expected, inserting the methanol into a site already containing water leads to a very large work almost 10x that of inserting into an empty site. Encouragingly, combining the move with GCMC water moves almost recovers the original work value. The histogram compares the distributions of the empty site insertions and the GCMC site insertions showing that using GCMC does not fully recover the same distribution but rather a more broad distribution at a slightly higher work value. This is not unexpected as the stochastic nature of both MD and the water GCMC means that a wider range of work measurements is accessible. Figure 9.4 gives a slightly more in-depth understanding of the process, showing that GCMC water moves are indeed removing the water molecules as the ligand interactions are being switched on.

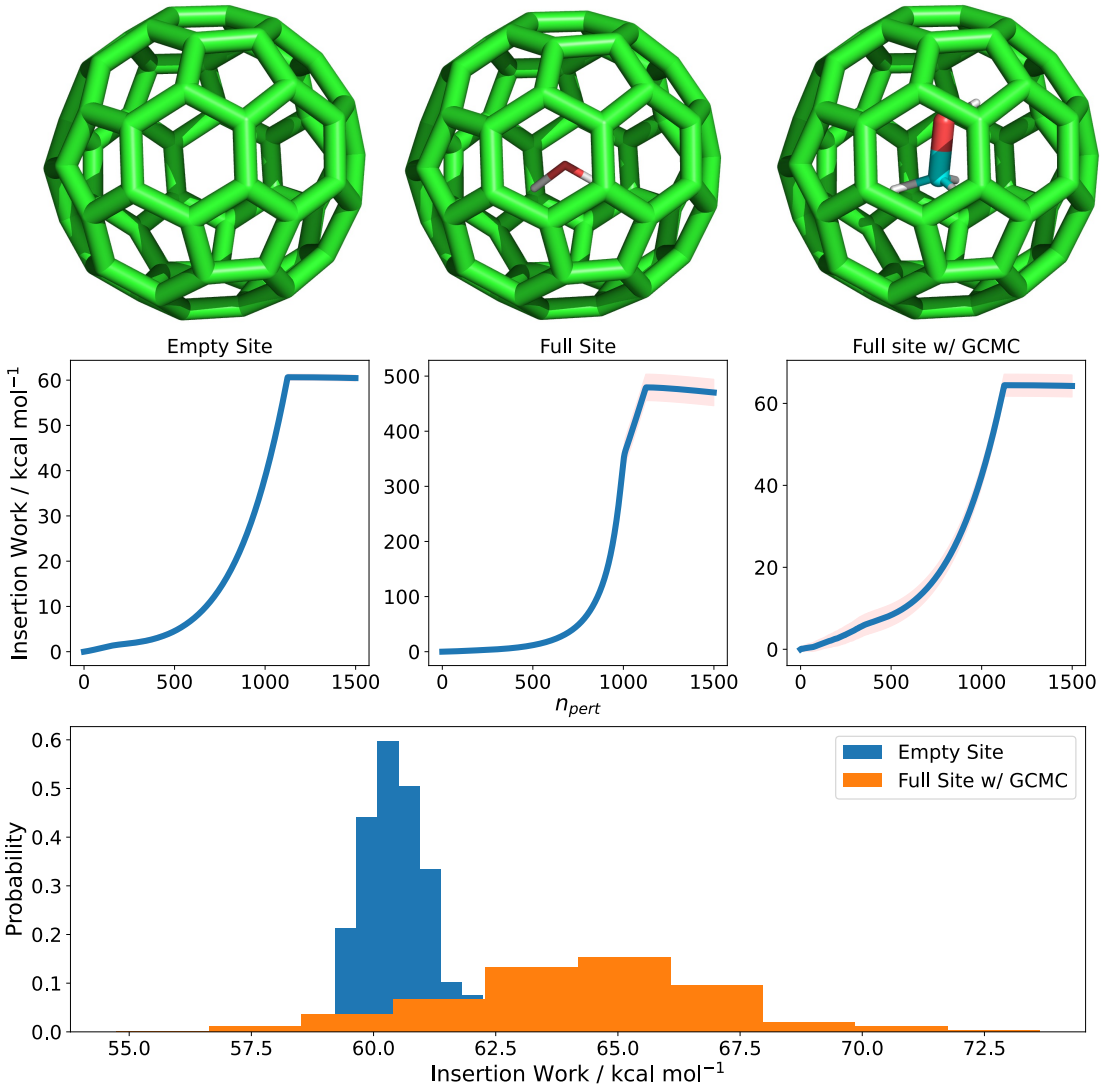


FIGURE 9.3: Comparison of insertion works for inserting a methanol molecule into C60. Inserting into C60 already containing a water leads to a large repulsive potential causing a large insertion work. Adding water GCMC to the protocol aids in removing water and recovering a more appropriate work measurement.

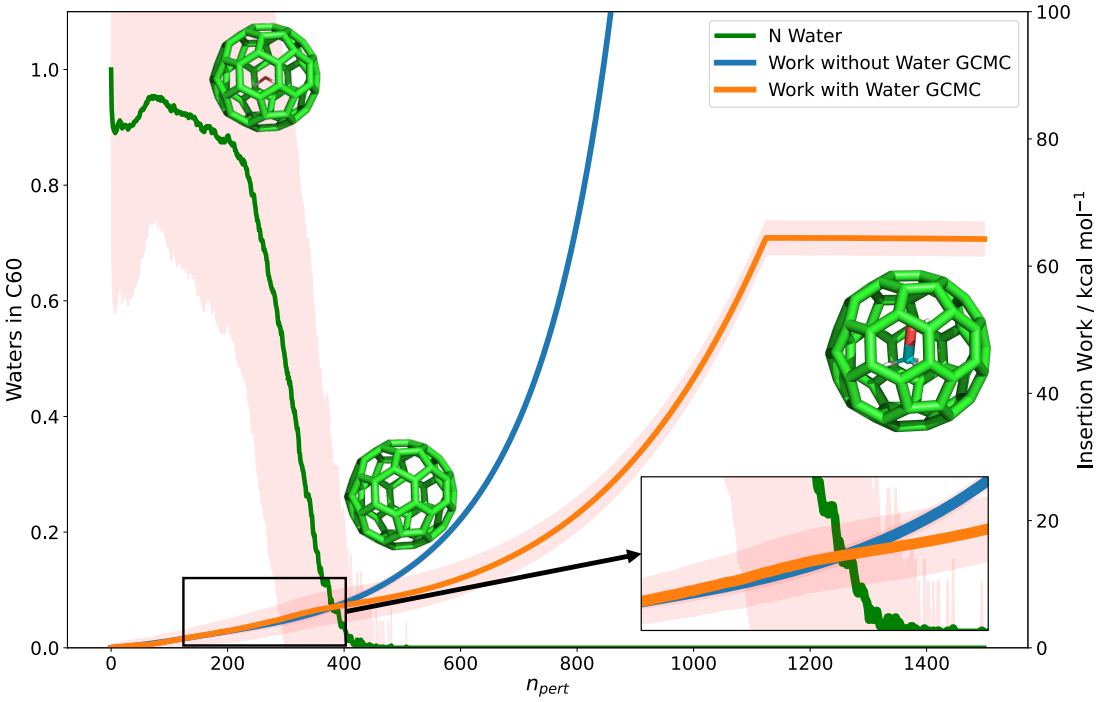


FIGURE 9.4: A more detailed view of the insertion moves into a hydrated C60 site. Using GCMC, the water molecule is quickly removed from the cavity as the ligand interactions are switched on. Inset: A zoomed in view of a potential inflection point where the ligand insertion works begin to deviate.

In another engineered system, we place an HSP90 ligand in its crystal pose in a system where several water molecules occupy the site, including in place of the ligand. This results in an overlap of the ligand and 3-4 water molecules. For this test, we do not run any MD integration, but rather we use a static structure giving a similar effect to an occluded pocket where neither the ligand nor water can diffuse away. The ligand interactions are then scaled (over 150 ps) to mimic an insertion move while simultaneously performing water GCMC moves at every perturbation. Owing to time constraints, the data presented is the average of 10 repeat moves only. The results of this test are shown in Figure 9.5 showing clearly that as the ligand is switched on, water molecules are easily removed by GCMC resulting in a lower work measurement.

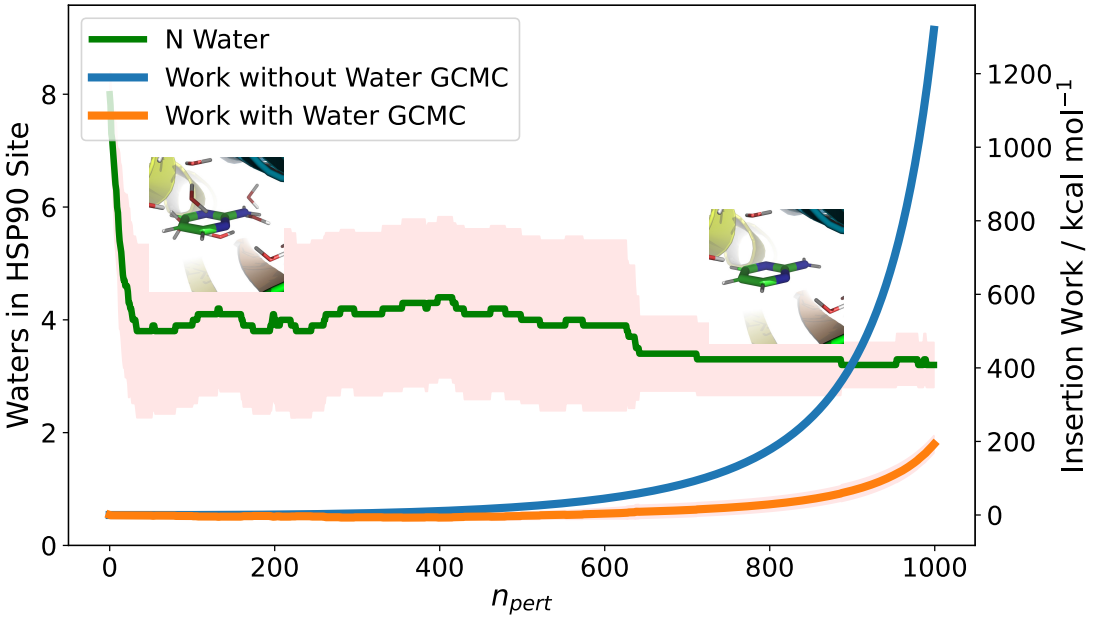


FIGURE 9.5: Detailed view of static insertion moves into HSP90. The ligand is initially placed in its crystal pose and dynamics are frozen. Using GCMC, water molecules are quickly removed from the cavity as the ligand interactions are switched on. The plotted data is the mean of ten repeats and the shaded regions represent the standard error of the mean.

These two proof of concept test systems show that water-based GCMC moves may be a useful complement to GCNCMC insertion and deletion moves in situations where the ligand may need to displace water, but the waters themselves are unable to escape.

However, this only solves part of the problem, as we first have to prevent the ligand from diffusing away in the early stages of the insertion move. In other words, using GCMC to remove water molecules is only useful when the ligand and water are competing for the same interactions. This is discussed further in the next section.

9.3.2 Exposed Pockets

Bound waters in exposed sites, in contrast to occluded sites, do in principle have means of escaping the pocket should the ligand succeed in displacing said waters. Chapter 4 showed that desolvation of pockets is possible, but only in a handful of moves compared to those which are automatically rejected. As mentioned before, the issues which arise in this scenario are associated with the balance between the interaction strengths of both molecules. Usually, it is easier for the weakly interacting ligand to diffuse away rather than compete with the bound waters. This, again, is exemplified by the host guest system, where we see a large proportion of moves, particularly at higher switching times, automatically rejected owing to the ligand leaving the GCMC region (65-95%).

It was shown in the previous section that provided the ligand and water molecules are overlapping and competing for the same interactions, then GCMC can be used to remove these waters. However, this scenario is much more common in occluded pockets as neither the water nor the ligand can escape, leading to unfavourable interactions. In exposed pockets, however, it is rare for this to happen as usually one of the two species will be forced to unbind. Within the alchemical framework, the fully interacting water will usually win, causing the ligand to diffuse away while in its low interacting state. Again, this is what is observed with the simple host-guest system where it was shown that valid moves do indeed remove water from the pocket, but more often than not, moves were rejected owing to leaving the GCMC sphere.

If it were possible to engineer the GCNCMC moves such that the ligand and bound waters had no choice but to compete for the binding interactions then it would make water-based GCMC moves more applicable to these exposed sites. Flat bottom restraints are one option which keeps the ligand in the binding site and may be sufficient in some cases. However, for relatively large pockets it is possible that the water could still win and push the ligand into an alternative sub-pocket or binding pose, or even force the ligand to sit at the edge of the restraint.

Another idea is to dissect the interactions of the ligand into those with only water and those with the rest of the system. In doing so, it allows the different interactions to be scaled independently making it possible to first turn on the ligand-protein interactions such that the ligand can find a stable binding pose without the influence of water molecules. At some point in the switch, the interactions with water can then be turned on such that any waters which may be overlapping at this point will either simply diffuse away by MD or be easily removed via GCMC. Figure 9.6 compares the current lambda schedule (1st row) with three proposed lambda schedules for this implementation. The first (2nd row), starts switching the ligand-protein VDW interactions for the first 25% of the move, with the ligand-water interactions then being turned on such that the ligand VDW interactions become fully on for all species at 75% of the move. In the second schedule (3rd row), the water interactions still start and finish at 25% and 75% respectively, but the ligand-protein VDW interactions are fully switched on at the halfway point of the move. This means that the ligand-protein VDW interactions are at full strength for part of the move. Finally, in the last option (4th row), the protein-ligand interactions are fully switched on before switching the ligand-water interactions. These lambda schedules will likely need further optimisation in the future and are included as a starting point.

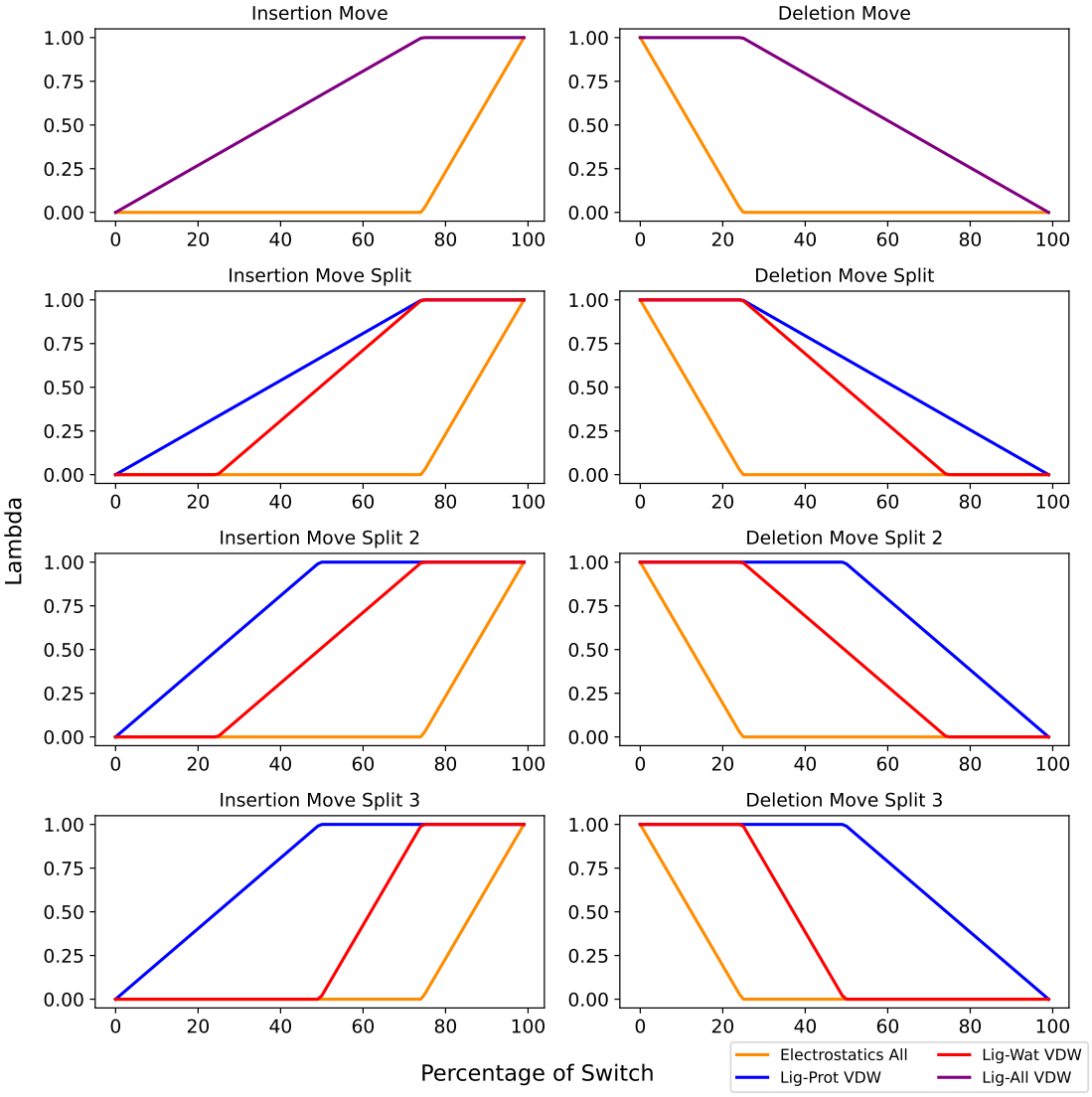


FIGURE 9.6: Comparison of the current lambda scheme and three new proposed schemes where the ligand interactions are split.

First, to validate the implementation we simply use the same engineered HSP90 system as before and again, couple the interactions while everything remains fixed. With no MD, it does not matter how the molecule is coupled, the overall nonequilibrium work associated with ligand coupling should remain the same - Figure 9.7 confirms this.

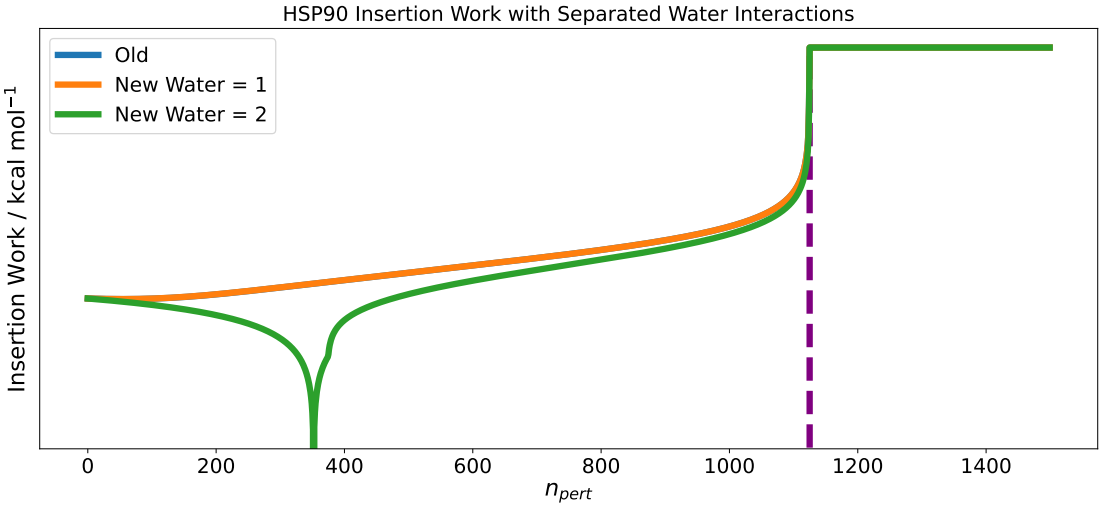


FIGURE 9.7: Comparison of the cumulative works between the standard implementation (blue) and the new implementation. Orange refers to the new code with water interactions not separated and green refers to them separated. The purple dashed line indicates the change from LJ switching to electrostatics. As expected, the work values are equal at the end of the LJ portion of the move. The green curve shows that the move begins favourably as the interactions are switched between only the ligand and the protein. The move then starts to become unfavourable as the interactions with the overlapping waters are switched on at approx. $n_{pert} = 250$. Note work values have been normalised between 0 and 1 and plotted on a log scale.

Next, to understand if splitting the interactions provides any benefit, we can reintroduce dynamics to the system. First, we can reintroduce the water dynamics while keeping the ligand and protein fixed. In this case, it is expected that the overlapping water molecules will unbind as the ligand is switched on regardless of whether the interactions are separated. This is simply because the waters are the only part of the system which can move to reduce the repulsive overlap, so acts as a simple sanity check (Fig. 9.8). This result implies that if the ligand were in place, and tightly bound enough to the protein, then any water molecules overlapping with the ligand would move to accommodate the ligand as the ligand-water interactions are turned on, even by just simple MD. In other words, the balance has shifted from the presence of water making ligand binding unfavourable, to the opposite. Encouragingly, waters leave the pocket later in the move when using the separated scheme.

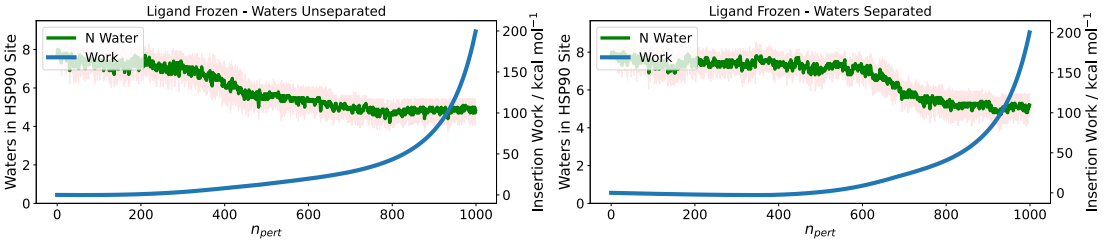


FIGURE 9.8: Comparison of insertion works and the desolvation of the HSP90 pocket as the ligand interactions are switched on between separated (Fig. 9.6 row 2) and non-separated water interactions. In these moves, only water is allowed to move by MD.

The next test is to perform the same ligand insertion moves but with full dynamics. As a reminder, these are ligand insertion moves starting from the bound configuration using the lambda schedule seen in the 2nd row of Figure 9.6. The results are shown in Figure 9.9. Unfortunately, in this case, the ligand still quickly unbinds (top) in its early lambda windows likely owing to the weak interactions with the protein. Using a flat bottom restraint (bottom right) results in a small amount of water being removed from the pocket and a consistent binding pose with an RMSD to the crystal pose of approximately 4 Å. However, this is also observed in moves where the interactions are not split (bottom left), prompting speculation as to whether the splitting is having the desired effect or if the addition of the restraint is having the most impact. That said, the average work done is lower in the separated scheme. Adding GCMC water moves to this protocol is a natural next step for future work.

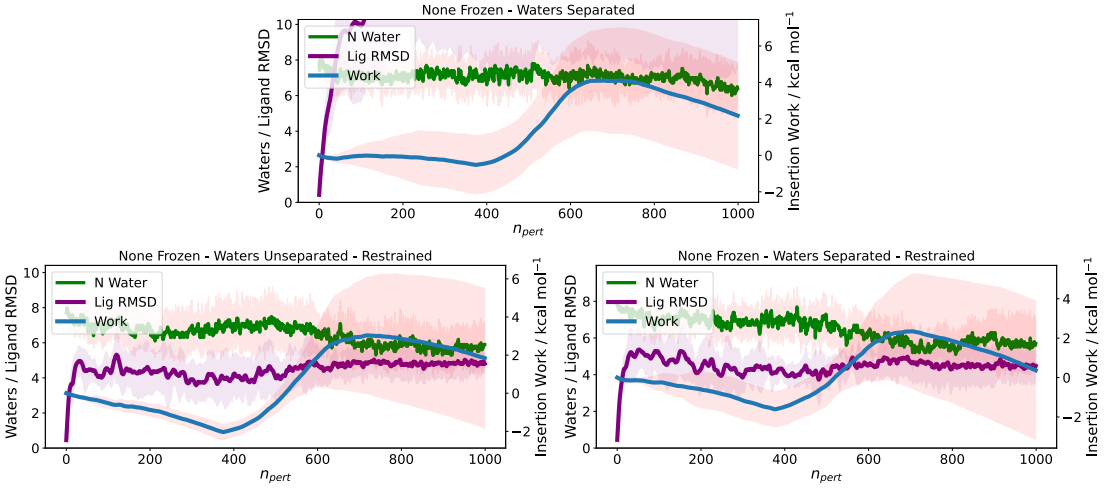


FIGURE 9.9: Comparison of insertion works and the desolvation of the HSP90 pocket as the ligand interactions are switched on between separated and non-separated water interactions. In these moves, all species are free to move. Using a flat bottom restraint keeps the ligand bound in both schemes, though not in the crystal pose.

9.4 Overlapping Side Chains and Cryptic Pockets

In a similar vein, sidechains which may overlap with the desired binding site, or cryptic pockets which may require a large conformational change to open, are poorly sampled by GCNCMC moves. Again, if these side chains exist in an occluded site, the move may result in high works, whereas in exposed sites, the ligand may diffuse away. Similarly, for cryptic pockets, these weakly interacting molecules are unable to induce the conformational changes required for binding. This is not unexpected as these changes occur over much longer timescales, even with fully interacting probes, and often require the use of other enhanced sampling techniques as exemplified in Chapter 7 where some cryptic pockets were explored.

One avenue for exploration may be to combine GCNCMC simulation with another system-agnostic enhanced sampling method such as Gaussian Accelerated MD (GaMD). It has already been stated in Chapter 7 that aMD simulations significantly improve the sampling of cryptic and occluded sites in MSMD simulations and therefore may be of use in the context of GCNCMC. The reweighting of GaMD simulations requires fewer data points to converge the error of the bias potentials compared to regular aMD, making it a more efficient approach.[245,257,272,273,275] Furthermore, it has been shown that GaMD can be flexible in terms of the potential energies to which a boost potential is applied.[245,272] It may be that localising the boost potential to protein atoms around the GCMC region may yield strong sampling improvements while not perturbing the system too much.

In an early proof of concept, we revisit T4L99A. Para-xylene, upon binding in the apolar site, causes the side chain of Val111 (χ_1) to flip from a trans conformation to a gauche (G-) conformation.[1,112,132,244,252] However, this movement is often not captured in traditional free energy calculations leading to an inaccurate result.[1,133,197] If the ligand is decoupled from its holo state, and the side chain fails to readjust to a trans conformation during the decoupling, an overly favourable free energy is calculated, as the effect of the side chain movement is not captured. Conversely, coupling the ligand from the apo state results in an overly high free energy if the side chain fails to adopt the favourable gauche conformation. This system has been the subject of many enhanced sampling method development studies including BLUES and AASMC.[112,132]

To this end, we have combined the openmm implementation of GaMD[275] with GCNCMC simulations of p-xylene and compare the sampling of the Val111 χ_1 dihedral across different protocols. Note this is preliminary work and the simulations have not been reweighted. It is presented here as proof of concept only.

First, we confirm that pure GaMD simulations with and without the ligand-bound maintain a gauche and trans valine conformation respectively. Figure 9.10 confirms that removing the ligand from the *holo* structure causes a quick relaxation of the Val111 conformation to the trans state, while when the ligand is bound, the simulation maintains the gauche state. Similar observations were made with vanilla MD (not shown).

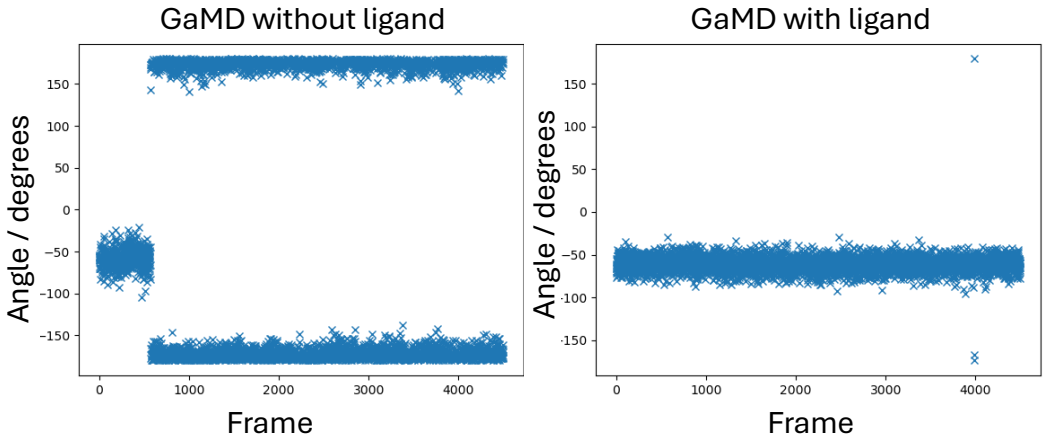


FIGURE 9.10: Time series of the Val111 χ_1 dihedral with and without a p-xylene ligand bound. Both simulations are initiated from the holo structure with the ligand removed or retained.

Performing GCNCMC insertions into the binding site with a trans valine would ideally induce the change from trans to gauche, however, the upper left plot of Figure 9.11 shows that GCNCMC insertions have little effect on the valine conformation. This is likely due to the large energy barrier preventing χ_1 rotation as the ligand is inserted. The purpose of accelerated methods is to add selective boost potentials to various energy terms to smooth the energy surface of that term, intuitively, applying a boost potential to the protein dihedrals in a system would result in enhanced sampling of rotameric states and has indeed been used in the past to sample rare events such as protein folding.[272,275] The default implementation of GaMD (upper right), uses a 'dual' boost which applies a boost potential to the total system energy and all of the protein dihedrals. In this case, we see minimal improvement in the sampling of the Val111 rotamer with trans still being the most dominant.

In accelerated methods, there are restrictions on how much boost can be applied to the system, with higher boosts making it more difficult to reweight. Using a dual boost on the whole system and all protein dihedrals means that the boost is spread across the whole system. For this use case, this is not required, and instead, we can focus the dihedral energy boost on just the protein residues in and close to our GCMC region. With this (bottom left), we see much more sampling of the valine side chain with gauche now the more dominant configuration. It is worth mentioning that this result may arise, in part, owing to templating from an insertion move causing the valine to flip and not allowing the system to recover the trans configuration after a forced deletion move. This was an oversight and should be investigated properly in the future. Interestingly, focusing the boost even further on just the specific valine residue results in less sampling of the gauche conformation and it is currently unclear why. However, focusing on specific residues requires prior knowledge of the system and is not conducive to the design philosophy of being system agnostic.

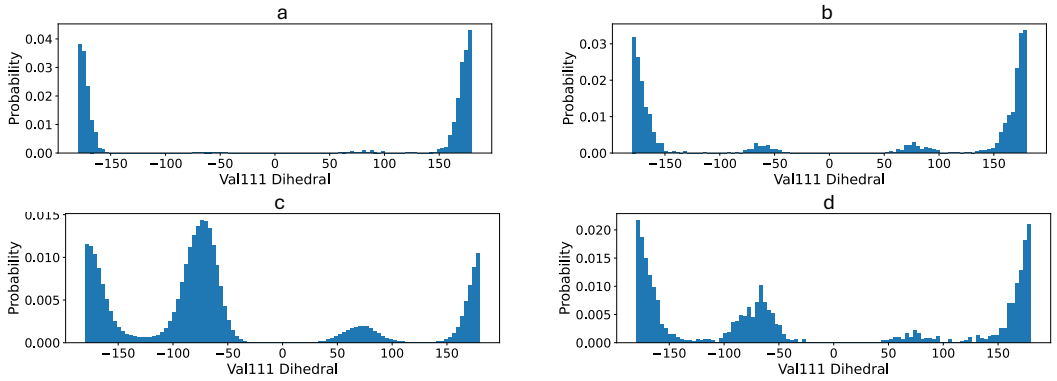


FIGURE 9.11: Comparison of Val111 χ_1 distributions for different MD and GaMD protocols. Dihedrals are recorded after each GCNCMC insertion move. A: Pure MD. B: Default GaMD Dual Boost. C: Focused GaMD dihedral boost on all protein residues within the GCMC sphere. D: A focused GaMD dihedral boost on just the valine residue.

In summary, applying GaMD to GCNCMC simulations may result in enhanced sampling of protein conformational changes. However, there is still much work to do to refine this protocol and method. Namely, a much more thorough study needs to be performed not only with side chain rotamers in binding sites but also pocket opening as in Chapter 7. A method of reweighting these simulations is crucial.

9.5 Summary

In this chapter, we have discussed some issues which have been encountered throughout this project. All can be attributed to poor sampling and are common amongst alchemical methods. Two of the most prevalent involve the displacement of bound water molecules and sampling protein conformational changes such as side chain rotamers. Ideas for solving the water issue have been presented by combining GCNCMC ligand moves with GCMC water moves similar to regular free energy calculations with enhanced water sampling. In the case of enhanced protein sampling, we have discussed the issue and suggested that accelerated MD methods may be of use. In either case, more theorising, testing and validation is required. It is stressed that solving these issues in a system-agnostic way will not only benefit the present method but will also be widely beneficial to alchemical-based methods in general, making it a powerful avenue for exploration and of significant interest to the wider community.

Chapter 10

Conclusions

10.1 Summary

This thesis has presented the development and application of the Grand Canonical Nonequilibrium Candidate Monte Carlo (GCNCMC) methodology and other computational approaches for predicting fragment binding sites, modes, and affinities. By addressing key challenges in fragment-based and structure-based drug design, the research outlined in this work contributes to the advancement of *in silico* methods in the field. An open source Python module for ligand-based GCNCMC, *grandlig*, has been developed and made publicly available for use by the wider community. The module is designed to act as a plugin to the popular MD engine, OpenMM, and should therefore garner much interest. A summary of the key findings of this thesis is presented below.

In Chapter 3, we apply a collection of simple methods to the ERK2 MiniFrag system. We show that the static structure method, FTMap,[88] performs well at identifying MiniFrag binding sites when using *holo* structures. However, the method begins to fail when using an *apo* structure. This highlights the limitations of static structure methods which do not sample pocket opening and closing. To that end, we then applied a basic mixed solvent MD protocol whereby short MD simulations of the *apo* structure solvated in a high concentration of MiniFrags are performed. In these simulations a marginal improvement over the FTMap results was observed, however, it was still difficult to discern the known binding sites from the background noise. This is believed to be due to the weak binding affinities of these fragments which was confirmed using ABFE calculations. Overall, this chapter highlighted the need for more enhanced sampling methods to sample the binding of weak fragments.

Chapter 4 introduces the application of GCNCMC to small molecule binding. Aspects which differ from the previous water-based implementation are discussed. Namely,

the calculation of, and the concentration dependence on, the excess chemical potential. It is shown that the presence of other molecules of the same species can affect the value of μ'_{sol} in a concentration dependant manner, however, it is concluded that this effect is minimal at sufficiently dilute concentrations. A validation of the method's sampling ability is performed by reproducing a simple ensemble property, concentration. It is shown that, for an appropriately calculated value of μ'_{sol}, system concentration is well reproduced. However, it is also shown that slight inaccuracies in the excess chemical potential can have adverse effects on the overall concentration of the system. We again conclude that these effects are much less pronounced when simulating dilute concentrations such as those used in protein-ligand applications. Finally, the method is applied to a simple host-guest system β-cyclodextrin to gain a basic understanding of the method's behaviour.

In Chapter 5, two novel methods for calculating binding affinity using the nonequilibrium work measurements derived from GCNCMC simulations are presented. These methods include GCNCMC titrations (Eq. 5.22) and the Grand Canonical Acceptance Ratio (GAR, Eq. 5.28). The performance of each estimator is demonstrated by calculating hydration free energies for a subset of molecules from the FreeSolv[153] database and the binding free energies of guest molecules to β-cyclodextrin. In both cases, identical results between the methods are observed and are in very good agreement with an alternative method, FEP. This not only provides empirical evidence of the accuracy of the two free energy estimators but also a wider validation of the GCNCMC method. These results prove that all the derivations and fundamental theories behind GCNCMC are rigorous and obey the rules of statistical mechanics.

Chapter 6 then builds on all the lessons learnt from the previous chapters and demonstrates how the method may be used in a structure-based design setting. We use two model systems with known occluded binding pockets, T4L99A and MUP1. In both cases, we begin by assuming no prior knowledge and using GCNCMC enhanced mixed solvent MD (MSMD) simulations to identify the binding pockets. The addition of GCNCMC to MSMD (GC-MSMD) resulted in the rapid identification of the pockets whereas regular MSMD failed. From this, we performed GCNCMC titration calculations, for a series of molecules binding to each protein, by exchanging molecules directly into or from the binding sites identified by GC-MSMD. In both cases, the free energy estimates are in good agreement with experimental values and crucially in line with a more established FEP protocol. Using toluene-T4L99A as a test system, we show that, as a byproduct of constant insertion and deletion moves, multiple fragment binding modes are naturally sampled in a GCNCMC simulation. This has implications for the final free energy estimates where, as binding modes and symmetrically equivalent modes are inherently sampled, there is no requirement for

prior knowledge of the existence of multiple binding modes nor the need to run multiple simulations at each binding mode as in traditional FEP.

Given the success of GC-MSMD simulations in identifying the binding sites of T4L99A and MUP1, the application of GCNCMC to mixed-solvent MD simulations was explored further in Chapter 7. This protocol was applied to several protein systems to enhance the mapping of protein binding sites. In the majority of systems, the GC-enhanced method resulted in improved mapping and the method's ability to distinguish between specific and non-specific binding interactions was demonstrated, representing an improvement over vanilla MSMD simulations. The greatest improvements between the two methods were seen in systems where the binding site is particularly occluded from the solvent. However, there is still scope for further development, particularly in systems with binding sites that require a large conformational change to form, as these were poorly sampled in both GC-MSMD and regular MSMD.

Finally, Chapter 8 highlighted the utility of GCNCMC in fragment-based drug discovery, showcasing its ability to screen fragments and identify high-affinity binders. This simple application of GCNCMC requires running one simulation at a user defined concentration. Fragments can then be deemed a hit or a miss based on the fragment's average occupancy in simulation. The protocol was applied to another model system T4L99A/M102Q and was shown to distinguish between known hits and misses with high accuracy. This application of GCNCMC is less mature and requires further development in the future but represents a relatively inexpensive screening method.

10.2 Future Work

While the research presented in this thesis has made significant advancements, several limitations and opportunities for future work remain. Despite the improvements introduced by GCNCMC, certain systems with cryptic or solvent exposed binding sites remain challenging. Some of these sampling challenges are outlined in Chapter 9.

It is shown that systems which require the concomitant binding or displacement of solvent molecules or protein conformational change are problematic as the alchemical insertion and deletion of fragment molecules often fail to sample these events. Further optimization of sampling protocols and integration with enhanced sampling methods, such as water-based GCMC and Gaussian accelerated MD, could address these limitations, with initial work and proof of concepts presented in Chapter 9. A key result is the use of a flat bottom restraint to prevent the diffusion of molecules from the GCMC region during coupling and decoupling. This not only results in more valid

moves but also allows for higher quality moves by restricting the sampling to only one pathway. This should be explored further in the future.

The computational expense associated with GCNCMC simulations also remains a barrier and future work should focus on optimizing the algorithm and methodology.

Finally, a broader validation using diverse protein-ligand datasets is necessary to confirm the method's generalizability and applying GCNCMC to novel and clinically relevant targets could further demonstrate its potential to address unmet needs in drug discovery.

10.3 Final Remarks

This thesis has yielded several insights and advances towards the GCNCMC method. The technique was introduced as a robust method for the prediction of binding sites, modes and affinities. This approach effectively mitigates some sampling limitations associated with classic MD simulations by enabling the quick and rigorous exchange of ligands within a defined region.

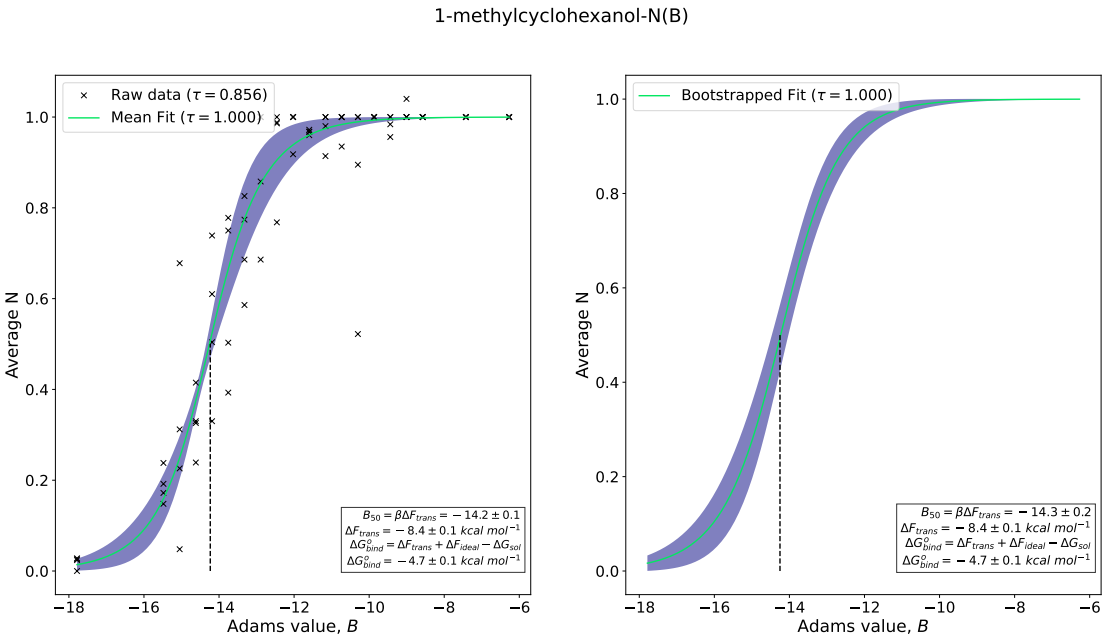
The integration of GCNCMC with molecular dynamics (MD) simulations provides a powerful framework for characterizing ligand binding. GCNCMC was successfully validated against benchmark systems, such as β-cyclodextrin host-guest systems, and applied to protein-ligand systems including T4 lysozyme and major urinary protein-1. These applications demonstrated the method's versatility and accuracy in predicting binding free energies and identifying key binding poses.

The findings of this thesis have broad implications for computational drug discovery. The ability to predict binding sites and affinities should enable more efficient fragment screening, reducing experimental costs and timelines. The integration of GCNCMC into fragment-based drug discovery workflows will hopefully improve hit identification and optimization. The detailed characterization of binding sites and hotspots supports the rational design of high-affinity ligands in structure-based drug design. GCNCMC's ability to sample deeply buried, occluded pockets makes it valuable for challenging SBDD targets. With further development, this method has the potential to unlock new opportunities in drug discovery.

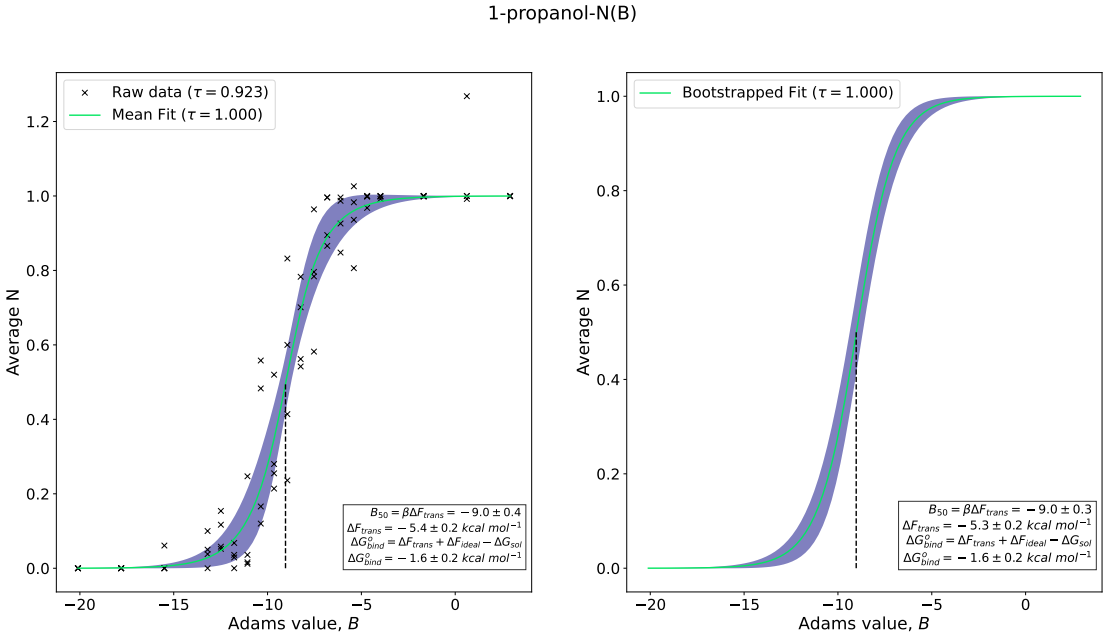
Appendix A

Titration Curves

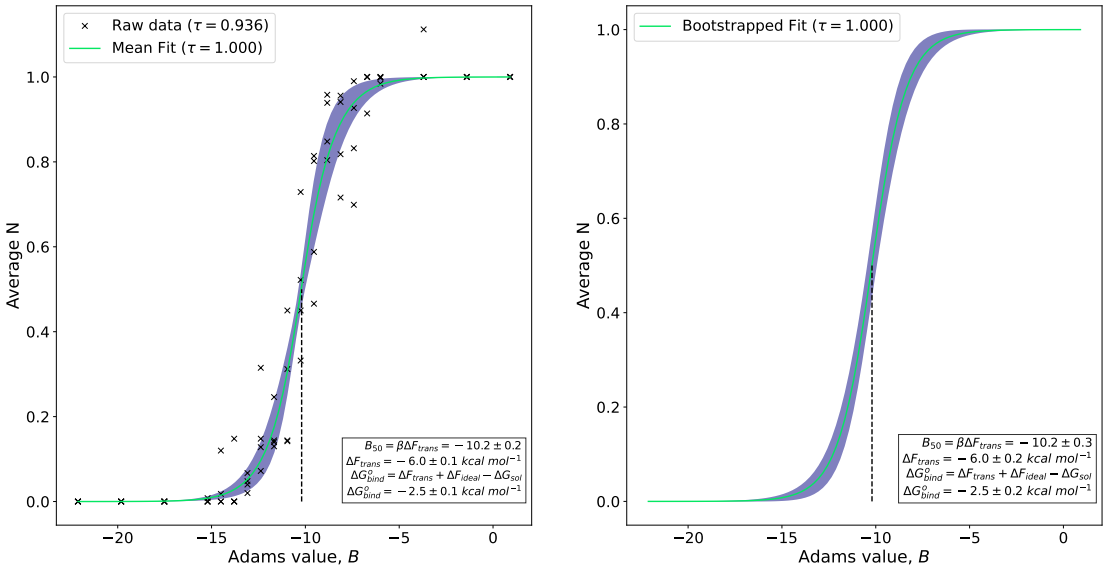
A.1 Host Guest Titrations



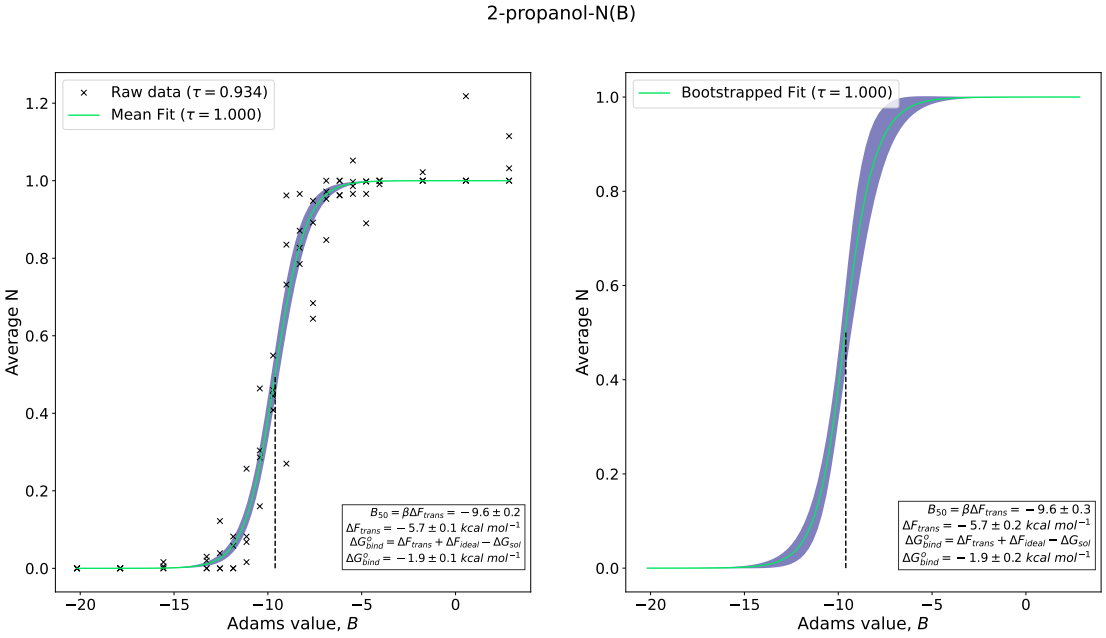
1-methylcyclohexanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



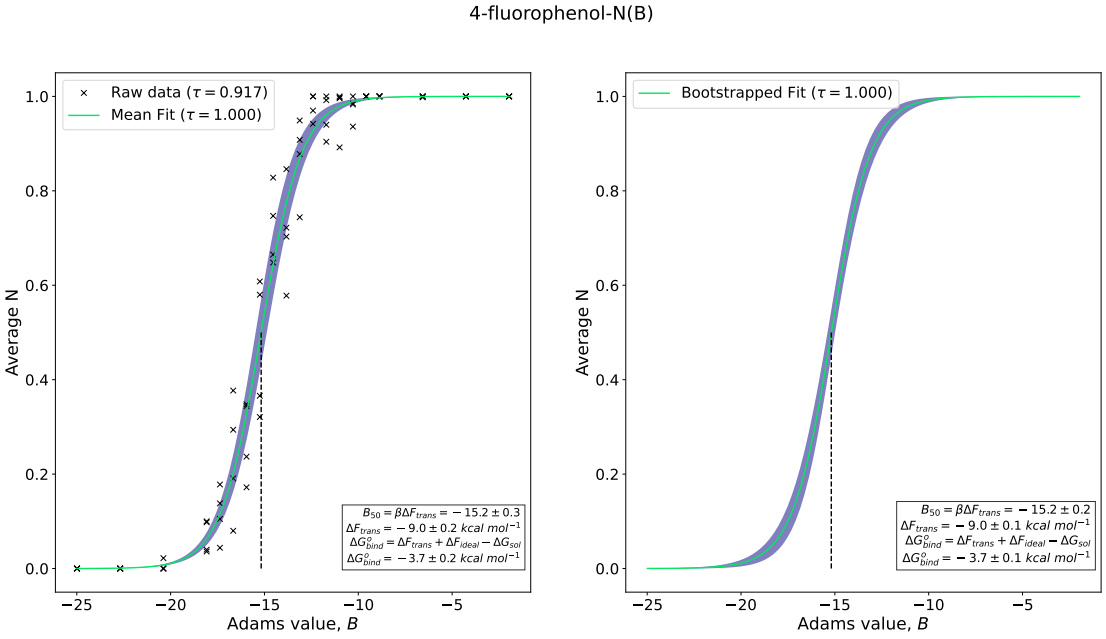
1-propanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



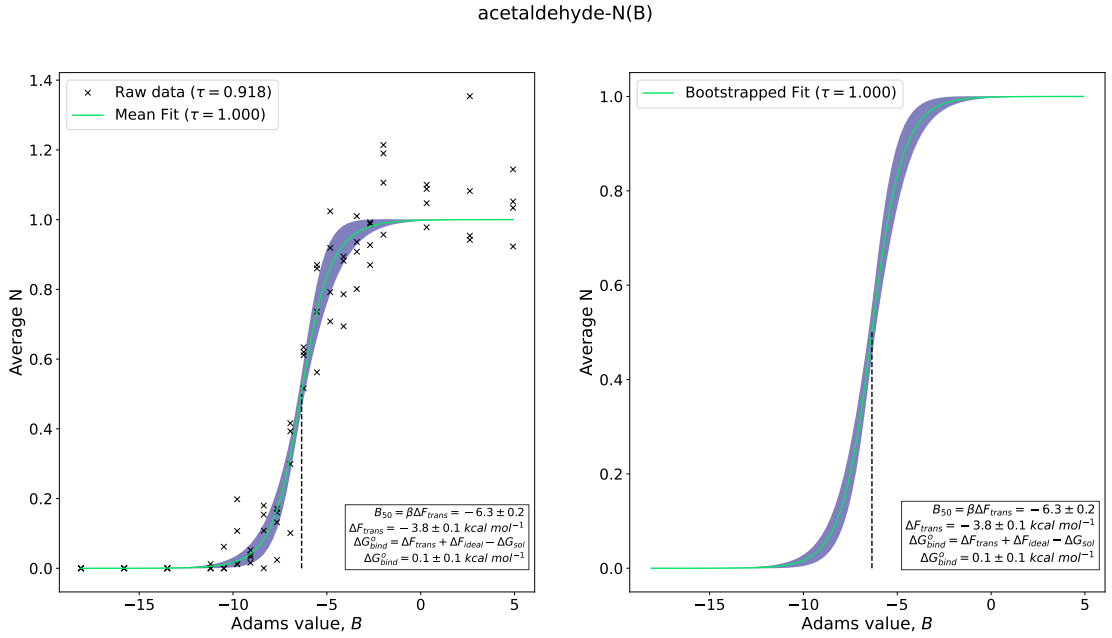
2-butanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



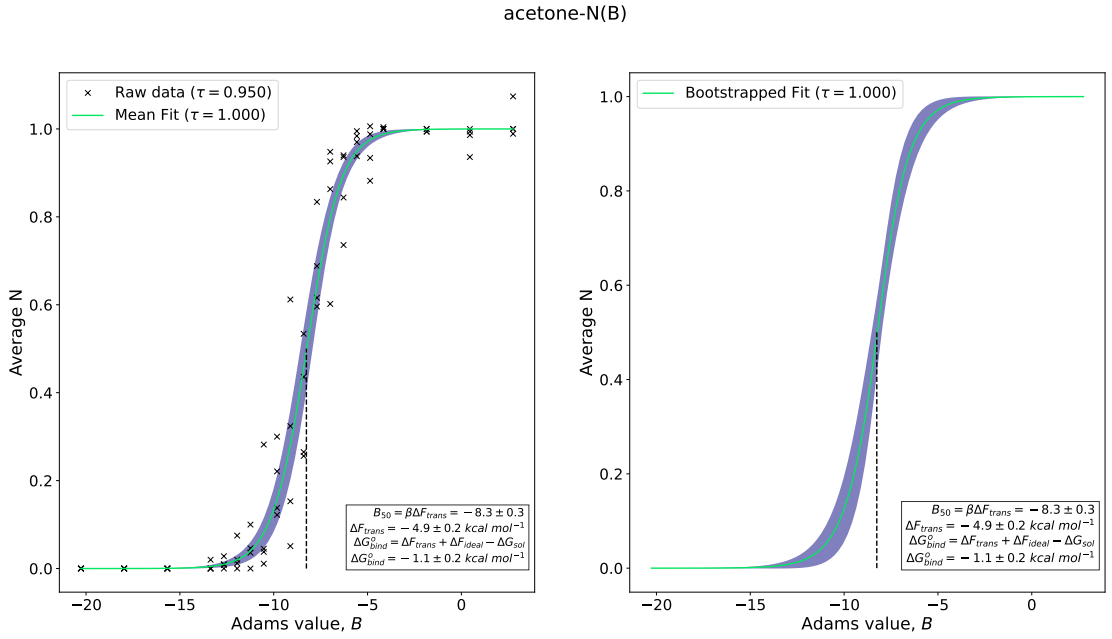
2-propanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



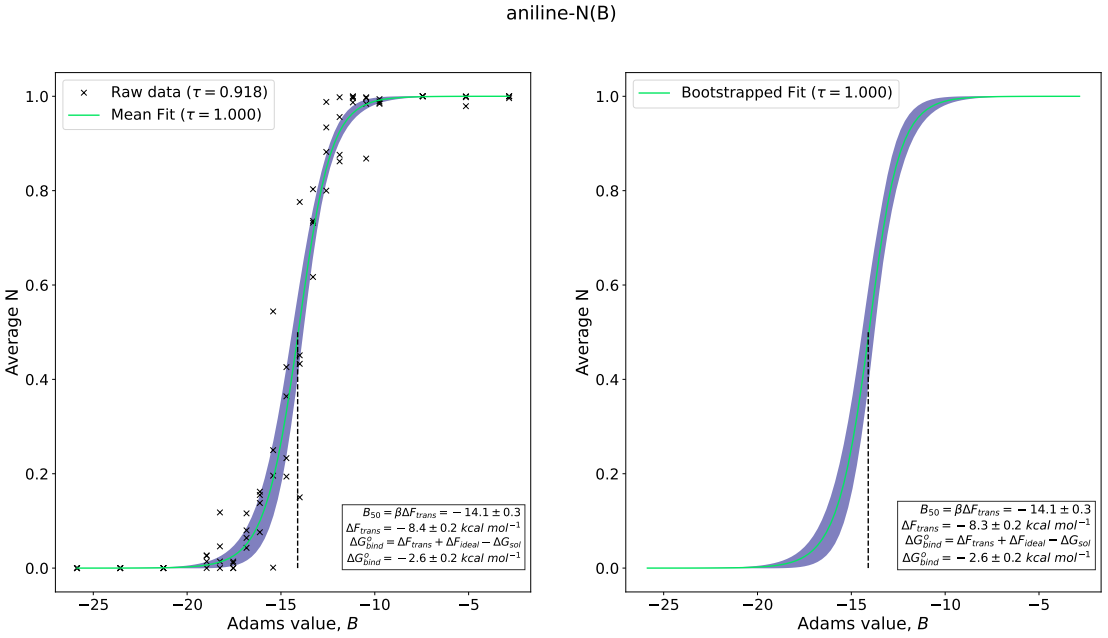
4-fluorophenol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



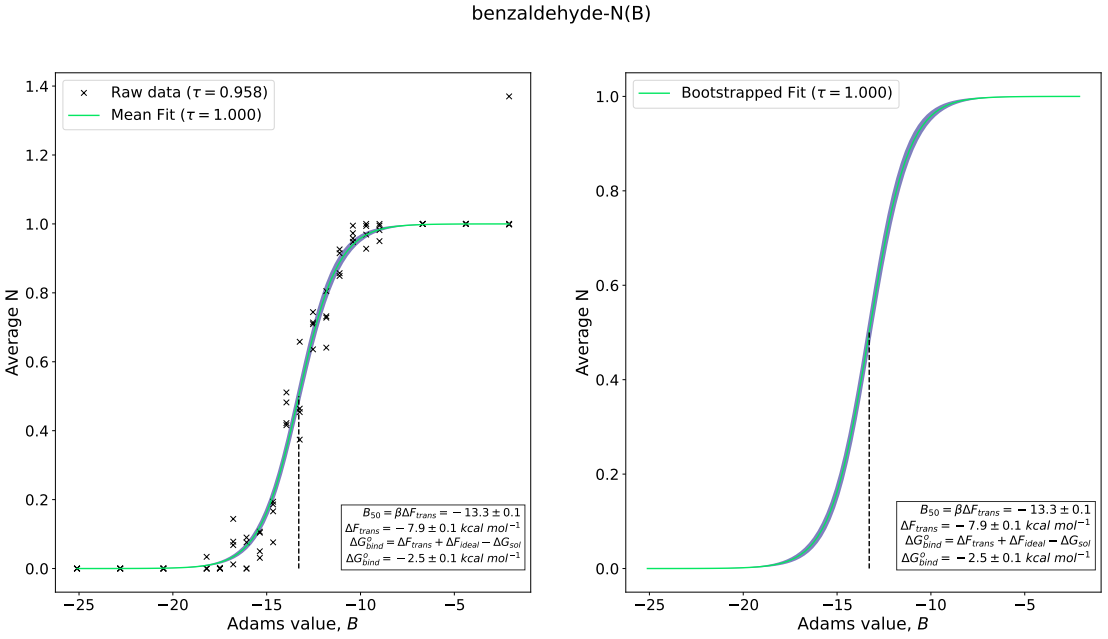
acetaldehyde titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



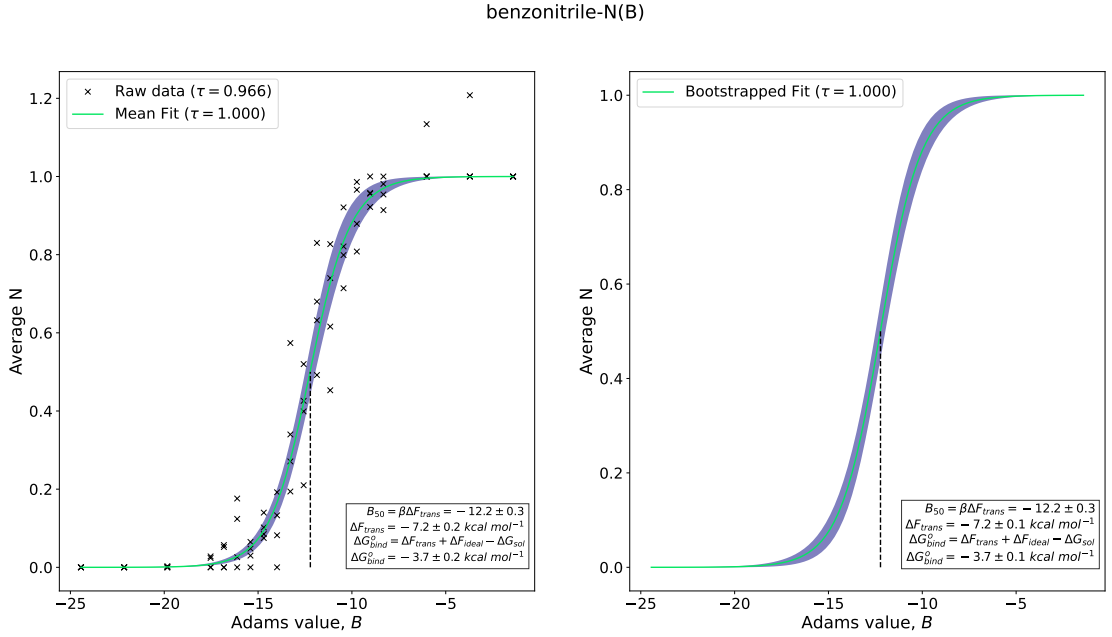
acetone titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



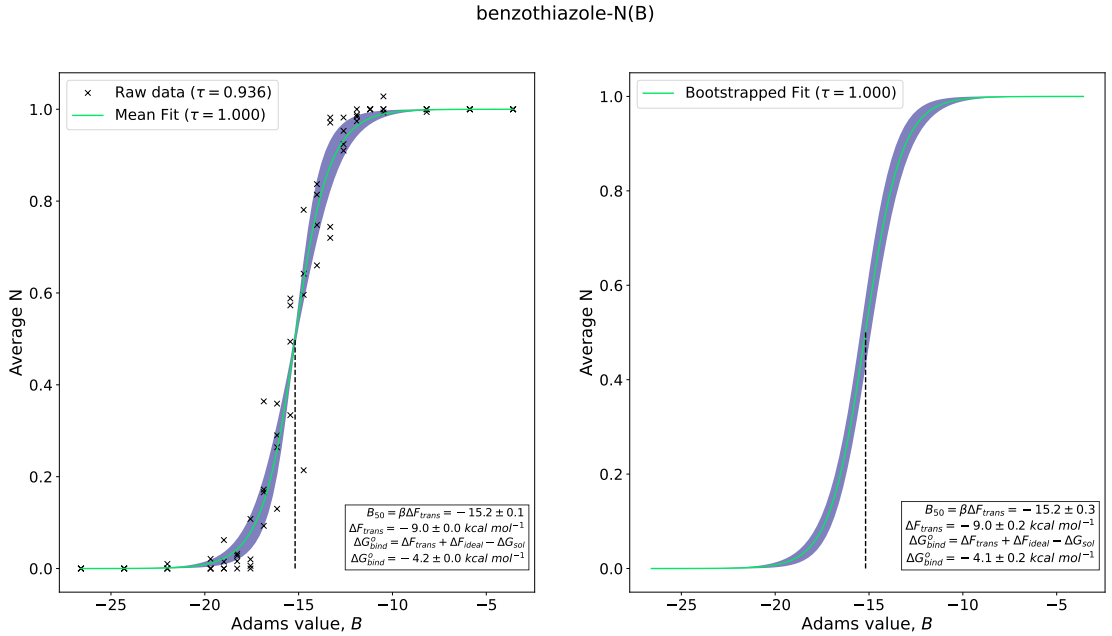
aniline titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



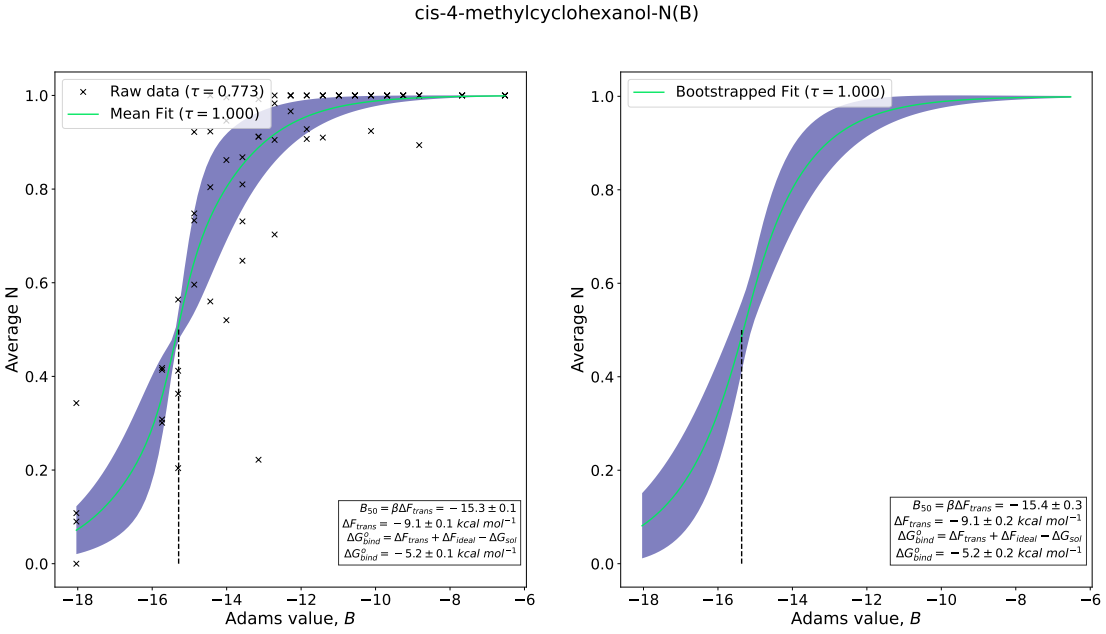
benzaldehyde titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



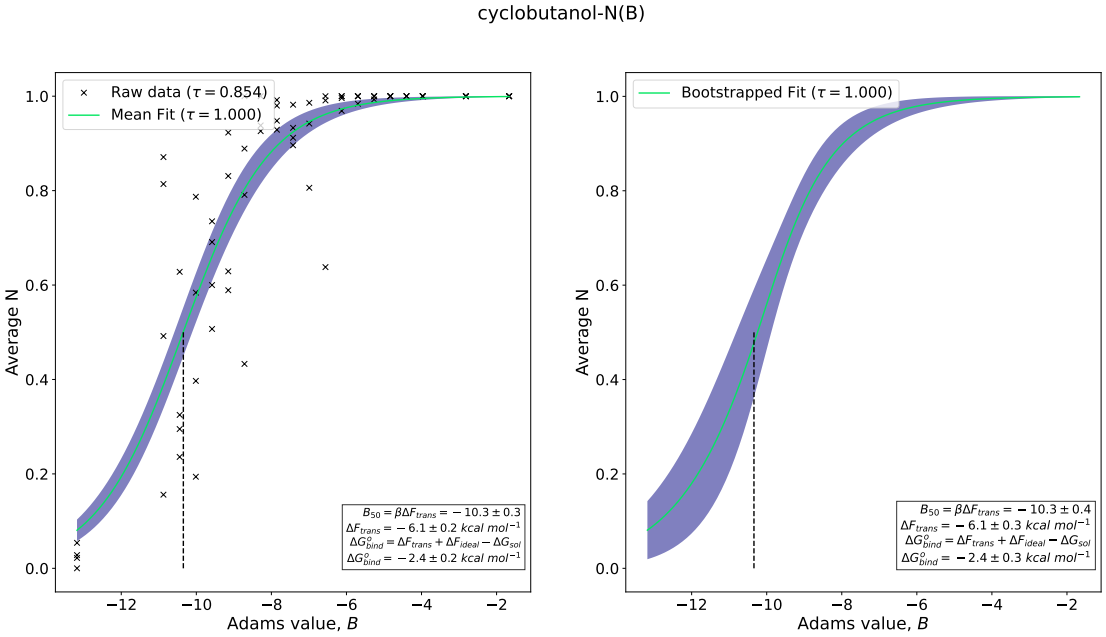
benzonitrile titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



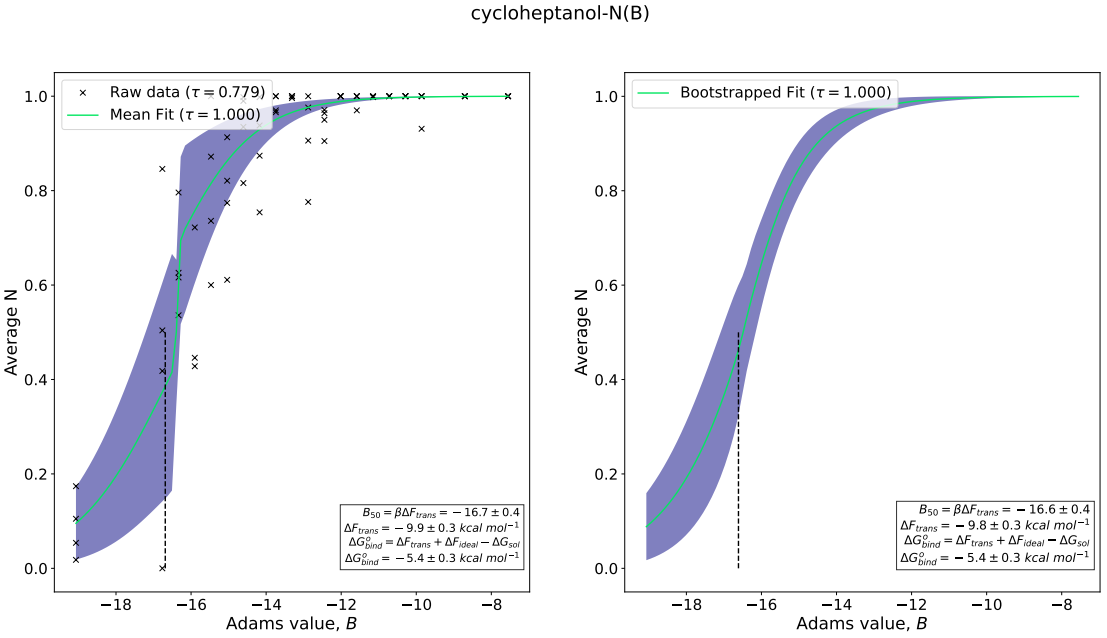
benzothiazole titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



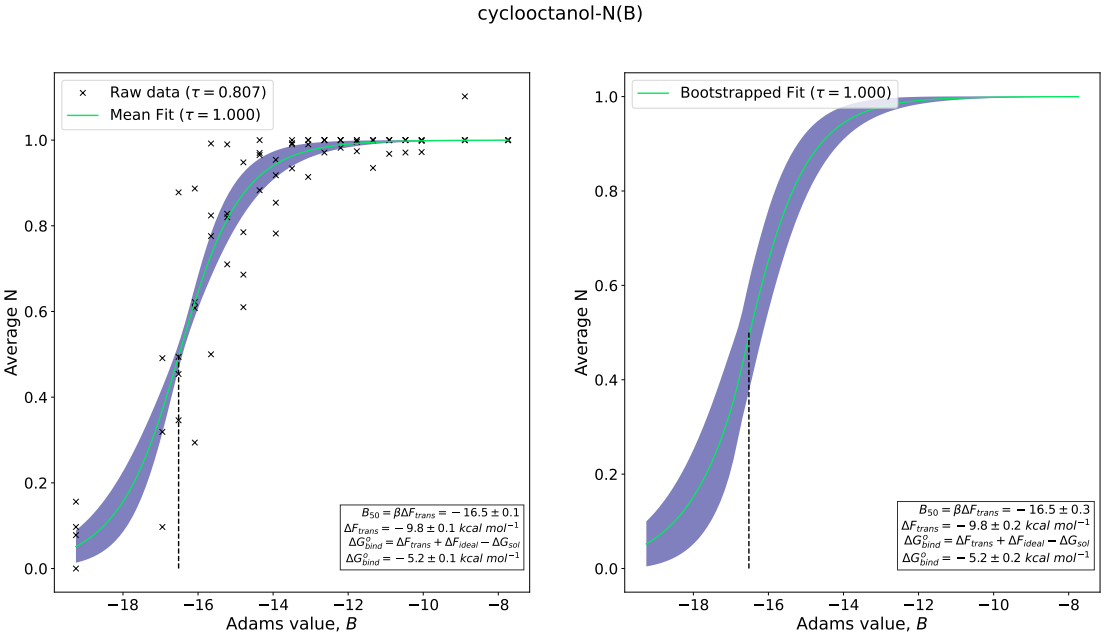
cis-4-methylcyclohexanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



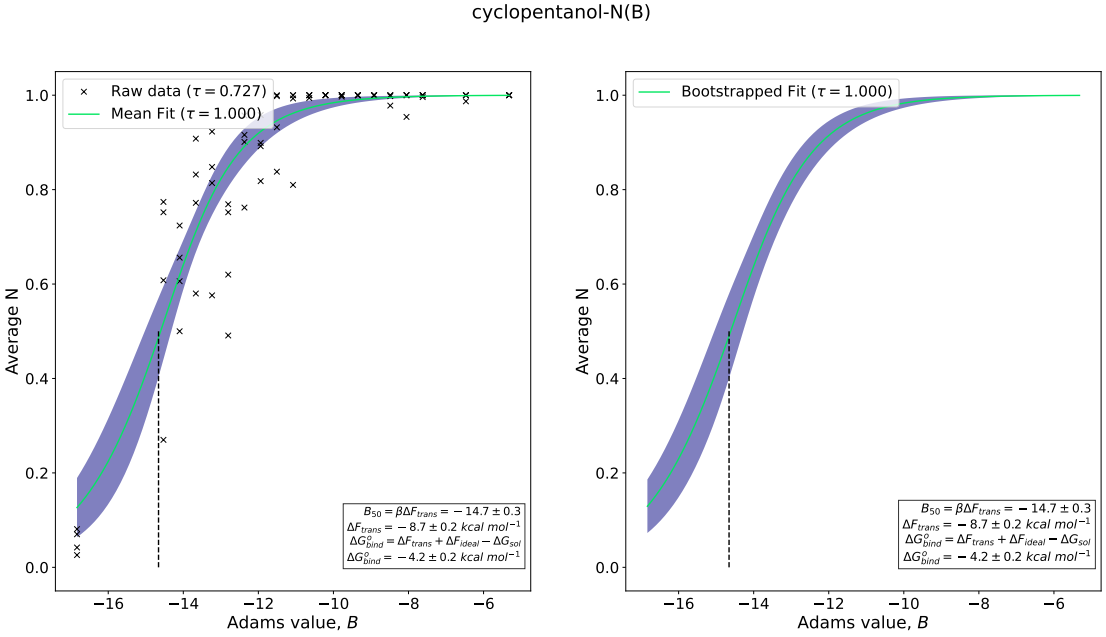
cyclobutanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



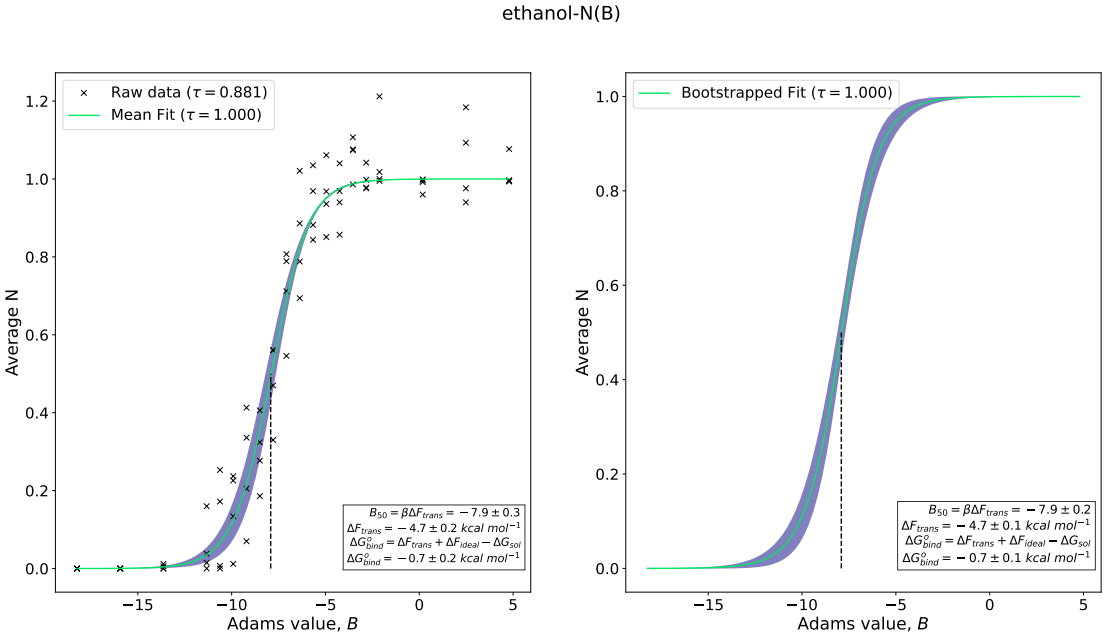
cycloheptanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



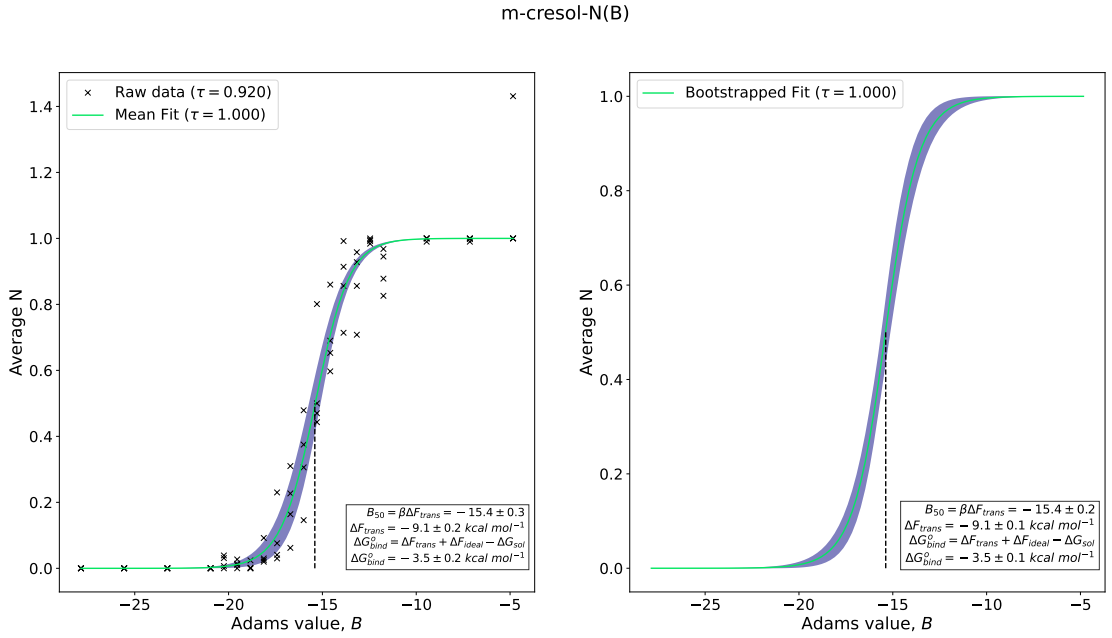
cyclooctanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



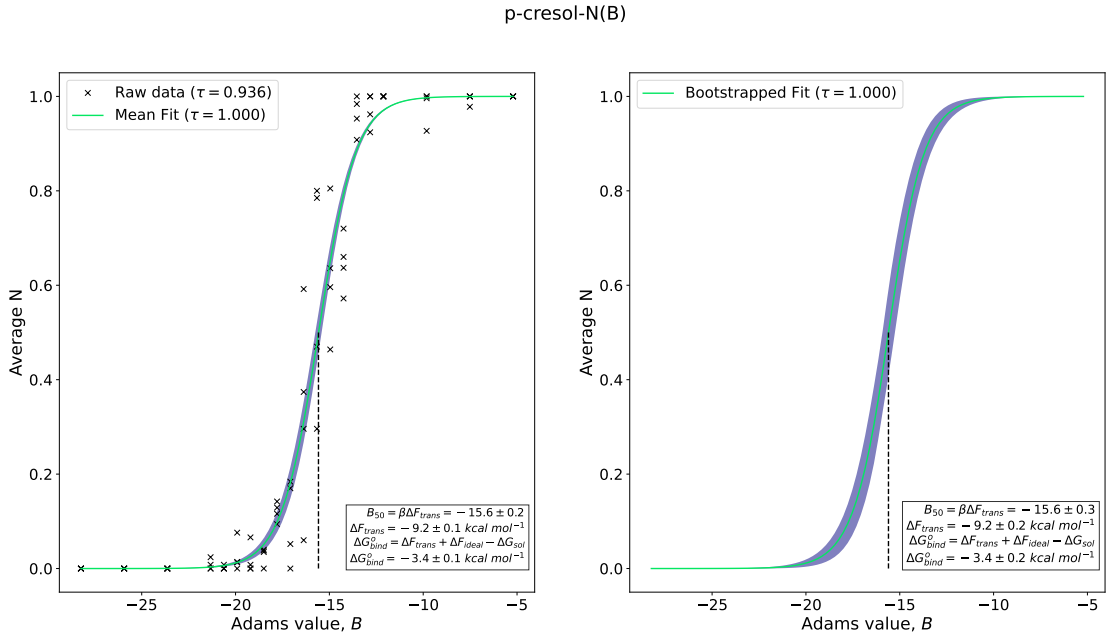
cyclopentanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



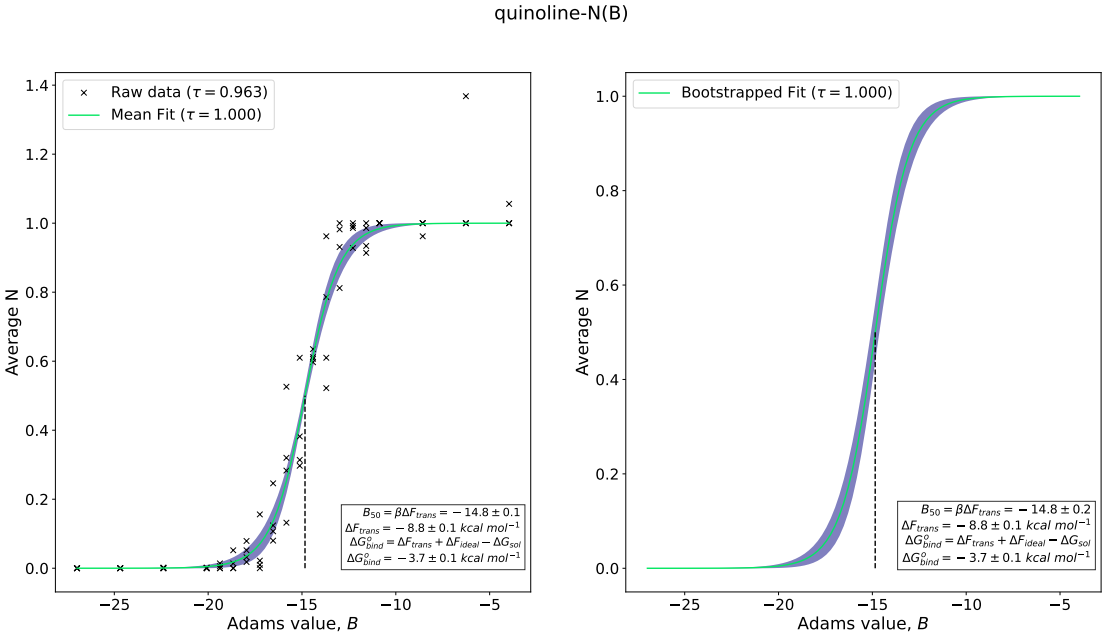
ethanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



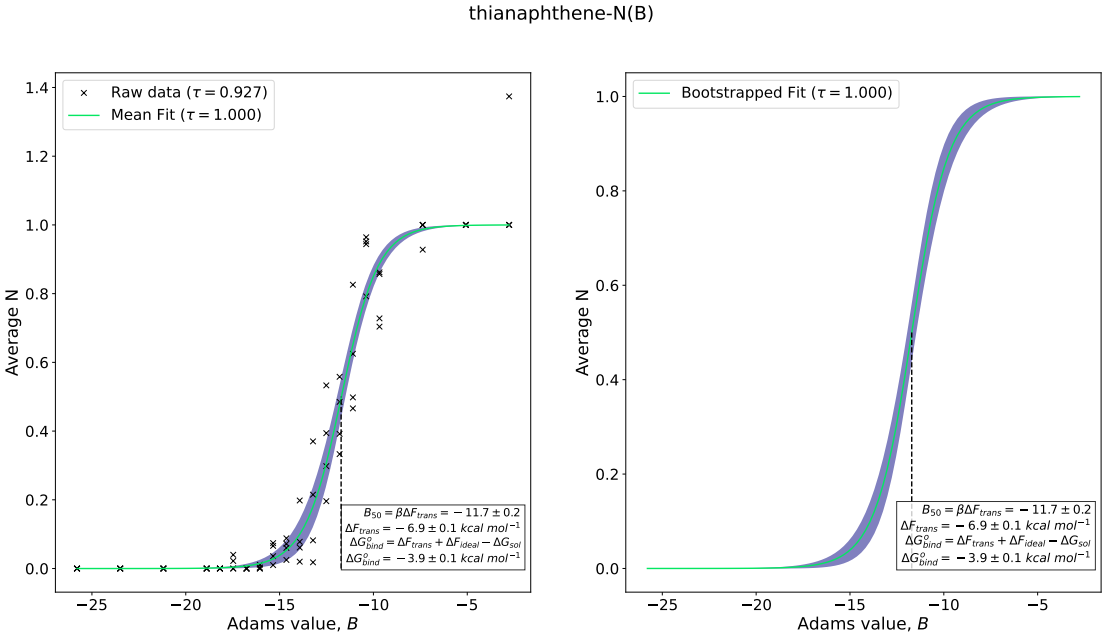
m-cresol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



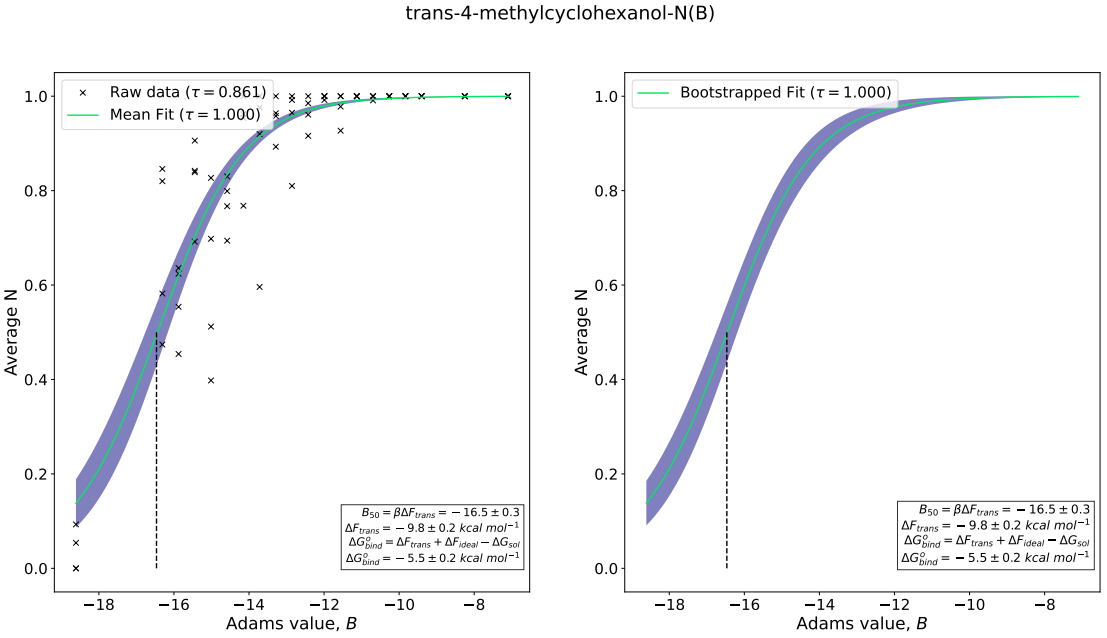
p-cresol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



quinoline titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.

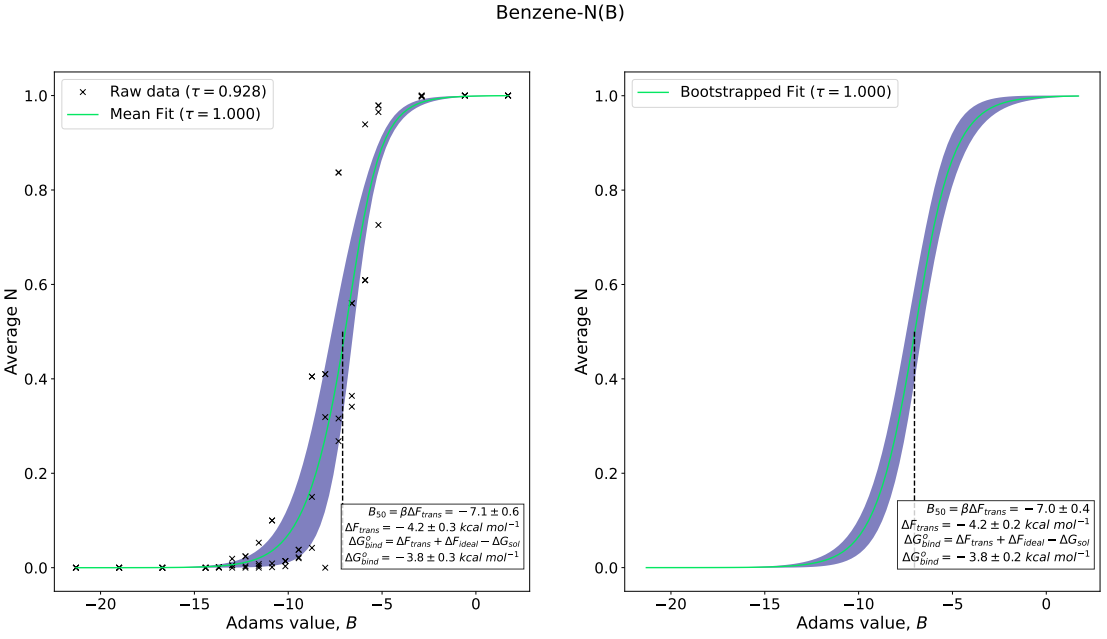


thianaphthene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.

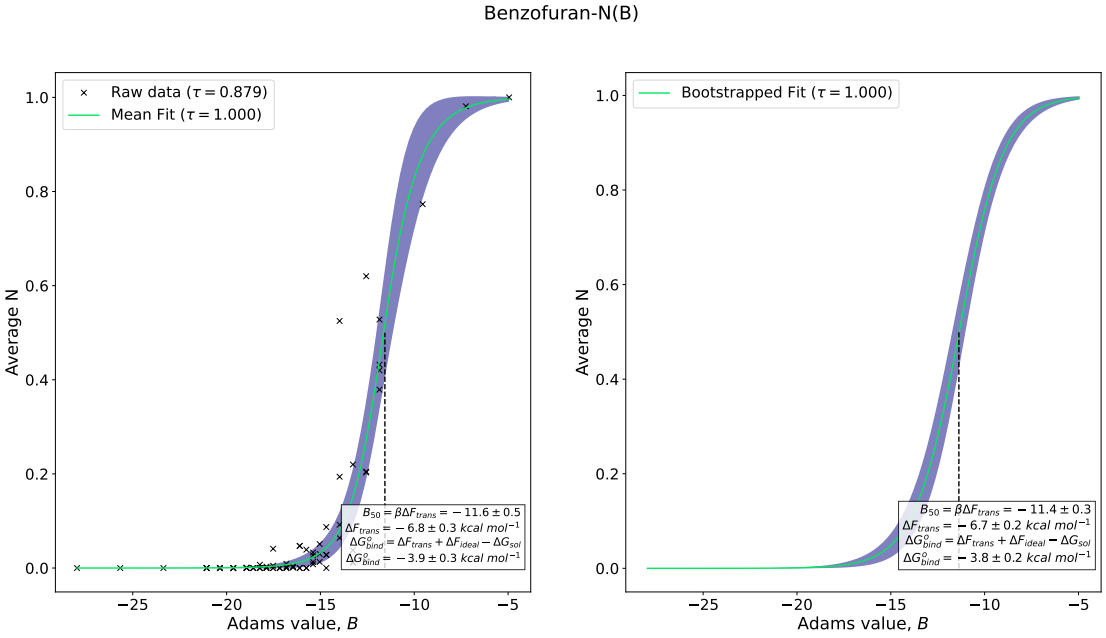


trans-4-methylcyclohexanol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.

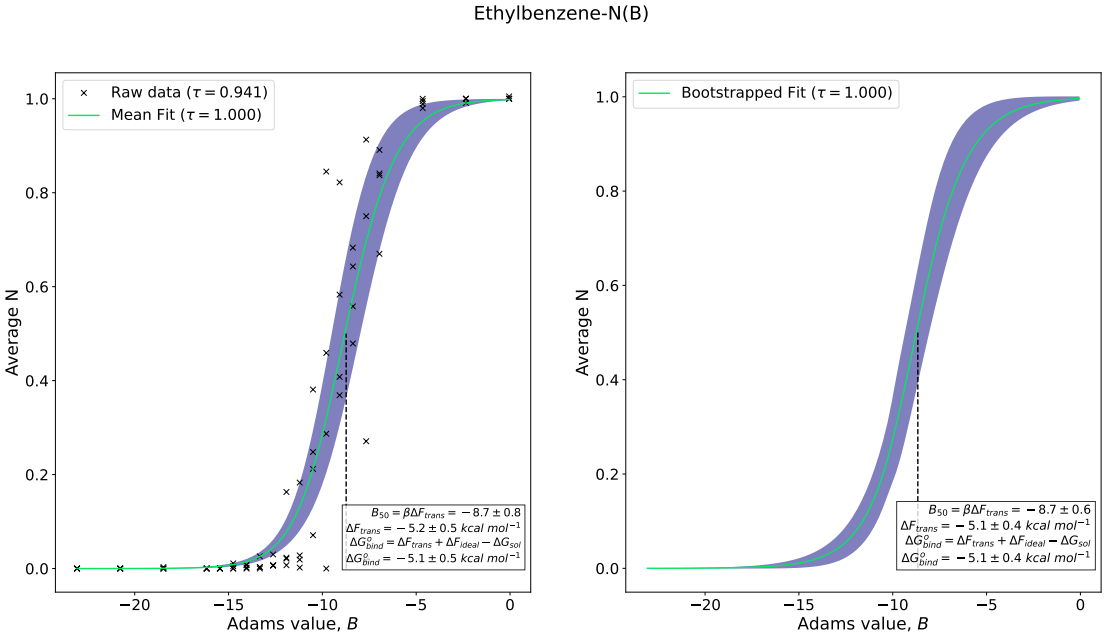
A.2 T4L99A Titrations



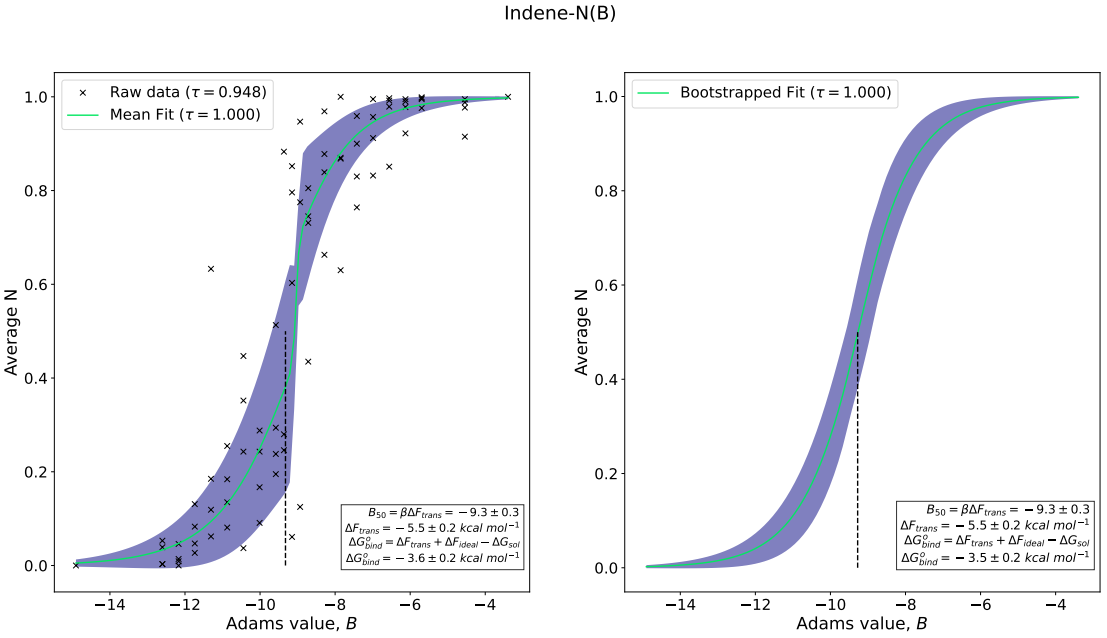
Benzene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



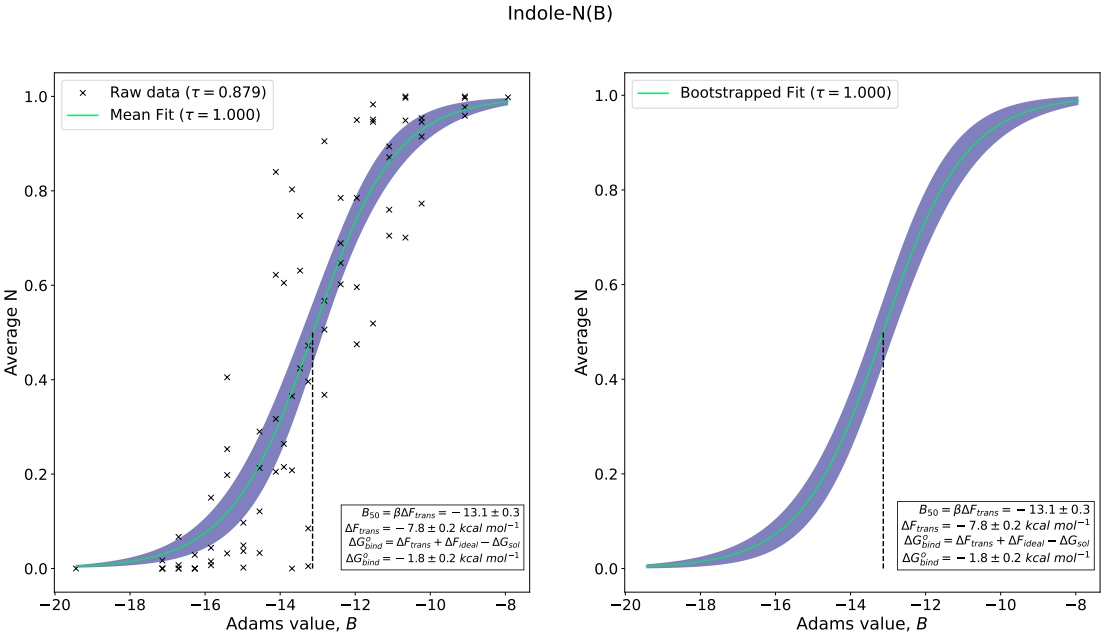
Benzofuran titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



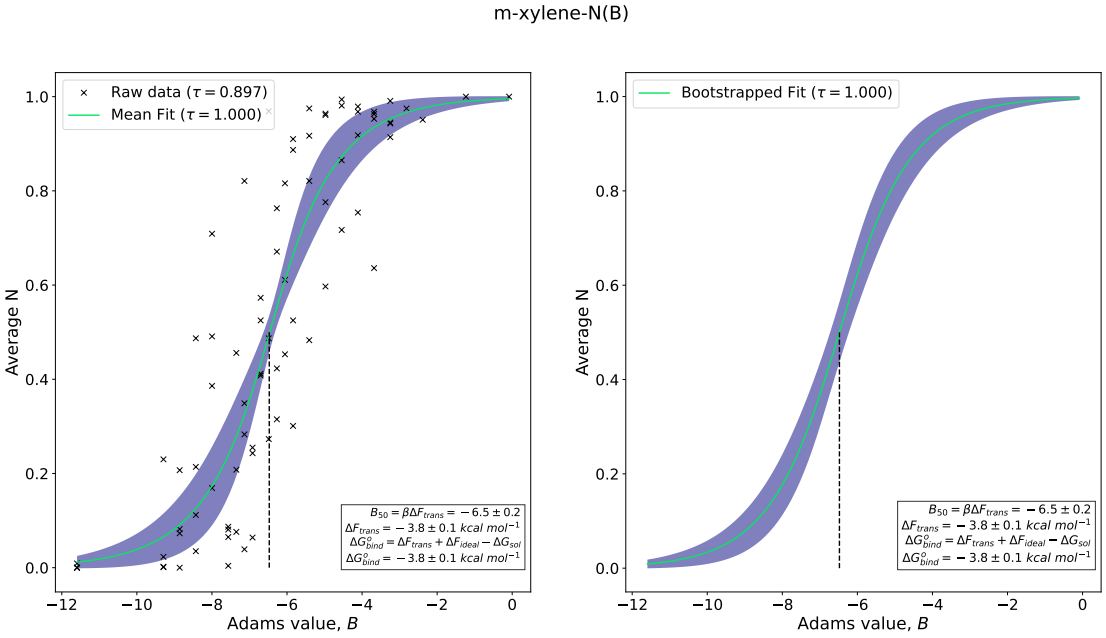
Ethylbenzene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



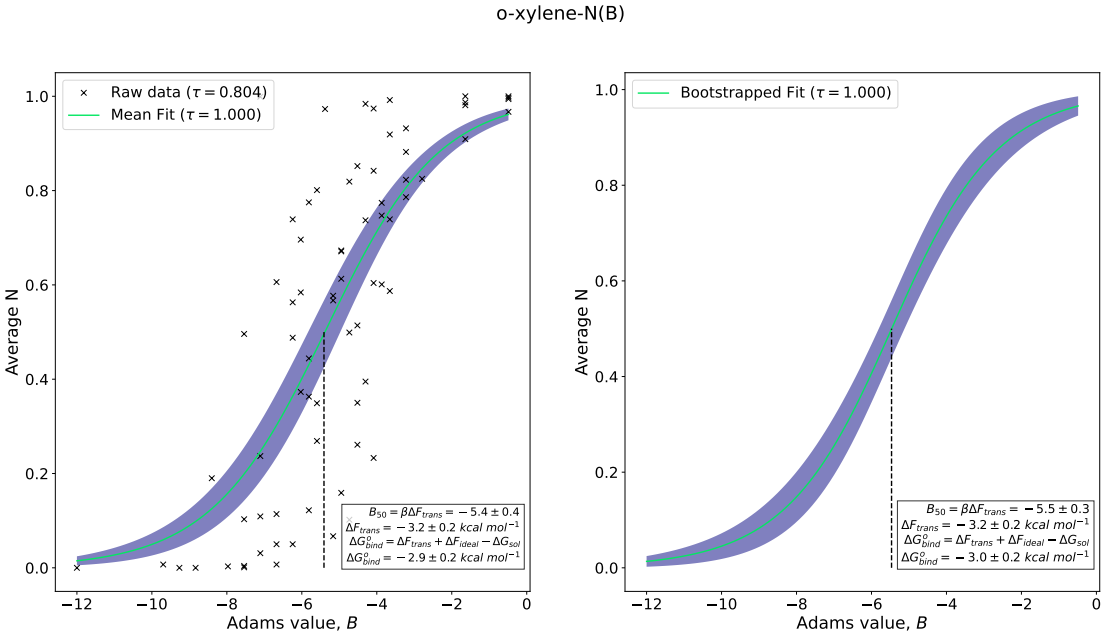
Indene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



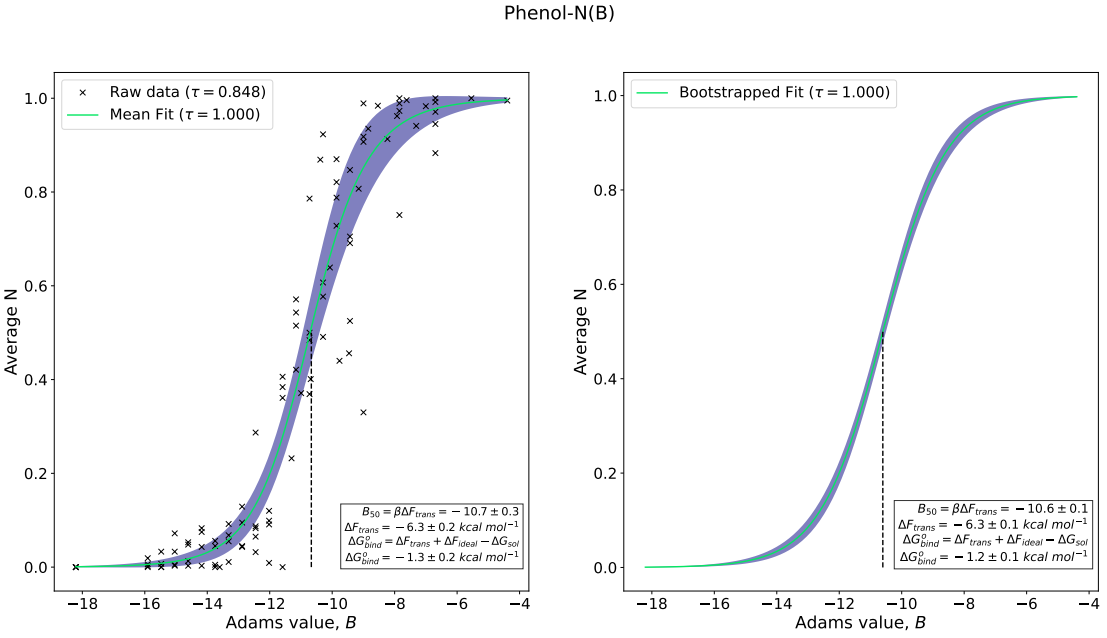
Indole titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



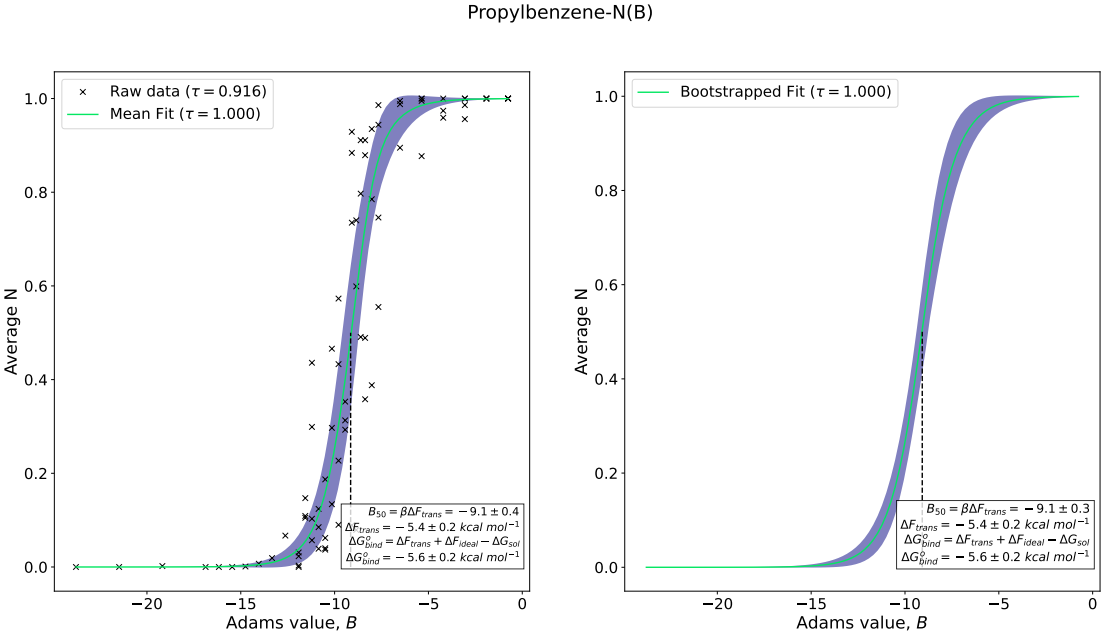
m-xylene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



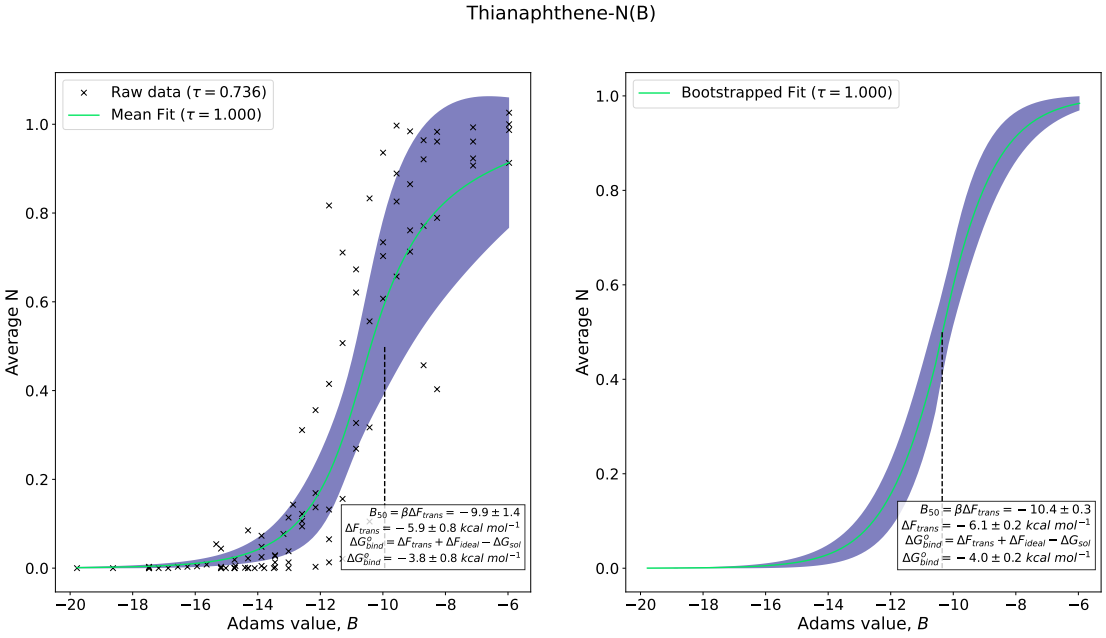
o-xylene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



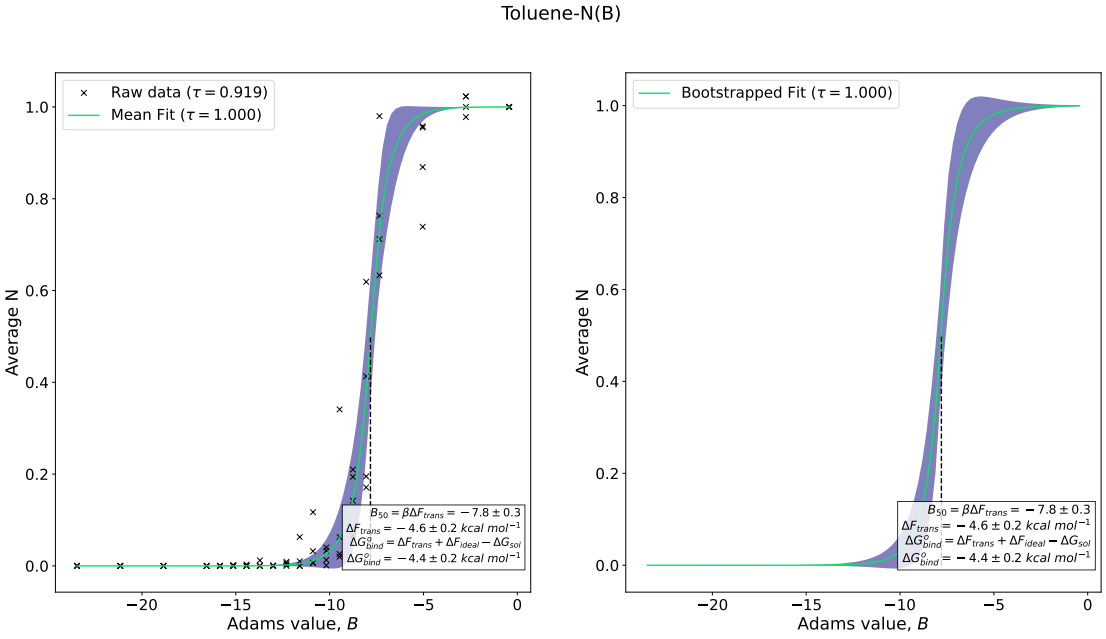
Phenol titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



Propylbenzene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.

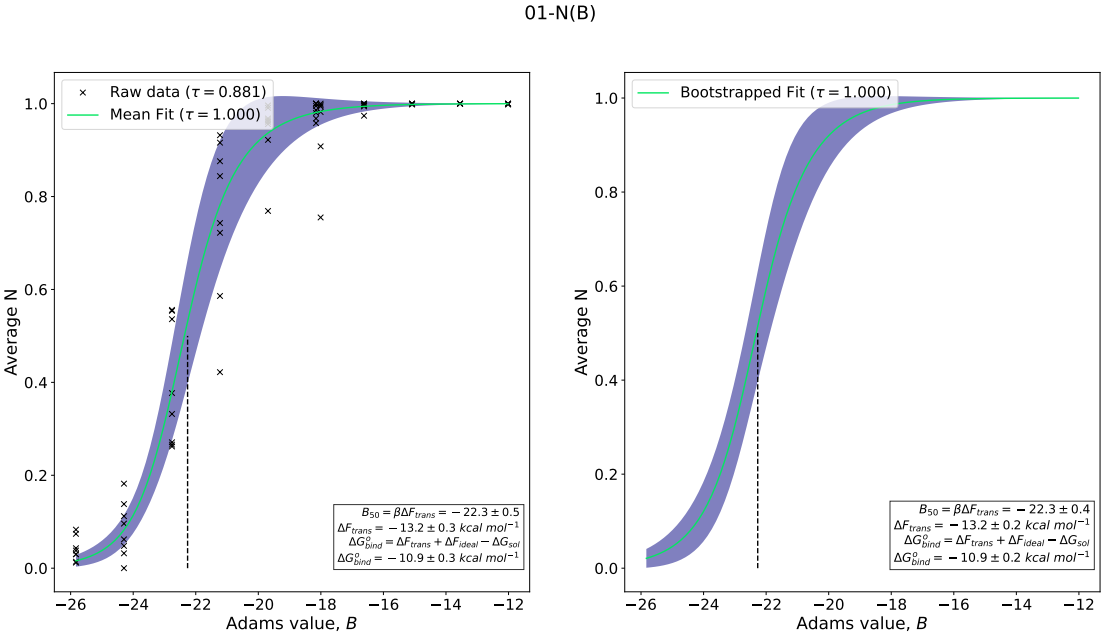


Thianaphthene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.

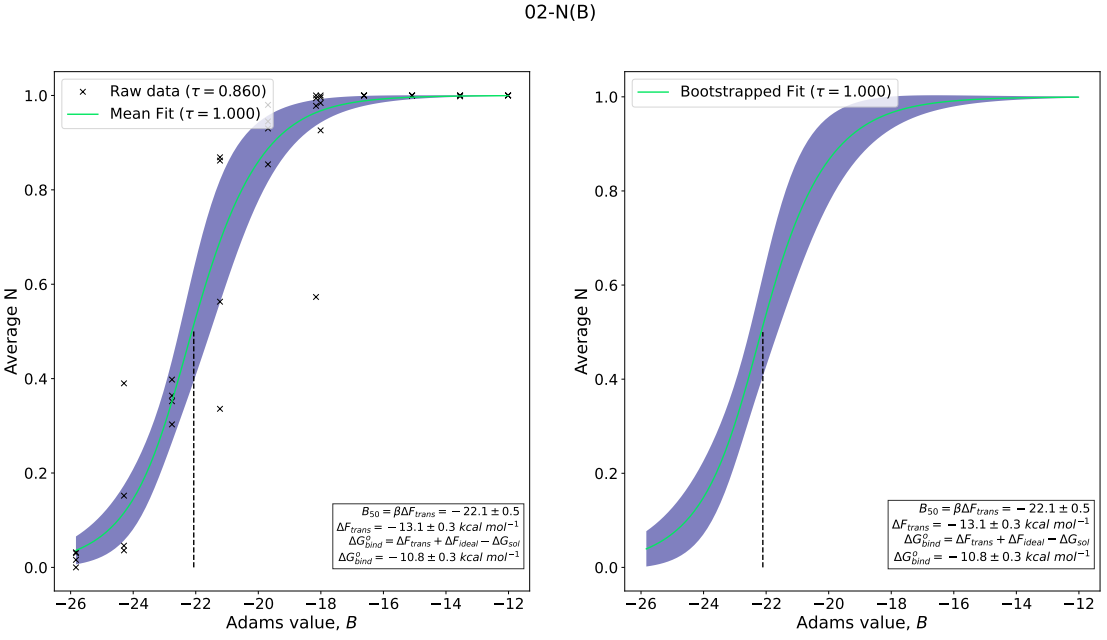


Toluene titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.

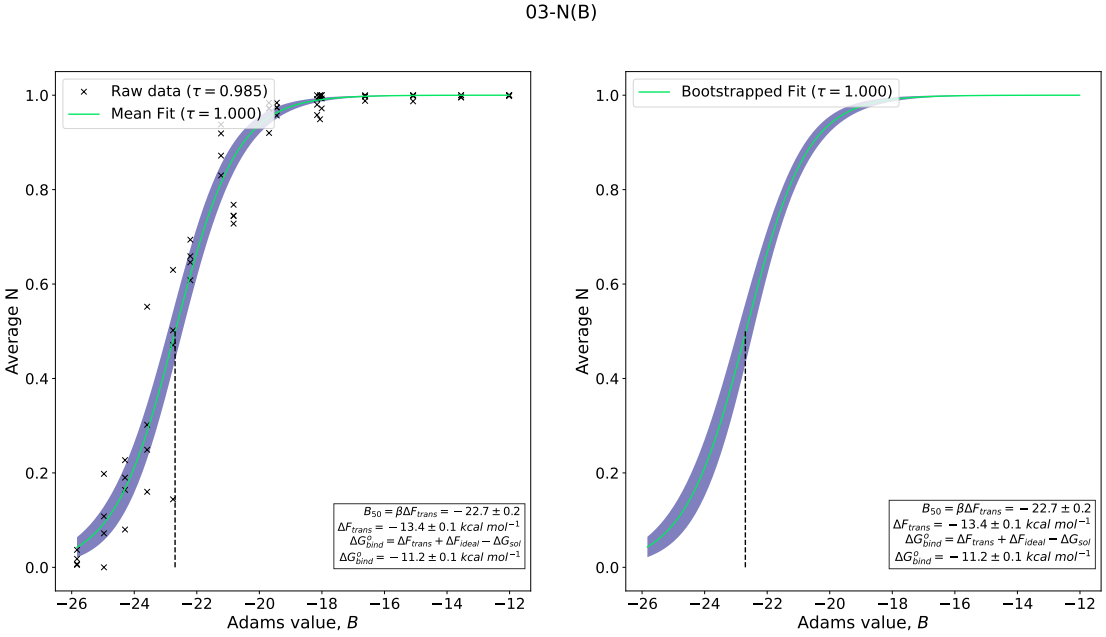
A.3 MUP1 Titrations



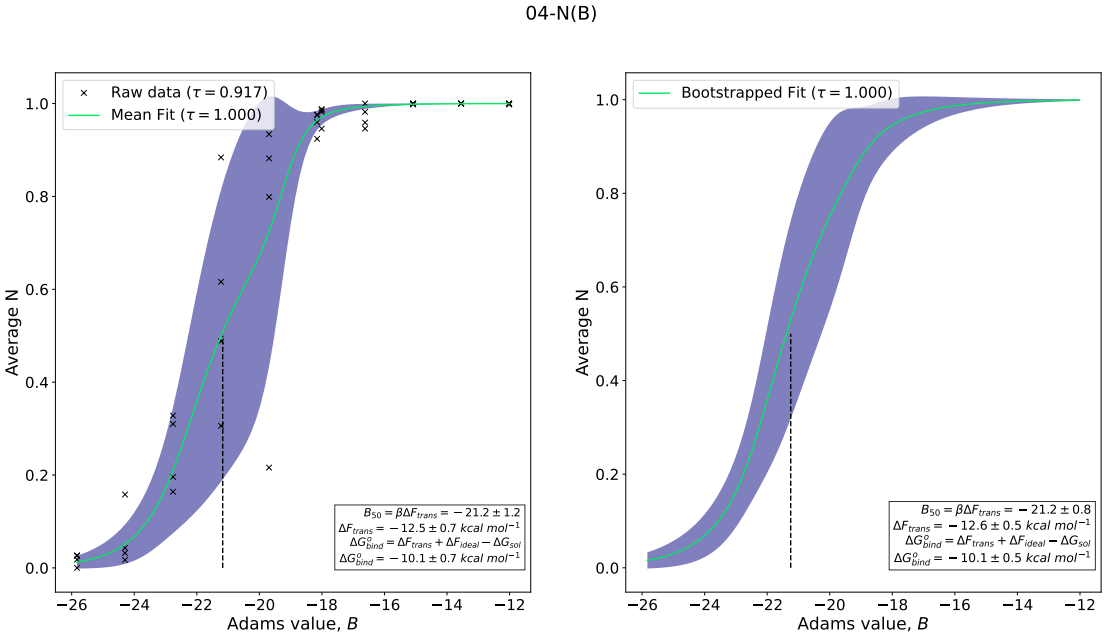
01 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



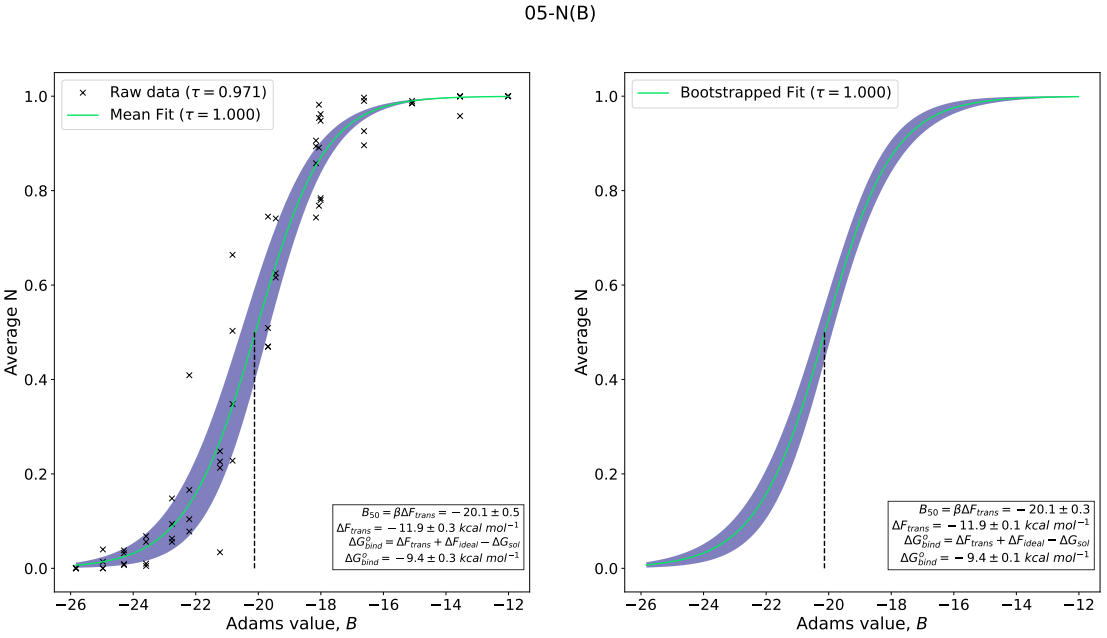
02 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



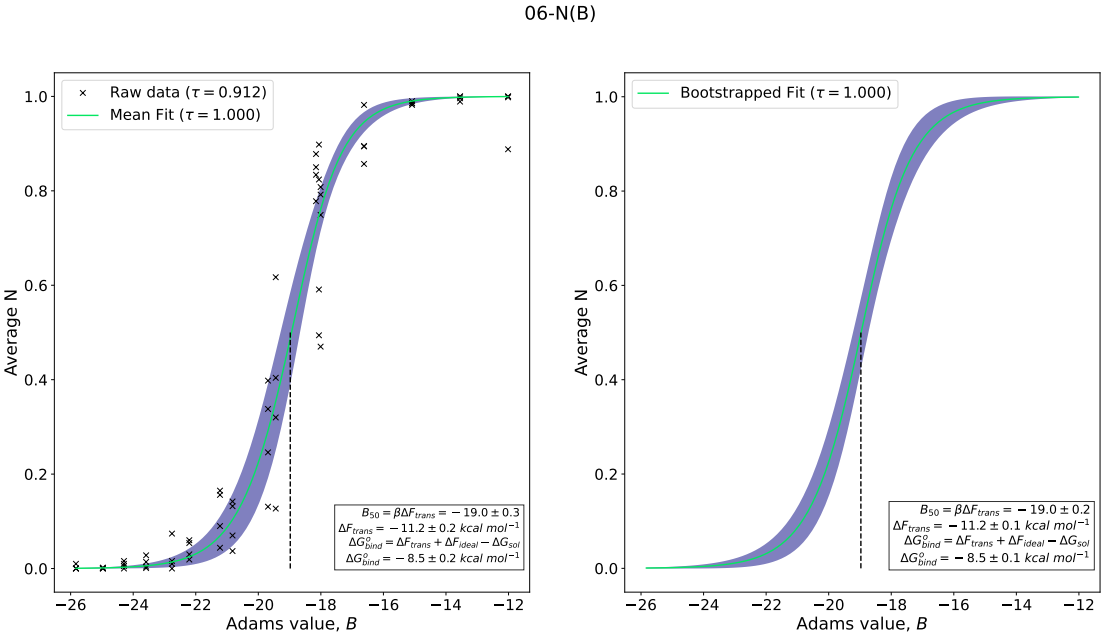
03 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



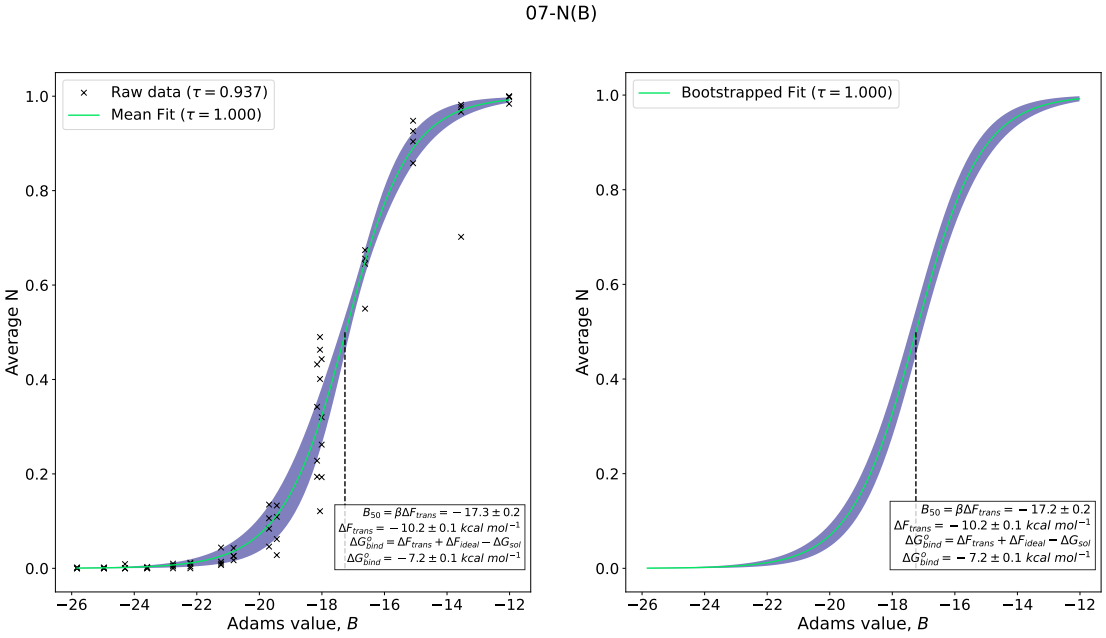
04 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



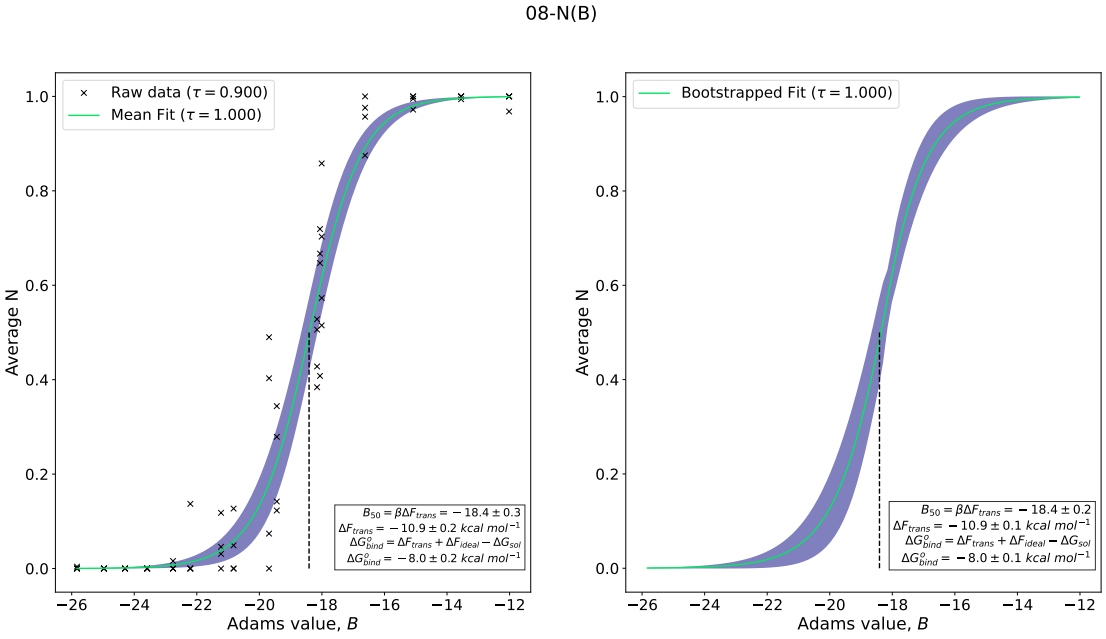
05 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



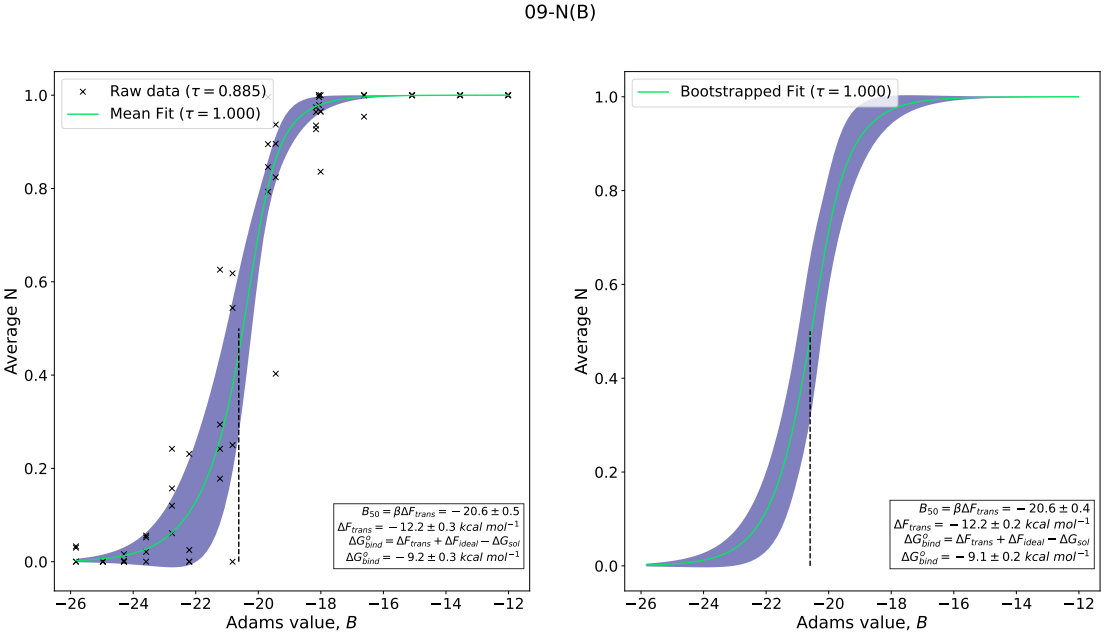
06 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



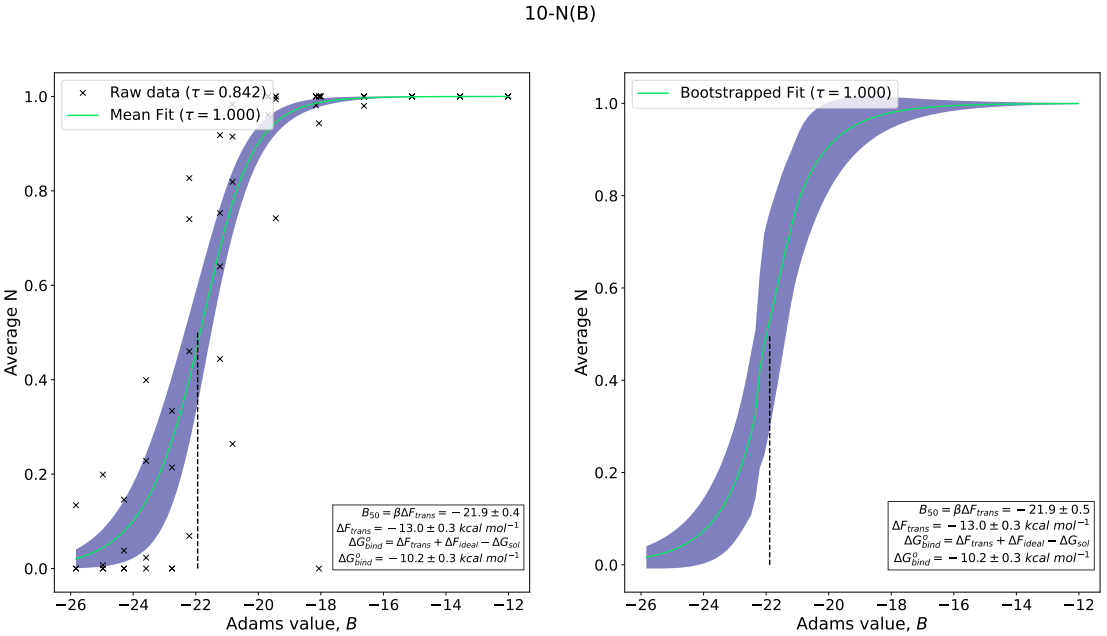
07 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



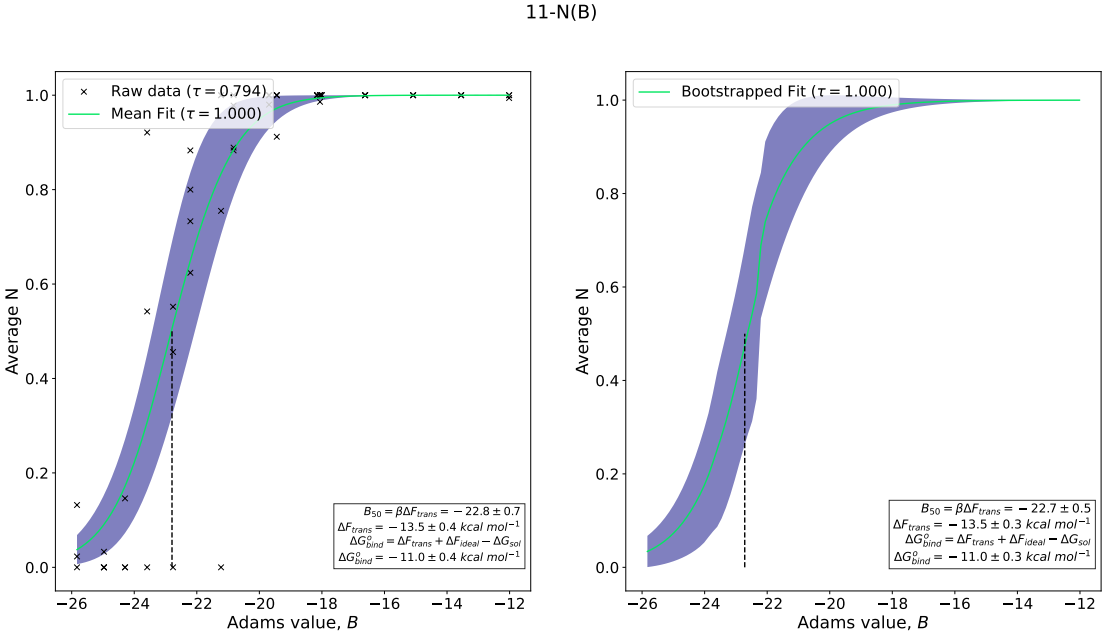
08 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



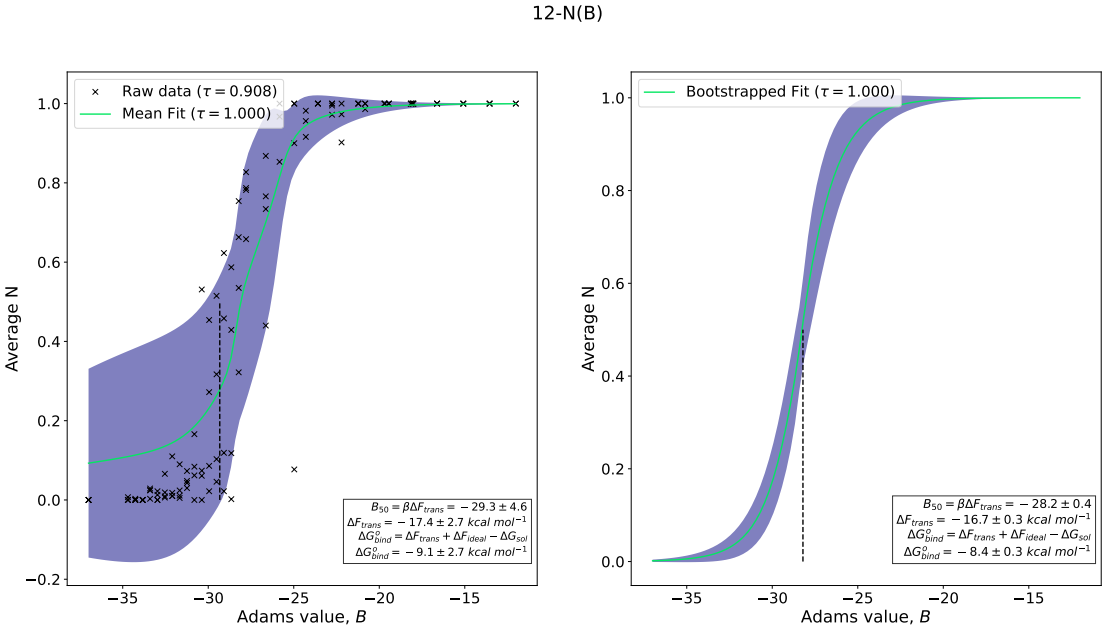
09 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



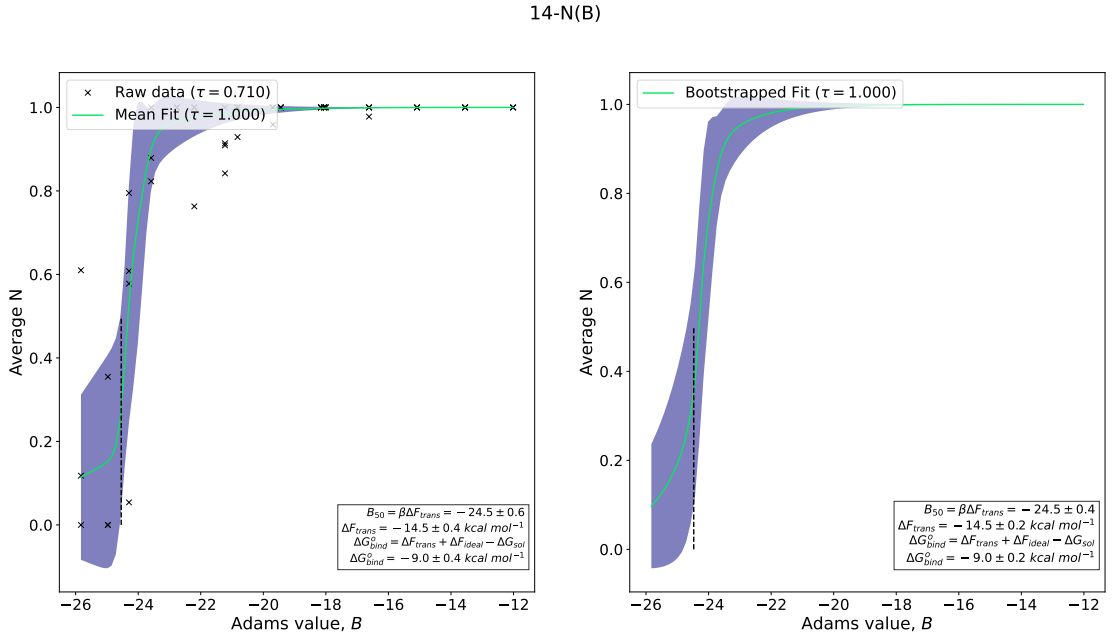
10 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



11 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



12 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.



14 titration curve. The shaded region, and errors, represent one standard deviation. Data is bootstrapped 1000 times.

References

[1] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet and K. A. Dill, *Journal of Molecular Biology*, 2007, **371**, 1118–1134.

[2] M. W. Y. Southey and M. Brunavs, *Frontiers in Drug Discovery*, 2023, **3**, year.

[3] C. R. Ganellin, R. Jefferis and S. M. Roberts, *Introduction to Biological and Small Molecule Drug Research and Development: Theory and Case Studies*, Academic Press, 2013.

[4] N. Berdigaliyev, and M. Aljofan, *Future Medicinal Chemistry*, 2020, **12**, 939–947.

[5] *Statistics | DrugBank Online*, `https://go.drugbank.com/stats`.

[6] A. Mullard, *Nature Reviews Drug Discovery*, 2025, **24**, 75–82.

[7] A. C. Anderson, *Chemistry & Biology*, 2003, **10**, 787–797.

[8] M. Batool, B. Ahmad and S. Choi, *International Journal of Molecular Sciences*, 2019, **20**, 2783.

[9] E. S. Lander *et al.*, *Nature*, 2001, **409**, 860–921.

[10] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, 2000, **28**, 235–242.

[11] M. O'Reilly, A. Cleasby, T. G. Davies, R. J. Hall, R. F. Ludlow, C. W. Murray, D. Tisi and H. Jhoti, *Drug Discovery Today*, 2019, **24**, 1081–1086.

[12] S. W. Kaldor *et al.*, *Journal of Medicinal Chemistry*, 1997, **40**, 3979–3985.

[13] J. N. Varghese, *Drug Development Research*, 1999, **46**, 176–196.

[14] E. E. Rutenber and R. M. Stroud, *Structure*, 1996, **4**, 1317–1324.

[15] T. Schindler, W. Bornmann, P. Pellicena, W. T. Miller, B. Clarkson and J. Kuriyan, *Science*, 2000, **289**, 1938–1942.

[16] *Drug Discovery Today*, 2013, **18**, 265–271.

[17] W. M. Pardridge, *NeuroRX*, 2005, **2**, 3–14.

[18] *Acta Pharmaceutica Sinica B*, 2022, **12**, 3049–3062.

[19] J. H. Van Drie, *Journal of Computer-Aided Molecular Design*, 2007, **21**, 591–601.

[20] J. Bajorath, *Computer-aided drug discovery*, 2015, https://f1000research.com/articles/4-630.

[21] D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke and H. Jhoti, *Nature Reviews Drug Discovery*, 2016, **15**, 605–619.

[22] I. V. Hinkson, B. Madej and E. A. Stahlberg, *Frontiers in Pharmacology*, 2020, **11**, 770.

[23] H. Dowden and J. Munro, *Nature Reviews Drug Discovery*, 2019, **18**, 495–496.

[24] Y. Bian and X.-Q. S. Xie, *The AAPS Journal*, 2018, **20**, 59.

[25] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg and A. L. Schacht, *Nature Reviews Drug Discovery*, 2010, **9**, 203–214.

[26] J. W. Scannell, A. Blanckley, H. Boldon and B. Warrington, *Nature Reviews Drug Discovery*, 2012, **11**, 191–200.

[27] R. K. Harrison, *Nature Reviews Drug Discovery*, 2016, **15**, 817–818.

[28] T. Takebe, R. Imai and S. Ono, *Clinical and Translational Science*, 2018, **11**, 597–606.

[29] G. K. Kiriiri, P. M. Njogu and A. N. Mwangi, *Future Journal of Pharmaceutical Sciences*, 2020, **6**, 27.

[30] D. Sun, W. Gao, H. Hu and S. Zhou, *Acta Pharmaceutica Sinica B*, 2022, **12**, 3049–3062.

[31] C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Advanced Drug Delivery Reviews*, 1996, **46**, 3–26.

[32] D. A. Erlanson, R. S. McDowell and T. O'Brien, *Journal of Medicinal Chemistry*, 2004, **47**, 3463–3482.

[33] R. S. Bohacek, C. McMartin and W. C. Guida, *Medicinal Research Reviews*, 1996, **16**, 3–50.

[34] G. Chessari and A. J. Woodhead, *Drug Discovery Today*, 2009, **14**, 668–675.

[35] C. W. Murray and D. C. Rees, *Nature Chemistry*, 2009, **1**, 187–192.

[36] P. Kirsch, A. M. Hartman, A. K. H. Hirsch and M. Empting, *Molecules*, 2019, **24**, 4309.

[37] M. Congreve, R. Carr, C. Murray and H. Jhoti, *Drug Discovery Today*, 2003, **8**, 876–877.

[38] A. L. Hopkins, C. R. Groom and A. Alex, *Drug Discovery Today*, 2004, **9**, 430–431.

[39] M. Congreve, G. Chessari, D. Tisi and A. J. Woodhead, *Journal of Medicinal Chemistry*, 2008, **51**, 3661–3680.

[40] S. Schultes, C. de Graaf, E. E. Haaksma, I. J. de Esch, R. Leurs and O. Krämer, *Drug Discovery Today: Technologies*, 2010, **7**, e157–e162.

[41] A. L. Hopkins, G. M. Keserü, P. D. Leeson, D. C. Rees and C. H. Reynolds, *Nature Reviews Drug Discovery*, 2014, **13**, 105–121.

[42] M. L. Verdonk and D. C. Rees, *ChemMedChem*, 2008, **3**, 1179–1180.

[43] P. D. Leeson and B. Springthorpe, *Nature Reviews. Drug Discovery*, 2007, **6**, 881–890.

[44] D. Erlanson, *Practical Fragments: Fragments in the clinic: 2024 edition*, 2024, https://practicalfragments.blogspot.com/2024/02/fragments-in-clinic-2024-edition.html.

[45] A. J. Woodhead, D. A. Erlanson, I. J. P. de Esch, R. S. Holvey, W. Jahnke and P. Pathuri, *Journal of Medicinal Chemistry*, 2024, **67**, 2287–2304.

[46] D. G. Brown, *Journal of Medicinal Chemistry*, 2023, **66**, 7101–7139.

[47] K. N. Allen, C. R. Bellamacina, X. Ding, C. J. Jeffery, C. Mattos, G. A. Petsko and D. Ringe, *The Journal of Physical Chemistry*, 1996, **100**, 2605–2611.

[48] A. B. Keeley *et al.*, *Journal of Medicinal Chemistry*, 2024, **67**, 572–585.

[49] H. Jhoti, A. Cleasby, M. Verdonk and G. Williams, *Current Opinion in Chemical Biology*, 2007, **11**, 485–493.

[50] J. Robson-Tull, *Bioscience Horizons: The International Journal of Student Research*, 2018, **11**, hzy015.

[51] G. Holdgate, S. Geschwindner, A. Breeze, G. Davies, N. Colclough, D. Temesi and L. Ward, in *Protein-Ligand Interactions: Methods and Applications*, ed. M. A. Williams and T. Daviter, Humana Press, Totowa, NJ, 2013, pp. 327–355.

[52] N. M. Pearce *et al.*, *Nature Communications*, 2017, **8**, 15123.

[53] A. C. Gibbs, M. C. Abad, X. Zhang, B. A. Tounge, F. A. Lewandowski, G. T. Struble, W. Sun, Z. Sui and L. C. Kuo, *Journal of Medicinal Chemistry*, 2010, **53**, 7979–7991.

[54] D. Hahn *et al.*, *Living Journal of Computational Molecular Science*, 2022, **4**, 1497–1497.

[55] A. Wlodawer, W. Minor, Z. Dauter and M. Jaskolski, *The FEBS Journal*, 2008, **275**, 1–21.

[56] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Research*, 2000, **28**, 235–242.

[57] A. M. Davis, S. J. Teague and G. J. Kleywegt, *Angewandte Chemie International Edition*, 2003, **42**, 2718–2736.

[58] E. Nogales and S. Scheres, *Molecular Cell*, 2015, **58**, 677–689.

[59] Y. Cheng, *Cell*, 2015, **161**, 450–457.

[60] E. H. Egelman, *Biophysical Journal*, 2016, **110**, 1008–1012.

[61] W. Kühlbrandt, *Science*, 2014, **343**, 1443–1444.

[62] H.-W. Wang and J.-W. Wang, *Protein Science*, 2017, **26**, 32–39.

[63] M. Saur, M. J. Hartshorn, J. Dong, J. Reeks, G. Bunkoczi, H. Jhoti and P. A. Williams, *Drug Discovery Today*, 2020, **25**, 485–490.

[64] S. B. Shuker, P. J. Hajduk, R. P. Meadows and S. W. Fesik, *Science*, 1996, **274**, 1531–1534.

[65] M. J. Harner, A. O. Frank and S. W. Fesik, *Journal of Biomolecular NMR*, 2013, **56**, 65–75.

[66] T. B. Almeida, S. Panova and R. Walser, *SLAS Discovery*, 2021, **26**, 1020–1028.

[67] M. P. Williamson, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2013, **73**, 1–16.

[68] L. Fielding, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2007, **51**, 219–242.

[69] O. Cala and I. Krimm, *Journal of Medicinal Chemistry*, 2015, **58**, 8739–8742.

[70] A. Viegas, J. Manso, F. L. Nobrega and E. J. Cabrita, *Journal of Chemical Education*, 2011, **88**, 990–994.

[71] S. Leavitt and E. Freire, *Current Opinion in Structural Biology*, 2001, **11**, 560–566.

[72] I. Navratilova and A. L. Hopkins, *ACS Medicinal Chemistry Letters*, 2010, **1**, 44–48.

[73] T. Neumann, H.-D. Junker, K. Schmidt and R. Sekul, *http://www.eurekaselect.com*.

[74] G. Senisterra, I. Chau and M. Vedadi, *ASSAY and Drug Development Technologies*, 2012, **10**, 128–136.

[75] M. Bissaro, M. Sturlese and S. Moro, *Drug Discovery Today*, 2020, **25**, 1693–1701.

[76] M. T. Muhammed and E. Aki-Yalcin, *Chemical Biology & Drug Design*, 2019, **93**, 12–20.

[77] A. Fiser, R. K. Do and A. Sali, *Protein Science: A Publication of the Protein Society*, 2000, **9**, 1753–1773.

[78] A. Waterhouse *et al.*, *Nucleic Acids Research*, 2018, **46**, W296–W303.

[79] M. Baek *et al.*, *Science*, 2021, **373**, 871–876.

[80] J. Abramson *et al.*, *Nature*, 2024, **630**, 493–500.

[81] M. Baek and D. Baker, *Nature Methods*, 2022, **19**, 13–14.

[82] J. Jumper *et al.*, *Nature*, 2021, **596**, 583–589.

[83] T. Saldaño *et al.*, *Bioinformatics*, 2022, **38**, 2742–2748.

[84] D. Sala, F. Engelberger, H. Mchaourab and J. Meiler, *Current Opinion in Structural Biology*, 2023, **81**, 102645.

[85] N. Raisinghani, V. Parikh, B. Foley and G. Verkhivker, *International Journal of Molecular Sciences*, 2024, **25**, 12968.

[86] P. J. Goodford, *Journal of Medicinal Chemistry*, 1985, **28**, 849–857.

[87] A. Miranker and M. Karplus, *Proteins: Structure, Function, and Bioinformatics*, 1991, **11**, 29–34.

[88] R. Brenke, D. Kozakov, G.-Y. Chuang, D. Beglov, D. Hall, M. R. Landon, C. Mattos and S. Vajda, *Bioinformatics*, 2009, **25**, 621–627.

[89] P. Ghanakota and H. A. Carlson, *Journal of Medicinal Chemistry*, 2016, **59**, 10383–10399.

[90] D. Alvarez-Garcia and X. Barril, *Journal of Chemical Theory and Computation*, 2014, **10**, 2608–2614.

[91] E. P. Raman, W. Yu, O. Guvench and A. D. MacKerell, *Journal of Chemical Information and Modeling*, 2011, **51**, 877–896.

[92] K. W. Lexa and H. A. Carlson, *Journal of the American Chemical Society*, 2011, **133**, 200–202.

[93] P. Ghanakota, H. Van Vlijmen, W. Sherman and T. Beuming, *Journal of Chemical Information and Modeling*, 2018, **58**, 784–793.

[94] S. K. Lakkaraju, E. P. Raman, W. Yu and A. D. J. MacKerell, *Journal of Chemical Theory and Computation*, 2014, **10**, 2281–2290.

[95] J. Tze-Yang Ng and Y. S. Tan, *Journal of Chemical Theory and Computation*, 2022, **18**, 1969–1981.

[96] R. D. Smith and H. A. Carlson, *Journal of Chemical Information and Modeling*, 2021, **61**, 1287–1299.

[97] V. Oleinikovas, G. Saladino, B. P. Cossins and F. L. Gervasio, *Journal of the American Chemical Society*, 2016, **138**, 14257–14263.

[98] F. Comitani and F. L. Gervasio, *Journal of Chemical Theory and Computation*, 2018, **14**, 3321–3331.

[99] Y. Sugita and Y. Okamoto, *Chemical Physics Letters*, 1999, **314**, 141–151.

[100] A. Borsatto, E. Gianquinto, V. Rizzi and F. L. Gervasio, *Journal of Chemical Theory and Computation*, 2024, **20**, 3335–3348.

[101] N. S. Pagadala, K. Syed and J. Tuszynski, *Biophysical Reviews*, 2017, **9**, 91–102.

[102] L. Chachulski and B. Windshügel, *Journal of Chemical Information and Modeling*, 2020, **60**, 6544–6554.

[103] D. M. Lorber and B. K. Shoichet, *Protein Science: A Publication of the Protein Society*, 1998, **7**, 938–950.

[104] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor, *Proteins: Structure, Function, and Bioinformatics*, 2003, **52**, 609–623.

[105] O. Trott and A. J. Olson, *Journal of Computational Chemistry*, 2010, **31**, 455–461.

[106] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard and J. L. Banks, *Journal of Medicinal Chemistry*, 2004, **47**, 1750–1759.

[107] J. Meiler and D. Baker, *Proteins*, 2006, **65**, 538–548.

[108] G. L. Warren *et al.*, *Journal of Medicinal Chemistry*, 2006, **49**, 5912–5931.

[109] S. C. Gill, N. M. Lim, P. B. Grinaway, A. S. Rustenburg, J. Fass, G. A. Ross, J. D. Chodera and D. L. Mobley, *The Journal of Physical Chemistry B*, 2018, **122**, 5579–5598.

[110] N. M. Lim, M. Osato, G. L. Warren and D. L. Mobley, *Journal of Chemical Theory and Computation*, 2020, **16**, 2778–2794.

[111] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh and J. D. Chodera, *Proceedings of the National Academy of Sciences*, 2011, **108**, E1009–E1018.

[112] M. Suruzhon, M. S. Bodnarchuk, A. Ciancetta, I. D. Wall and J. W. Essex, *Journal of Chemical Theory and Computation*, 2022, **18**, 3894–3910.

[113] J. Mondal, N. Ahalawat, S. Pandit, L. E. Kay and P. Vallurupalli, *PLOS Computational Biology*, 2018, **14**, e1006180.

[114] A. C. Pan, H. Xu, T. Palpant and D. E. Shaw, *Journal of Chemical Theory and Computation*, 2017, **13**, 3372–3377.

[115] P. Macek, P. Novák, L. Žídek and V. Sklenář, *The Journal of Physical Chemistry B*, 2007, **111**, 5731–5739.

[116] A. J. Clark, P. Tiwary, K. Borrelli, S. Feng, E. B. Miller, R. Abel, R. A. Friesner and B. J. Berne, *Journal of Chemical Theory and Computation*, 2016, **12**, 2990–2998.

[117] S. Genheden and U. Ryde, *Expert Opinion on Drug Discovery*, 2015, **10**, 449–461.

[118] E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J. Z. H. Zhang and T. Hou, *Chemical Reviews*, 2019, **119**, 9478–9508.

[119] F. Godschalk, S. Genheden, P. Söderhjelm and U. Ryde, *Physical Chemistry Chemical Physics*, 2013, **15**, 7731–7739.

[120] T. Tuccinardi, *Expert Opinion on Drug Discovery*, 2021, **16**, 1233–1237.

[121] Z. Cournia, B. Allen and W. Sherman, *Journal of Chemical Information and Modeling*, 2017, **57**, 2911–2937.

[122] L. Wang *et al.*, *Journal of the American Chemical Society*, 2015, **137**, 2695–2703.

[123] T. B. Steinbrecher, M. Dahlgren, D. Cappel, T. Lin, L. Wang, G. Krilov, R. Abel, R. Friesner and W. Sherman, *Journal of Chemical Information and Modeling*, 2015, **55**, 2411–2420.

[124] G. A. Ross, C. Lu, G. Scarabelli, S. K. Albanese, E. Houang, R. Abel, E. D. Harder and L. Wang, *Communications Chemistry*, 2023, **6**, 1–12.

[125] Z. Cournia, C. Chipot, B. Roux, D. M. York and W. Sherman, in *Free Energy Methods in Drug Discovery: Current State and Future Directions*, American Chemical Society, 2021, vol. 1397 of ACS Symposium Series, pp. 1–38.

[126] I. Alibay, A. Magarkar, D. Seeliger and P. C. Biggin, *Communications Chemistry*, 2022, **5**, 1–13.

[127] S. Boresch, F. Tettinger, M. Leitgeb and M. Karplus, *The Journal of Physical Chemistry B*, 2003, **107**, 9535–9551.

[128] D. L. Mobley, J. D. Chodera and K. A. Dill, *The Journal of chemical physics*, 2006, **125**, 084902.

[129] R. J. Gowers, I. Alibay, D. W. Swenson, M. M. Henry, B. Ries, H. M. Baumann and J. R. B. Eastwood, *The Open Free Energy library*, 2023, https://zenodo.org/records/8344248.

[130] M. Suruzhon, T. Senapathi, M. S. Bodnarchuk, R. Viner, I. D. Wall, C. B. Barnett, K. J. Naidoo and J. W. Essex, *Journal of Chemical Information and Modeling*, 2020, **60**, 1917–1921.

[131] V. Gapsys, S. Michielssens, D. Seeliger and B. L. de Groot, *Journal of Computational Chemistry*, 2015, **36**, 348–354.

[132] K. H. Burley, S. C. Gill, N. M. Lim and D. L. Mobley, *Journal of chemical theory and computation*, 2019, **15**, 1848–1862.

[133] M. Suruzhon, M. L. Samways and J. W. Essex, in *Free Energy Methods in Drug Discovery: Current State and Future Directions*, American Chemical Society, 2021, vol. 1397 of ACS Symposium Series, pp. 109–125.

[134] M. Suruzhon, M. S. Bodnarchuk, A. Ciancetta, R. Viner, I. D. Wall and J. W. Essex, *Journal of Chemical Theory and Computation*, 2021, **17**, 1806–1821.

[135] H. E. Bruce Macdonald, C. Cave-Ayland, G. A. Ross and J. W. Essex, *Journal of Chemical Theory and Computation*, 2018, **14**, 6586–6597.

[136] J. Thompson, W. P. Walters, J. A. Feng, N. A. Pabon, H. Xu, M. Maser, B. B. Goldman, D. Moustakas, M. Schmidt and F. York, *Artificial Intelligence in the Life Sciences*, 2022, **2**, 100050.

[137] Y. Khalak, G. Tresadern, D. F. Hahn, B. L. de Groot and V. Gapsys, *Journal of Chemical Theory and Computation*, 2022, **18**, 6259–6270.

[138] P. Ghanakota, P. H. Bos, K. D. Konze, J. Staker, G. Marques, K. Marshall, K. Leswing, R. Abel and S. Bhat, *Journal of Chemical Information and Modeling*, 2020, **60**, 4311–4325.

[139] R. Gorantla, A. Kubincová, B. Suutari, B. P. Cossins and A. S. J. S. Mey, *Journal of Chemical Information and Modeling*, 2024, **64**, 1955–1965.

[140] D. Adams, *Molecular Physics*, 1974, **28**, 1241–1252.

[141] H.-J. Woo, A. R. Dinner and B. Roux, *The Journal of Chemical Physics*, 2004, **121**, 6392–6400.

[142] M. Mezei, *Molecular Physics*, 1980, **40**, 901–906.

[143] G. A. Ross, M. S. Bodnarchuk and J. W. Essex, *Journal of the American Chemical Society*, 2015, **137**, 14930–14943.

[144] M. Clark, F. Guarnieri, I. Shkurko and J. Wiseman, *Journal of Chemical Information and Modeling*, 2006, **46**, 231–242.

[145] M. S. Bodnarchuk, M. J. Packer and A. Haywood, *ACS Medicinal Chemistry Letters*, 2020, **11**, 77–82.

[146] M. L. Samways, H. E. Bruce Macdonald, R. D. Taylor and J. W. Essex, *Journal of Chemical Information and Modeling*, 2023, **63**, 387–396.

[147] M. L. Samways, H. E. Bruce Macdonald and J. W. Essex, *Journal of Chemical Information and Modeling*, 2020, **60**, 4436–4441.

[148] G. A. Ross, E. Russell, Y. Deng, C. Lu, E. D. Harder, R. Abel and L. Wang, *Journal of Chemical Theory and Computation*, 2020, **16**, 6061–6076.

[149] M. Samways, *phd*, University of Southampton, 2021.

[150] O. J. Melling, M. L. Samways, Y. Ge, D. L. Mobley and J. W. Essex, *Journal of Chemical Theory and Computation*, 2023, **19**, 1050–1062.

[151] G. A. Ross, H. E. Bruce Macdonald, C. Cave-Ayland, A. I. Cabedo Martinez and J. W. Essex, *Journal of Chemical Theory and Computation*, 2017, **13**, 6373–6381.

[152] P. Eastman *et al.*, *PLoS Computational Biology*, 2017, **13**, e1005659.

[153] G. Duarte Ramos Matos, D. Y. Kyu, H. H. Loeffler, J. D. Chodera, M. R. Shirts and D. L. Mobley, *Journal of Chemical & Engineering Data*, 2017, **62**, 1559–1569.

[154] A. R. Leach, *Molecular modelling: principles and applications*, Prentice Hall, Harlow, England ; New York, 2nd edn., 2001.

[155] C. W. Hopkins, S. Le Grand, R. C. Walker and A. E. Roitberg, *Journal of Chemical Theory and Computation*, 2015, **11**, 1864–1874.

[156] S. Boothroyd *et al.*, *Journal of Chemical Theory and Computation*, 2023, **19**, 3251–3275.

[157] J. T. Horton, S. Boothroyd, P. K. Behara, D. L. Mobley and D. J. Cole, *Digital Discovery*, 2023, **2**, 1178–1187.

[158] Y. Wang *et al.*, *Chemical Science*, 2022, **13**, 12016–12033.

[159] K. Takaba *et al.*, *Machine-learned molecular mechanics force field for the simulation of protein-ligand systems and beyond*, 2023, http://arxiv.org/abs/2307.07085, arXiv:2307.07085.

[160] J. W. Ponder *et al.*, *The Journal of Physical Chemistry B*, 2010, **114**, 2549–2564.

[161] P. Eastman *et al.*, *The Journal of Physical Chemistry B*, 2024, **128**, 109–116.

[162] H. A. Lorentz, *Annalen der Physik*, 1881, **248**, 127–136.

[163] A. Rahman, *Physical Review*, 1964, **136**, A405–A411.

[164] L. Verlet, *Physical Review*, 1967, **159**, 98–103.

[165] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *The Journal of Chemical Physics*, 1982, **76**, 637–649.

[166] W. F. V. Gunsteren and H. J. C. Berendsen, *Molecular Simulation*, 1988, **1**, 173–185.

[167] D. Beeman, *Journal of Computational Physics*, 1976, **20**, 130–139.

[168] J. A. Izaguirre, D. P. Catarello, J. M. Wozniak and R. D. Skeel, *The Journal of Chemical Physics*, 2001, **114**, 2090–2098.

[169] G. Bussi, D. Donadio and M. Parrinello, *The Journal of Chemical Physics*, 2007, **126**, 014101.

[170] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *The Journal of Chemical Physics*, 1984, **81**, 3684–3690.

[171] E. Braun, S. M. Moosavi and B. Smit, *Journal of Chemical Theory and Computation*, 2018, **14**, 5262–5272.

[172] H. C. Andersen, *The Journal of Chemical Physics*, 1980, **72**, 2384–2393.

[173] S. Nosé, *Molecular Physics*, 1984, **52**, 255–268.

[174] W. G. Hoover and B. L. Holian, *Physics Letters A*, 1996, **211**, 253–257.

[175] B. Leimkuhler and C. Matthews, *Applied Mathematics Research eXpress*, 2013, **2013**, 34–56.

[176] B. Leimkuhler and C. Matthews, in *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, ed. B. Leimkuhler and C. Matthews, Springer International Publishing, Cham, 2015, pp. 261–328.

[177] A. Rojnuckarin, S. Kim and S. Subramaniam, *Proceedings of the National Academy of Sciences*, 1998, **95**, 4288–4292.

[178] J. Fass, D. A. Sivak, G. E. Crooks, K. A. Beauchamp, B. Leimkuhler and J. D. Chodera, *Entropy*, 2018, **20**, 318.

[179] K.-H. Chow and D. M. Ferguson, *Computer Physics Communications*, 1995, **91**, 283–289.

[180] J. Åqvist, P. Wennerström, M. Nervall, S. Bjelic and B. O. Brandsdal, *Chemical Physics Letters*, 2004, **384**, 288–294.

[181] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, Academic Press, San Diego, 2nd edn., 2002.

[182] I. G. Tironi, R. Sperb, P. E. Smith and W. F. van Gunsteren, *The Journal of Chemical Physics*, 1995, **102**, 5451–5459.

[183] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *The Journal of Chemical Physics*, 1995, **103**, 8577–8593.

[184] P. P. Ewald, *Annalen der Physik*, 1921, **369**, 253–287.

[185] A. Y. Toukmaji and J. A. Board, *Computer Physics Communications*, 1996, **95**, 73–92.

[186] Y. Ge, D. F. Hahn and D. L. Mobley, *Journal of Chemical Information and Modeling*, 2021, **61**, 1048–1052.

[187] M. E. Tuckerman, *Statistical mechanics: theory and molecular simulation*, Oxford University Press, Oxford ; New York, 2010.

[188] B. Widom, *The Journal of Chemical Physics*, 1963, **39**, 2808–2812.

[189] M. K. Gilson, J. A. Given, B. L. Bush and J. A. McCammon, *Biophysical Journal*, 1997, **72**, 1047–1069.

[190] C.-E. Chang, M. J. Potter and M. K. Gilson, *The Journal of Physical Chemistry B*, 2003, **107**, 1048–1055.

[191] A. S. J. S. Mey *et al.*, *Living Journal of Computational Molecular Science*, 2020, **2**, 18378.

[192] J. Mondal, N. Ahalawat, S. Pandit, L. E. Kay and P. Vallurupalli, *PLOS Computational Biology*, 2018, **14**, e1006180.

[193] D. E. Hyre, I. Le Trong, E. A. Merritt, J. F. Eccleston, N. M. Green, R. E. Stenkamp and P. S. Stayton, *Protein Science*, 2006, **15**, 459–467.

[194] A. Basavapathruni *et al.*, *Chemical Biology & Drug Design*, 2012, **80**, 971–980.

[195] V. Gapsys, A. Yildirim, M. Aldeghi, Y. Khalak, D. van der Spoel and B. L. de Groot, *Communications Chemistry*, 2021, **4**, 1–13.

[196] Y. Khalak, G. Tresadern, M. Aldeghi, H. M. Baumann, D. L. Mobley, B. L. d. Groot and V. Gapsys, *Chemical Science*, 2021, **12**, 13958–13971.

[197] H. M. Baumann, V. Gapsys, B. L. de Groot and D. L. Mobley, *The Journal of Physical Chemistry B*, 2021, **125**, 4241–4261.

[198] C. Jarzynski, *Physical Review E*, 1997, **56**, 5018–5035.

[199] G. E. Crooks, *Journal of Statistical Physics*, 1998, **90**, 1481–1487.

[200] M. R. Shirts, E. Bair, G. Hooker and V. S. Pande, *Physical Review Letters*, 2003, **91**, 140601.

[201] B. P. Cossins, S. Foucher, C. M. Edge and J. W. Essex, *The Journal of Physical Chemistry B*, 2009, **113**, 5508–5519.

[202] R. W. Zwanzig, *The Journal of Chemical Physics*, 1954, **22**, 1420–1426.

[203] M. R. Shirts and V. S. Pande, *The Journal of Chemical Physics*, 2005, **122**, 144107.

[204] C. Jarzynski, *Physical Review E*, 2006, **73**, 046105.

[205] H. Resat and M. Mezei, *The Journal of Chemical Physics*, 1993, **99**, 6052–6061.

[206] C. H. Bennett, *Journal of Computational Physics*, 1976, **22**, 245–268.

[207] M. R. Shirts and J. D. Chodera, *The Journal of Chemical Physics*, 2008, **129**, year.

[208] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *The Journal of Chemical Physics*, 1953, **21**, 1087–1092.

[209] W. K. Hastings, *Biometrika*, 1970, **57**, 97–109.

[210] S. Sasmal, S. C. Gill, N. M. Lim and D. L. Mobley, *Journal of Chemical Theory and Computation*, 2020, **16**, 1854–1865.

[211] T. D. Bergazin, I. Y. Ben-Shalom, N. M. Lim, S. C. Gill, M. K. Gilson and D. L. Mobley, *Journal of computer-aided molecular design*, 2021, **35**, 167–177.

[212] G. A. Ross, A. S. Rustenburg, P. B. Grinaway, J. Fass and J. D. Chodera, *The Journal of Physical Chemistry B*, 2018, **122**, 5466–5486.

[213] P. Cimermancic *et al.*, *Journal of Molecular Biology*, 2016, **428**, 709–719.

[214] C.-H. Ngan, D. R. Hall, B. Zerbe, L. E. Grove, D. Kozakov and S. Vajda, *Bioinformatics (Oxford, England)*, 2012, **28**, 286–287.

[215] D. Kozakov, L. E. Grove, D. R. Hall, T. Bohnuud, S. E. Mottarella, L. Luo, B. Xia, D. Beglov and S. Vajda, *Nature Protocols*, 2015, **10**, 733–755.

[216] L. E. Grove, D. R. Hall, D. Beglov, S. Vajda and D. Kozakov, *Bioinformatics (Oxford, England)*, 2013, **29**, 1218–1219.

[217] O. Khan, G. Jones, M. Lazou, D. Joseph-McCarthy, D. Kozakov, D. Beglov and S. Vajda, *Journal of Chemical Information and Modeling*, 2024, **64**, 2084–2100.

[218] *Journal of Molecular Biology*, 2022, **434**, 167587.

[219] J. Seco, F. J. Luque and X. Barril, *Journal of Medicinal Chemistry*, 2009, **52**, 2363–2371.

[220] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *Journal of Chemical Theory and Computation*, 2015, **11**, 3696–3713.

[221] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *Journal of Chemical Physics*, 1983, **79**, 926–935.

[222] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157–1174.

[223] A. Jakalian, D. B. Jack and C. I. Bayly, *Journal of Computational Chemistry*, 2002, **23**, 1623–1641.

[224] V. Le Guilloux, P. Schmidtke and P. Tuffery, *BMC Bioinformatics*, 2009, **10**, 168.

[225] K. W. Lexa and H. A. Carlson, *Journal of Chemical Information and Modeling*, 2013, **53**, 391–402.

[226] P. M. U. Ung, P. Ghanakota, S. E. Graham, K. W. Lexa and H. A. Carlson, *Biopolymers*, 2016, **105**, 21–34.

[227] J. L. Thomaston, M. L. Samways, A. Konstantinidi, C. Ma, Y. Hu, H. E. Bruce Macdonald, J. Wang, J. W. Essex, W. F. DeGrado and A. Kolocouris, *Biochemistry*, 2021, **60**, 2471–2482.

[228] J. L. I. Kulp, J. L. J. Kulp, D. L. Pompliano and F. Guarnieri, *Journal of the American Chemical Society*, 2011, **133**, 10740–10743.

[229] N. M. Henriksen and M. K. Gilson, *Journal of Chemical Theory and Computation*, 2017, **13**, 4253–4269.

[230] L. Wickstrom, P. He, E. Gallicchio and R. M. Levy, *Journal of Chemical Theory and Computation*, 2013, **9**, 3136–3150.

[231] H. Zhang, C. Yin, H. Yan and D. van der Spoel, *Journal of Chemical Information and Modeling*, 2016, **56**, 2080–2092.

[232] J. Yin, A. T. Fenley, N. M. Henriksen and M. K. Gilson, *The Journal of Physical Chemistry B*, 2015, **119**, 10145–10155.

[233] M. V. Rekharsky, M. P. Mayhew, R. N. Goldberg, P. D. Ross, Y. Yamashoji and Y. Inoue, *The Journal of Physical Chemistry B*, 1997, **101**, 87–100.

[234] D. L. Mobley and M. K. Gilson, *Annual Review of Biophysics*, 2017, **46**, 531–558.

[235] C. Cézard, X. Trivelli, F. Aubry, F. Djedaïni-Pilard and F.-Y. Dupradeau, *Physical Chemistry Chemical Physics*, 2011, **13**, 15103–15121.

[236] K. N. Kirschner and R. J. Woods, *Proceedings of the National Academy of Sciences*, 2001, **98**, 10541–10545.

[237] K. N. Kirschner and R. J. Woods, *The Journal of Physical Chemistry A*, 2001, **105**, 4150–4155.

[238] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712–725.

[239] C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *The Journal of Physical Chemistry*, 1993, **97**, 10269–10280.

[240] Z. Huai, Z. Shen and Z. Sun, *Journal of Chemical Information and Modeling*, 2021, **61**, 284–297.

[241] T. Suzuki, *Journal of Chemical Information and Computer Sciences*, 2001, **41**, 1266–1273.

[242] S. E. Boyce, D. L. Mobley, G. J. Rocklin, A. P. Graves, K. A. Dill and B. K. Shoichet, *Journal of Molecular Biology*, 2009, **394**, 747–763.

[243] N. M. Lim, L. Wang, R. Abel and D. L. Mobley, *Journal of chemical theory and computation*, 2016, **12**, 4620–4631.

[244] W. Jiang and B. Roux, *Journal of chemical theory and computation*, 2010, **6**, 2559–2565.

[245] J. Wang and Y. Miao, *Journal of Chemical Theory and Computation*, 2023, **19**, 733–745.

[246] R. Malham, S. Johnstone, R. J. Bingham, E. Barratt, S. E. V. Phillips, C. A. Laughton and S. W. Homans, *Journal of the American Chemical Society*, 2005, **127**, 17061–17067.

[247] *PDBFixer*, 2023, `https://github.com/openmm/pdbfixer`, original-date: 2013-08-29T22:29:24Z.

[248] J. Lan *et al.*, *Nature*, 2020, **581**, 215–220.

[249] I. S. Joung and T. E. I. Cheatham, *The Journal of Physical Chemistry B*, 2008, **112**, 9020–9041.

[250] S. Träger, G. Tamò, D. Aydin, G. Fonti, M. Audagnotto and M. Dal Peraro, *Bioinformatics*, 2021, **37**, 921–928.

[251] M. K. Gilson and K. K. Irikura, *The Journal of Physical Chemistry B*, 2010, **114**, 16304–16317.

[252] A. Morton, W. A. Baase and B. W. Matthews, *Biochemistry*, 1995, **34**, 8564–8575.

[253] S. R. Kimura, H. P. Hu, A. M. Ruvinsky, W. Sherman and A. D. Favia, *Journal of Chemical Information and Modeling*, 2017, **57**, 1388–1401.

[254] S. K. Lakkaraju, W. Yu, E. P. Raman, A. V. Hershfeld, L. Fang, D. A. Deshpande and A. D. MacKerell, *Journal of Chemical Information and Modeling*, 2015, **55**, 700–708.

[255] J. Liang, C. Woodward and H. Edelsbrunner, *Protein Science*, 1998, **7**, 1884–1897.

[256] D. Schmidt, M. Boehm, C. L. McClendon, R. Torella and H. Gohlke, *Journal of Chemical Theory and Computation*, 2019, **15**, 3331–3343.

[257] D. Hamelberg, J. Mongan and J. A. McCammon, *The Journal of Chemical Physics*, 2004, **120**, 11919–11929.

[258] K. W. Lexa, G. B. Goh and H. A. Carlson, *Journal of Chemical Information and Modeling*, 2014, **54**, 2190–2199.

[259] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon and N. A. Baker, *Nucleic Acids Research*, 2004, **32**, W665–W667.

[260] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski and J. H. Jensen, *Journal of Chemical Theory and Computation*, 2011, **7**, 525–537.

[261] L. Martínez, R. Andrade, E. G. Birgin and J. M. Martínez, *Journal of Computational Chemistry*, 2009, **30**, 2157–2164.

[262] D. A. Case *et al.*, *Journal of Chemical Information and Modeling*, 2023, **63**, 6183–6191.

[263] W. Humphrey, A. Dalke and K. Schulten, *Journal of Molecular Graphics*, 1996, **14**, 33–38.

[264] Z. Wang, G. Zhu, Q. Huang, M. Qian, M. Shao, Y. Jia and Y. Tang, *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 1998, **1384**, 335–344.

[265] P. Ghanakota and H. A. Carlson, *The Journal of Physical Chemistry B*, 2016, **120**, 8685–8695.

[266] J. R. Stephenson Clarke, L. R. Douglas, P. J. Duriez, D.-I. Balourdas, A. C. Joerger, R. Khadiullina, E. Bulatov and M. G. J. Baud, *ACS Pharmacology & Translational Science*, 2022, **5**, 1169–1180.

[267] A. Hafner, M. L. Bulyk, A. Jambhekar and G. Lahav, *Nature Reviews Molecular Cell Biology*, 2019, **20**, 199–210.

[268] V. Chasov, R. Mirgayazova, E. Zmievskaya, R. Khadiullina, A. Valiullina, J. Stephenson Clarke, A. Rizvanov, M. G. J. Baud and E. Bulatov, *Frontiers in Oncology*, 2020, **10**, year.

[269] S. E. Jackson, in *Molecular Chaperones*, ed. S. Jackson, Springer, Berlin, Heidelberg, 2013, pp. 155–240.

[270] C. N. Nguyen, T. Kurtzman Young and M. K. Gilson, *The Journal of Chemical Physics*, 2012, **137**, 044101.

[271] A. D. MacKerell, S. Jo, S. K. Lakkaraju, C. Lind and W. Yu, *Biochimica et Biophysica Acta (BBA) - General Subjects*, 2020, **1864**, 129519.

[272] Y. Miao, V. A. Feher and J. A. McCammon, *Journal of Chemical Theory and Computation*, 2015, **11**, 3584–3595.

[273] S. C. Turner, *phd*, University of Southampton, 2024.

[274] S. Bloodworth *et al.*, *Angewandte Chemie International Edition*, 2019, **58**, 5038–5043.

[275] M. M. Copeland, H. N. Do, L. Votapka, K. Joshi, J. Wang, R. E. Amaro and Y. Miao, *The Journal of Physical Chemistry B*, 2022, **126**, 5810–5820.