UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Physical Sciences
School of Chemistry

Development and Application of Grand Canonical Methods for Molecular Dynamics Simulations

by

Marley Luke Samways
MChem
ORCiD: 0000-0001-9431-8789

A thesis for the degree of
Doctor of Philosophy

November 2021

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Chemistry

Doctor of Philosophy

Development and Application of Grand Canonical Methods for Molecular Dynamics Simulations

by Marley Luke Samways

The work presented in this thesis focuses on the use of grand canonical Monte Carlo (GCMC) sampling during molecular dynamics (MD) simulations (referred to as GCMC/MD), which is used in this work with the aim of enhancing the sampling of water molecules at buried protein-ligand interfaces. Several developments in both the methodology and implementation of GCMC are presented, as well as insights into the binding of drugs to an influenza protein.

First, a Python module (*grand*) is presented in chapter 3, which was developed during this work to allow GCMC sampling of water molecules to be carried out with the OpenMM software package. This implementation of GCMC was thoroughly tested in terms of reproduction of bulk water densities, as well as a rigorous statistical validation.

In chapter 4, GCMC/MD simulations are applied to the M2 protein, which is an influenza drug target, where water is thought to play a key role in ligand binding. Insights are provided into how water affects the binding of different ligand enantiomers to the M2 channel, as well as the possible role of water networks in the resistance of M2 to some drugs, which may aid in the design of future inhibitors.

In chapter 5, it is shown that nonequilibrium candidate Monte Carlo (NCMC) can be used to drastically increase the acceptance rates of GCMC moves — referred to as grand canonical nonequilibrium candidate Monte Carlo (GCNCMC) — by allowing the environment to relax in response to a proposed water insertion or deletion. Whilst these moves are more expensive, they can be up to five times more efficient than traditional GCMC. In chapter 6, it is shown that this improvement greatly facilitates grand canonical sampling of molecules larger than water, indicating that GCNCMC sampling of molecular fragments could have applications in computer-aided drug design.

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:
    - M. L. Samways, H. E. Bruce Macdonald and J. W. Essex, *J. Chem. Inf. Model.*, 2020, **60**, 4436-4441
    - M. L. Samways, R. D. Taylor, H. E. Bruce Macdonald and J. W. Essex, *Chem. Soc. Rev.*, 2021, **50**, 9104-9120
    - 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

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

Acknowledgements

First, I'd like to thank all current and former members of the Essex Group for providing such a great working atmosphere — I've learned so much by working with you all over the years. Not only is there always someone who is able to help with a given issue, but it is so much easier to work hard when surrounded by a great group of colleagues and friends. Specifically, I'd like to thank Dr Hannah Bruce Macdonald, for helping me to learn the fundamentals of GCMC; Ollie Melling and Will Poole for their help with the *grand* module; Miroslav Suruzhon and Khaled Maksoud for always being willing to discuss difficult and abstract problems; and Mabel Wong, Anna Cavalleri and Jack Sawdon for listening to my many complaints over the past few years. I'd also like to extend my gratitude to members of other research groups — particularly the Frey Group — who have supported me during this journey.

I'd also like to thank my supervisor, Prof. Jonathan Essex, for all of his support, not only during my PhD, but also during the summer projects that I carried out in his group as an undergraduate. The advice that you've given me over the years, whilst also allowing me to work so independently, has given me the opportunity and freedom to learn and develop as a scientist. Thank you for having the confidence to give me such a challenging, but rewarding, project to work on.

I'd also like to thank Dr Richard Taylor of UCB, with whom I have collaborated on a number of interesting projects over the past few years. Whilst results from these projects are not presented in this thesis, the experience gained, and advice received, have shaped my outlook during my PhD, allowing me to focus my work on GCMC where it would have most relevance to drug design.

Finally, I want to thank my family — especially my parents, Gary and Mel, and my fiancée, Ashleigh — who, in spite of not knowing what I actually do, have supported me unwaveringly throughout.

Definitions and Abbreviations

$\langle \dots \rangle$	Ensemble average
B	Adams parameter
B_{equil}	Adams parameter at equilibrium
β	Thermodynamic beta, defined as $(k_B T)^{-1}$
$\mathrm{d}\mathbf{r}^N$	Shorthand for $\mathrm{d}x_1\mathrm{d}y_1\mathrm{d}z_1 \dots \mathrm{d}x_N\mathrm{d}y_N\mathrm{d}z_N$
E	Total energy
F	Helmholtz free energy
G	Gibbs free energy
h	Planck constant
k_B	Boltzmann constant
Λ	Thermodynamic wavelength
Λ_p	NCMC protocol
λ	Alchemical coupling parameter
m	Mass
μ	Chemical potential
μ^{ex}	Excess chemical potential
N	Number of particles
Ω	Grand potential
$\mathbf{p}^N$	Momentum vector for N particles
π	Equilibrium probability (when written as a function)
Q_{NVT}	Canonical partition function
$\mathbf{r}^N$	Position vector for N particles
ρ	Density (of probability or number, depending on context)
$\mathbf{s}^N$	Scaled position vector for N particles
T	Temperature
τ	Nonequilibrium switching time
U	Potential energy
V	Volume
V_{GCMC}	Volume over which GCMC sampling is carried out
W	Work
W_p	Protocol work
$\Xi_{\mu VT}$	Grand canonical partition function

Z_{NPT}	Isothermal-isobaric partition function
BAR	Bennett acceptance ratio
BPTI	Bovine pancreatic trypsin inhibitor
GCMC	Grand canonical Monte Carlo
GCMC/MD	A simulation which combines GCMC sampling with MD
GCNCMC	Grand canonical nonequilibrium candidate Monte Carlo
GCNCMC/MD	A simulation which combines GCNCMC sampling with MD
M2	Matrix 2 protein
MBAR	Multistate Bennett acceptance ratio
MC	Monte Carlo
MD	Molecular dynamics
NCMC	Nonequilibrium candidate Monte Carlo
PDB	Protein Data Bank
PME	Particle mesh Ewald
QM	Quantum mechanics/mechanical
WT	Wild type

Chapter 1

Introduction

1.1 Significance of Protein-Bound Water Molecules

The binding of small molecules to protein targets is very important in many biological processes, as well as in drug design — a 2017 study estimated that small molecules represent approximately 85 % of approved drugs.[1] Accordingly, there are a number of computational methods which can be used in order to study protein-ligand complexes, ranging from docking, which seeks to provide a rapid prediction of the structure and stability of a complex,[2] to free energy calculations, which typically aim to rigorously predict differences in binding affinity between pairs of ligands.[3] As the vast majority of biology occurs in an aqueous environment, water also plays an important role in protein-small molecule binding. For example, if two molecules form equally favourable interactions with the protein, the less hydrophilic of these will bind more strongly, owing to the greater difference in stability between the solvated and protein-bound states — this is known as the hydrophobic effect.[4] However, water is not just the solvent in which biological events take place, as it also plays an active role in many processes.[5] Of particular interest here is the impact that protein-bound water molecules can have on the thermodynamics of ligand binding.[5–10] The prevalence of water molecules at protein-ligand interfaces was highlighted by a 2007 study of 392 high resolution crystal structures, which found that over 85 % of the complexes contained at least one water bridge between the protein and ligand.[11]

Water molecules which bind at protein-ligand interfaces are typically much more restricted in terms of their motion than water molecules in bulk solution. This means that such ordered water sites bind to the protein with a loss of entropy, thereby requiring a negative enthalpy change in order for the binding to be associated with a favourable change in free energy:

$$\Delta G = \Delta H - T\Delta S \tag{1.1}$$

where ΔG is the change in Gibbs free energy, ΔH is the change in enthalpy, T is the temperature and ΔS is the change in entropy. This entropic effect is of great interest in drug design,[10] because if a compound modification can be designed to cause the displacement of an ordered water site (shown schematically in Fig. 1.1), then the gain in entropy associated with releasing the water into bulk solution will contribute favourably to the binding affinity of the compound. It has been suggested that the maximum possible value of this entropic effect is around 2 kcal mol^{-1}, based on the entropy associated with the transition from ice to liquid water.[6] However, it should be noted that water displacement is likely to be enthalpically unfavourable, and that the interactions that the water makes with the complex must be adequately replaced by the compound modification in order for the displacement to have a net favourable effect on the ligand binding affinity. In some cases, it may be more effective to conserve the water, and treat it as part of the protein binding site, with which the interactions involving the ligand should be optimised.

There are many examples in the literature of cases where water displacement has been associated with both increases[12–14] and decreases[15,16] in ligand binding affinity — a widely recognised example is HIV-1 protease, where the displacement of a highly ordered water site is correlated with a large increase in binding affinity for cyclic urea inhibitors.[17] As well as simple changes in binding affinity, there are also a number of cases where the hydration pattern of a protein binding site plays a key role in the selectivity[18–21] or promiscuity[22] of the site. However, it should be noted that the exact balance between the enthalpic and entropic contributions of a given water site is not known *a priori*, and therefore, it is very difficult to predict whether compound development is best served by displacement or conservation of the site.

The sections below describe some of the experimental and computational methods which can be used for the study of water in protein complexes.

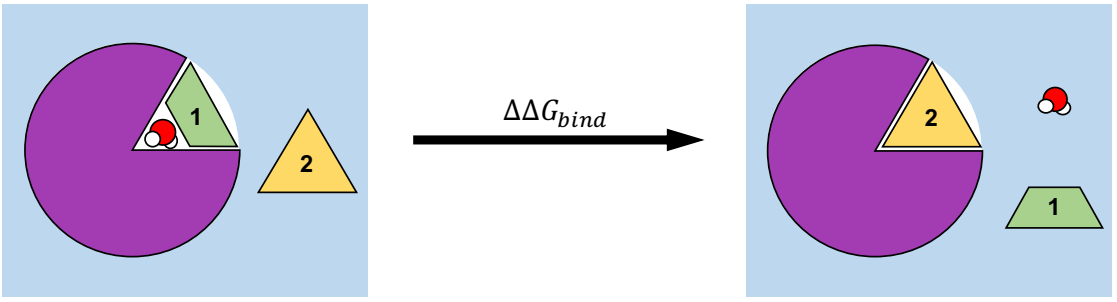


FIGURE 1.1: Schematic showing a hypothetical compound modification, where compound 2 was designed, based on compound 1, in order to displace the buried water site. $\Delta\Delta G_{bind}$ is the difference in binding free energies of the two compounds.

1.2 Experiment

X-ray crystallography is the primary experimental method used for the structural analysis of protein-bound water molecules. In these experiments, the electron density distribution is inferred from the X-ray diffraction pattern of a single crystal, and a structural model is then proposed and refined to fit the electron density.[23] One potential issue with this method is that the conditions under which the crystals were formed may be very different to physiological conditions, which could cause a distortion in the protein structure.[24] However, there are also a number of practical limitations which affect the identification of water binding sites within protein structures. First, it is important to note that water is isoelectronic with a number of commonly used ions (Na^+, Mg^{2+}, NH_4^+ and F^-), from which it cannot be easily distinguished in an electron density map[23] — although the identity may sometimes be inferred from the surrounding environment, such as nearby charged species and/or hydrogen bonding groups. Secondly, if a water site is highly mobile, or partially occupied, the electron density observed could be very weak, causing the site to be missed. Additionally, the positions and occupancies of water sites are sometimes inadvertently fitted to the noise in the electron density, thereby 'absorbing' errors in the model, by reducing the amount of 'unexplained' density.[25] The cumulative impact of these issues has been demonstrated by studies in which very different water distributions were assigned to the same structure by different crystallographers.[26,27] Nevertheless, X-ray crystallography is preferable to the alternatives (discussed below), and is the source of the bulk of experimental data on water locations within protein structures.

The positions of hydrogen atoms are rarely assigned in X-ray crystal structures, as they have very little impact on the scattering of X-rays, and are therefore not easily detected. However, hydrogen/deuterium atoms are much better at scattering neutrons, allowing their positions to be identified much more easily using neutron diffraction.[28,29] Therefore, these structures can resolve some of the aforementioned issues with determining water binding locations, by making it easier to distinguish water sites from isoelectronic groups and structural noise, via the characteristic 'boomerang' shape typically observed for water molecules in nuclear density maps.[30] However, neutron structures are more difficult to obtain than X-ray structures, in part owing to the requirement for much larger crystals.[28] The impact of this limitation is evidenced by the fact that only 179 of 176,528 structures in the Protein Data Bank (PDB)[31,32] are labelled as neutron diffraction structures (as of 10$^{\text{th}}$ April 2021). It should be noted that water locations can also be inferred from nuclear magnetic resonance (NMR) spectroscopy, by identifying water-protein interactions via the nuclear Overhauser effect.[33–36] However, this also suffers from a number of limitations, one of which is that the interactions are not easily studied between water sites and protein sites involving labile hydrogen atoms,

as the signal is significantly disturbed by hydrogen exchange between water and the protein.[33,34,37]

1.3 Computation

As well as some of the aforementioned limitations of experimental methods for the identification of water binding locations, it should also be noted that it is not possible to experimentally measure the thermodynamics of individual water sites. Owing to these limitations, computation can play a very useful role in the molecular-level analysis of water positions and thermodynamics within a protein structure.[38,39] To this end, a large number of methods have been developed over the years, to offer computational prediction and analysis of protein-bound water sites, at varying levels of both theoretical rigour and computational cost.[40] A number of these methods are described below.

1.3.1 Knowledge-Based Methods

Owing to the wealth of available crystallographic data regarding protein-bound water sites,[31,32] a large number of knowledge-based methods have been developed which attempt to predict water locations in protein structures by extrapolating from these data. The more common approach adopted by these methods is to model the structural features surrounding crystallographic water sites, which can then be used to identify likely water locations within a protein of interest. The AQUARIUS method calculates the observed 3D distributions of water about each of the amino acids from a knowledge base of structures, and then maps these onto the amino acid coordinates of the query structure, yielding a distribution of water sites.[41,42] Similarly, the SuperStar method extracts fragment-water interaction distributions from the IsoStar database,[43] which are then superimposed onto instances of those molecule fragments within a protein structure, generating a map of water positions around the structure.[44,45] AcquaAlta identifies likely water locations based on the hydrogen bonding geometries which would be formed with the protein at those locations (using a knowledge base of hydrogen bond geometries), whilst also using quantum mechanical data to prioritise different hydrogen bond types, based on their calculated strength.[46] The WarPP method also searches the space around the protein, and ranks specific points, based on the quality of hydrogen bonds formed (using observed hydrogen bond geometries for reference), water sites are placed at optimal locations and their positions are then refined together.[47] Xiao *et al.* developed a method which identifies tetrahedral units within a knowledge base, where one vertex is a water site and the other three are protein atoms, from which water molecules are then mapped onto triplets of protein atoms within a structure of interest.[48] A somewhat different approach is employed by the wPMF method, which constructs a knowledge-based potential, based on radial distribution functions between

waters and different types of protein atom. These potentials are then used to calculate the probability of a water site at points on a lattice, and the more favourable points are clustered to identify water binding locations.[49] Knowledge-based methods for water site prediction often achieve good accuracy in terms of the identification of crystallographic water sites, as they are trained on crystallographic data. However, the quality of the predictions is dependent on the quality of the training data, and as such, these methods will suffer from the same limitations regarding water placement as X-ray crystallography (discussed in section 1.2). Additionally, these methods may have difficulty with protein-ligand interfaces, if the chemistry of the ligand is not adequately represented in the training set.

As well as the prediction of water binding locations, knowledge-based methods have also been developed to characterise observed water binding sites — typically, this involves predicting whether a site is likely to be conserved or displaced upon ligand binding. A number of methods employ clustering of water sites from a series of superimposed structures of very similar proteins, where the cluster occupancy (i.e. the fraction of structures containing a given water) is assumed to be linked to the stability of a water site — these methods include ProBiS H2O,[50] PyWATER,[51] WatCH[52] and the method described by Bottoms *et al.*[53] The Consolv method predicts whether a water site observed in an *apo*-structure is likely to be conserved or displaced upon ligand binding, using a *k*-nearest neighbours algorithm. Each water molecule is modelled by a series of structural descriptors (atom density, hydrophilicity, number of hydrogen bonds and crystallographic temperature factor), and then compared against a knowledge base of waters, where the *k* most similar sites (note that the value of *k* is a parameter of the method) 'vote' on whether the query water would be conserved or displaced.[54] WaterScore uses a logistic regression of the temperature factor, number of protein contacts and solvent-accessible surface area of water sites to predict their probability of being conserved in corresponding *holo*-structures.[55] Ross *et al.* trained a tree-based machine learning model to predict whether water sites would be conserved or displaced, based on terms describing the hydrogen bond interactions, hydrophilicity and lipophilicity of the site.[56] Again, these methods are limited by the relevance of the data used to train and parameterise the models. Additionally, many of these methods assume that the likelihood of water displacement is inherent to the water site, when in reality, the nature of the ligand will also play a significant role. It should also be noted that those methods which make use of the temperature factor of a water site (such as Consolv and WaterScore) can only be applied to crystallographic water sites, and not predicted sites.

1.3.2 Interaction-Based Site Prediction

For a given model of how water interacts with a protein environment, water sites can be predicted by identifying the most energetically favourable binding locations within a structure of interest. One approach is to superimpose a 3D lattice onto the structure and then sample the water interaction energy at each point, as first proposed by Goodford with the GRID method.[57] The 3D reference interaction site model (3D-RISM) uses an integral equation theory (based on statistical mechanics) in order to resolve a continuous solvent distribution onto a grid.[58–60] Methods such as Placevent[61] and GAsol[62] can then place explicit water sites around the structure in order to best represent the distribution calculated using 3D-RISM. The method proposed by Setny and Zacharias identifies grid cells as occupied or unoccupied by water, based on their interactions with the protein, and then refines this grid, taking water-water interactions into account (via a continuum model).[63] Ben-Shalom *et al.* developed a method which allows water to be translated between points on a grid (covering both protein and bulk solvent), thereby allowing waters to be exchanged much more rapidly during molecular dynamics simulations.[64]

Rather than sampling regular points on a grid, a number of other methods instead place waters by 'flooding' protein cavities, and then refining these randomly generated sites (typically independently of each other). The multiple copy simultaneous search (MCSS) method first discards water sites with interaction energies which do not meet some threshold, then optimises the positions of those which remain, and discards any overlapping sites.[65] WATGEN scores water sites based on their hydrogen bonds formed with the protein, then selects from these, starting from those with the highest scores (any waters which clash with higher scoring sites are discarded).[66] The Dowser method minimises each water site, and then discards those which do not meet a defined threshold[67] — various modifications have been made to optimise the parameters of this method.[68,69] WaterDock makes use of a docking program to insert waters into a protein structure, then, after running this many times, the docked water locations are clustered into discrete sites[56] — WaterDock 2.0 reduces the number of false predictions by filtering water sites based on the quality of their hydrogen bonds with the ligand.[70]

In general, these methods offer a significant degree of control over the balance between the computational cost and the rigour of the predictions made. For example, increasing the quality of the interaction model used, and sampling the space around the protein more extensively would be expected to give better quality predictions, albeit at increased computational cost. Conversely, if the speed of the predictions is prioritised (such as when screening a large number of structures), then a more approximate interaction model might be used in combination with a less extensive search. It should be

noted that some of these methods produce water density grids, rather than water locations, which might be more difficult to interpret in a drug design context. Additionally, some methods (Dowser, GRID and WaterDock) consider water molecules in isolation, and therefore may fail to capture water sites which are stabilised by other waters.

1.3.3 Inclusion of Waters in Ligand Docking

Protein-ligand docking is a widely used tool in computer-aided drug design, as the rapid predictions of ligand binding modes, and associated estimation of binding affinity (referred to as a docking score), make this a very useful tool for virtual screening.[2] Owing to the influence that structural water molecules can have on ligand binding, the majority of docking programs now include some treatment of binding site waters. Most approaches include explicit water molecules as part of the protein structure (often requiring pre-determined water locations), with varying degrees of water flexibility — in some cases, the water positions are fixed, and many methods allow these waters to be displaced during ligand docking. FlexX includes fixed water molecules as the ligand is grown from fragments, with the presence of the waters evaluated throughout, in order to determine if they should be bound or displaced.[71] The SLIDE method first attempts to resolve ligand-water clashes by translation of the water site, and, if this is not reasonably feasible, the water is displaced and the conservation probability of the site (predicted with Consolv[54]) is used to apply an energetic penalty.[72] Glide XP uses grid-based sampling to dock water molecules after the ligand has been docked to the protein[73] — the WScore method builds upon Glide XP by including flexible, displaceable water sites during the ligand docking procedure, using results from a WaterMap[74,75] analysis (see the following section) to determine the thermodynamic effect of these displacements.[76] The GOLD program includes rotationally flexible water sites which can be displaced during ligand docking, where conserved water sites result in an entropic penalty for the complex[77] — a similar approach is taken by the FITTED method.[78] It is also possible to use grid-resolved water densities and/or thermodynamics from the GIST method[79] (see the following section), which has been used to influence ligand docking in both the AutoDock[80] and DOCK[81,82] programs. A distinct approach is to treat the water molecules as a part of the ligand, rather than the protein. In these methods, water molecules are attached to the ligand during the docking procedure, based on hydrogen bond geometries.[83] This approach is available in AutoDock[84] and RosettaLigand.[85]

The key advantage of these docking methods is the speed with which they can be executed, making them very useful for screening. They also differ from many of the other categories of methods discussed, in that they do not require the binding mode of the

ligand to be known *a priori*, as this is precisely what they are used to predict. The authors of the methods discussed in this section consistently reported that a more rigorous treatment of water in docking improves the measured performance of the algorithm, thereby highlighting the importance of water molecules in protein-ligand binding.

1.3.4 Thermodynamic Analysis of Water Sites

As previously mentioned, experimental methods cannot directly measure the thermodynamics associated with the binding of individual water molecules at protein-ligand interfaces. Fortunately, computational methods do not suffer from this limitation, and a number of methods have been developed to calculate water binding free energies, with varying degrees of theoretical rigour — many of these methods require water locations to be determined prior to analysis. Widely considered to be the gold standard in the calculation of water binding free energies,[86] is the double decoupling method.[87] This involves calculating the free energies associated with removing a water molecule (the concept of alchemical decoupling is described in section 2.4) from the protein binding site, and also from bulk water, where the difference between these gives the binding free energy of the site. This typically involves the application of restraints/constraints which prevent the removed water site from being replaced with another water molecule (which is not possible in experiments) — the calculated free energy must then be corrected for this.[87] Using this method, Barillari *et al.* have shown that, in general, water molecules with more negative binding free energies are less likely to be displaced by a ligand.[88]

A number of methods employ inhomogeneous fluid solvation theory (IFST)[89,90] for the analysis of water configurations obtained from molecular dynamics (MD) simulations. Here, the binding enthalpy of a water site is typically calculated based on the average interaction energy observed (a widely used assumption), and the entropy is estimated from correlation functions of the water positions and orientations. This methodology can be used to calculate thermodynamics for discrete water sites, as implemented in WaterMap,[74,75] for example, or the solvation thermodynamics can be resolved onto a lattice, yielding grid-based inhomogeneous solvation theory[79] (GIST). In IFST, the entropy is typically calculated from two contributions: distributions in the water positions relative to a solute, and distributions of water orientations in the solute frame of reference (these distributions are assumed to be uniform in bulk water). However, this is a truncation of the solvation entropy as described by IFST — the full expression involves integrating over all possible combinations of water molecules[79] — and the GIST method has since been expanded to take second-order correlations between water molecules into account.[91] Many other approaches also exist for estimation of the entropy of protein-bound water sites from simulation data. WATsite approximates

the entropy using probability distributions of the water positions and orientations[92] — these data can also be used by the DeepWATsite algorithm to rescore structures obtained from docking.[93] The SPAM approach estimates the water binding free energy from the distribution of interaction energies observed, from which the entropic contribution is inferred by subtracting the average interaction energy from the free energy.[94] Another approach is cell theory, which calculates orientational, vibrational and librational entropic effects for water using the number of orientations, forces and torques observed for each water, relative to the corresponding values for bulk water[95] — like IFST, these data can be resolved onto a grid (known as grid cell theory).[96,97] Unlike IFST, cell theory does not involve a large expansion of the entropic term, as it makes use of a mean field approximation,[96] which improves the efficiency of the entropy calculation.

JAWS is a Monte Carlo method which allows the prediction of water locations and estimation of their binding free energies.[98] Here, water sites are partially decoupled via a parameter which is free to vary over the course of a simulation, and provides a route for waters to gradually bind and unbind from the system, whilst also allowing them to better explore the protein. The favourability of water binding in different regions of the protein can also be determined from the extent to which the corresponding waters are fully coupled.[98] Grand canonical Monte Carlo (GCMC) simulations are carried out at constant chemical potential, volume and temperature, which allow the number of particles to fluctuate according to these constraints.[99–102] This involves the insertion and deletion of particles, which can be used to allow waters to rapidly bind and unbind from a protein over the course of a simulation.[103,104] The binding free energy of a water network can be calculated from simulations at a range of chemical potential values,[104,105] whilst also predicting water binding locations.

These methods are usually much slower than the other classes of methods discussed, as they typically involve complex descriptions of molecular interactions and simulations which can be very computationally expensive. A common limitation of these methods is that many of those which analyse molecular dynamics data are based on the assumption that the simulated water distributions are equilibrated, which may not be the case, as water exchange between proteins and solution can be very slow.[106] Additionally, some of these approaches analyse water sites independently of each other, meaning that cooperative effects between waters are not captured. Despite these limitations, computational analysis of the thermodynamics of protein-bound water sites is recognised as a useful tool in modern structure-based drug design.[38,39]

1.4 Objectives

The method which underpins the work presented in this thesis is grand canonical Monte Carlo (GCMC). As discussed briefly in the previous section, this is a simulation method which allows the number of particles in a simulation to fluctuate according to a defined chemical potential (which is constant).[99–102] When applied to protein binding sites, this allows waters to rapidly bind and unbind, without kinetic limitations.[21,103–105,107–110] This is useful, as the timescales of water exchange between a protein binding site and bulk solution can be as long as milliseconds,[106] which can be very problematic for molecular dynamics simulations.[64] GCMC simulations have been found to predict the locations of crystallographic water sites in protein binding sites very well,[104] which, as previously discussed, can be very useful in computer-aided drug design. Beyond simple prediction of water binding locations, GCMC simulations can also be used to rigorously calculate the binding free energy of water networks to protein binding sites (giving results in good agreement with double decoupling)[105] which offers insight into the thermodynamics associated with binding site hydration — it should also be noted that cooperative effects between water sites are implicitly captured by GCMC. Additionally, GCMC can be used to automatically capture the effects of water displacement when carrying out free energy calculations where one ligand is perturbed into another[107,109,110] — otherwise, the accuracy of these calculations can be significantly impaired if the displaced water is not expelled from the protein binding site within the timescale of the simulation.[111] The underlying theory of GCMC is described in detail in section 2.6.3.

Despite the power of grand canonical simulation methods, they are not yet widely used for the simulation of biomolecular systems — their utility is, however, recognised in the computational study of adsorbent binding to porous materials.[112] The objectives of this work were to increase the usability of GCMC simulations of protein-bound water sites, apply these methods to protein systems of interest, and to build upon the recent theoretical and methodological developments.[104,105,109] In order to facilitate these aims, the *grand* Python module was developed during this work,[113] which serves as a 'bolt-on' tool to allow GCMC sampling to be carried out with the OpenMM[114] simulation engine. This approach was chosen because OpenMM has a large and growing user base, and the highly customisable framework makes it very suitable to the development of prototype code for novel methods. Chapter 3 describes the development and validation of the GCMC implementation in *grand*. This module was then used to study the thermodynamics of water binding to the matrix 2 protein (an influenza A drug target), in order to investigate the effects of the water network on ligand binding to the transmembrane domain, as presented in chapter 4. Chapter 5 presents a novel development, in which nonequilibrium candidate Monte Carlo[115] (NCMC) is used to drastically improve the efficiency of GCMC by allowing the insertion and deletion of water sites to

be carried out in a smoother fashion — this method is referred to as grand canonical nonequilibrium candidate Monte Carlo (GCNCMC). It is then shown in chapter 6 that GCNCMC makes grand canonical sampling of small organic molecules feasible, which could have applications in computational fragment-based drug design.

Chapter 2

Theory and Methods

2.1 Modelling Molecular Interactions

The total energy, E, of a molecular system can be calculated using the time-independent Schrödinger equation:[116]

$$\hat{\mathcal{H}}\psi = E\psi \tag{2.1}$$

where $\hat{\mathcal{H}}$ is the Hamiltonian operator and ψ is the wavefunction of the system. However, this equation cannot be solved analytically for multi-electron systems, and the cost of numerical methods increases very rapidly with the size of the system.[116] Therefore, to simulate 'large' systems, such as those of interest in biological processes, much of computational chemistry employs classical Hamiltonians, in order to make the simulations tractable. However, this comes at the cost of reduced accuracy, and a neglect of quantum mechanical effects (such as bond-breaking and electron transfer, for example) in these simulations. This also means that energies for these classical simulations cannot be determined from first principles, and require a number of fitted parameters. The combination of a particular Hamiltonian with a particular set of parameters is known as a force field. This section describes the general structure of modern force fields.

In a classical simulation of a system containing N particles, with positions, $\mathbf{r}^N$, and momenta, $\mathbf{p}^N$, the total energy is calculated as the sum of the potential and kinetic energies:[116,117]

$$E(\mathbf{r}^N, \mathbf{p}^N) = U(\mathbf{r}^N) + \sum_{i=1}^{N} \frac{|\mathbf{p_i}|^2}{2m} \tag{2.2}$$

where m is the particle mass, and $U(\mathbf{r}^N)$ is the potential energy calculated as a function of the particle positions. The potential energy can be decomposed into a series of

different contributions:

$$U = U_{bonds} + U_{angles} + U_{dihedrals} + U_{ele} + U_{vdW} \tag{2.3}$$

where U_{bonds}, U_{angles} and $U_{dihedrals}$ are the bonded terms, which arise from bond stretching, angle bending and dihedral rotation, respectively; and U_{ele} and U_{vdW} are the non-bonded terms, which describe electrostatic and van der Waals interactions, respectively. These two groups of terms are described in more detail below.

2.1.1 Bonded Interactions

The potential energy associated with bond stretching is modelled as a harmonic potential, where the bonded interaction between atoms i and j is calculated as a function of the bond length, ℓ_{ij}:

$$U_{ij}^{bond}(\ell_{ij}) = \frac{1}{2}k_{ij}^{\ell}(\ell_{ij} - \ell_{ij}^{eq})^2 \tag{2.4}$$

where ℓ_{ij}^{eq} is the equilibrium bond length, and k_{ij}^{ℓ} is the force constant. These parameters might be extracted from experimental data, such as crystallographic structures for the equilibrium bond length and infrared spectroscopy for the force constant, or they might be fitted to quantum mechanical (QM) data. This potential is a good approximation of the bond interaction energy at bond lengths close to the equilibrium value, but the approximation breaks down at very short and very long bond lengths.[116] However, the majority of simulations are only concerned with allowing bonds to fluctuate about their equilibrium lengths, and as Eq. 2.4 only requires two parameters, this is a convenient, and widely employed, approximation. The value of U_{bonds} is calculated as the sum of bonded interactions over all bonds.

Similarly, the angle bending interactions are also typically modelled via a harmonic potential, where the interaction for an angle, θ_{ijk}, involving atoms i, j and k is calculated as:[116]

$$U_{ijk}^{angle}(\theta_{ijk}) = \frac{1}{2}k_{ijk}^{\theta}(\theta_{ijk} - \theta_{ijk}^{eq})^2 \tag{2.5}$$

where θ_{ijk}^{eq} is the equilibrium value of the angle, and k_{ijk}^{θ} is the force constant. The parameters can again be inferred from experimental measurements, or fitted to QM data. Alternatively, some equilibrium angle values might be determined using chemical intuition — for example, the angles about an sp^2-hybridised carbon centre might be set to 120°. Again, this is a reasonable approximation, as we expect the angles to fluctuate about their equilibrium values. The value of U_{angles} is calculated as the sum of angle bending interactions over all bond angles.

The modelling of dihedral interactions is less trivial than that of bond stretching and angle bending, for two reasons. Firstly, it is almost always the case that dihedral interactions have multiple local energy minima, and secondly, they must also satisfy a periodicity compatible with 360°, in order to avoid discontinuities. Both of these requirements can be satisfied via a Fourier series consisting of M cosine functions, where the interaction energy for the dihedral of atoms i, j, k and l is given by:[116]

$$U_{ijkl}^{dihedral}(\varphi_{ijkl}) = \sum_{m=1}^{M} k_m \left[\cos(n_m \varphi_{ijkl} - \gamma_m) + 1\right] \tag{2.6}$$

where φ_{ijkl} is the dihedral angle, and k_m, n_m and γ_m are parameters associated with the m^{th} cosine function. These parameters are often fitted to a potential energy surface calculated at the quantum mechanical level. However, as dihedral parameters are typically the last to be fitted for a force field, they may unintentionally absorb errors which arise from deficiencies in other aspects of the model. $U_{dihedrals}$ is calculated as the sum of all dihedral interaction energies.

In order to make the parameters more suitable for general use, most force fields make use of atom types, which provide a qualitative description of the chemistry of each atom (such as whether a carbon atom is aromatic or aliphatic, for example). The parameters can then be determined for different combinations of atom types, rather than exact combinations of atoms. For example, a generalised set of bonded parameters might be determined for bonds between two sp^3-hybridised carbon atoms, which are then applied to all bonds of this type, eliminating the need to specifically parameterise every carbon-carbon bond encountered. This aims to reduce the number of parameters required, and improve the transferability of the force field to new molecules. However, recent developments, such as the Open Force Field Initiative, have sought to develop force fields which do not employ atom types, as some complex molecules can be difficult to describe within the constraints imposed by the atom types available within a force field.[118,119] This is based on the concept of direct chemical perception, where the force field parameters are inferred directly from the molecular chemistry,[119] rather than indirectly via the set of pre-defined atom types.[118] The SMIRNOFF99Frosst force field, developed using this approach, was observed to provide similar accuracy to the general AMBER force field[120] (GAFF) for the prediction of small molecule hydration free energies, but requiring many fewer force field parameters to do so.[118]

2.1.2 Nonbonded Interactions

The majority of force fields describe the charge distribution of a system by using point charges centred on each atom. Therefore, the interactions between all of these point

charges can be calculated using Coulomb's law:[116,117]

$$U_{ele} = \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.7}$$

where q_i is the charge on atom i, ϵ_0 is the vacuum permittivity and r_{ij} is the interparticle separation for particles i and j. The charges are often determined using some form of quantum mechanical analysis. One approach is to rapidly calculate a set of charges from a semi-empirical quantum mechanical calculation (which make use of empirical parameters and are therefore more approximate than *ab initio* methods). The AM1-BCC method rapidly calculates an initial set of atomic charges from AM1 semi-empirical calculations,[121] and then applies a bond charge correction (BCC) to these charges, which takes into account the atoms to which each atom is bonded.[122,123] Bond types are defined based on pairs of atom types, and the corrections for each bond type were parameterised by fitting to the molecular electrostatic potential (calculated at a higher level of theory).[122,123] More rigorous (and computationally expensive) approaches make use of *ab initio* QM calculations, such as the restrained electrostatic potential (RESP) method, which optimises the atomic charges (within some restraints) to reproduce the molecular electrostatic potential calculated using high level QM calculations.[124] Distributed multipole analysis is another method, which uses a multipole expansion about each atom to fit the calculated charge distribution.[125] Other approaches involve artificially 'sharing' the electron density across the atoms — some of which are based on the quantum theory of atoms in molecules[126] — which can then be combined with the nuclear charges to generate a net charge for each atom. There are a number of methods for the distribution of the electron density, including Mulliken population analysis,[127] Hirshfeld analysis[128] and iterative stockholder analysis.[129–131]

However, the use of atom-centred point charges to represent a charge distribution is rather limited. There are two issues with the representation of charge in most force fields: firstly, these charges often lack the complexity required to describe molecular charge distributions (the quadrupolar nature of benzene is a good example), and secondly, these charges are typically fixed, and do not change in response to their environment. Two approaches exist for the first issue: one of which is to include higher order multipoles at the atom centres, and the other is to consider additional point charges, positioned slightly off-centre.[132,133] As for the second issue, polarisation is very difficult to efficiently include as it must be iteratively solved, owing to the interdependence of the induced multipoles of different atoms. The AMOEBA force field offers a solution to both of these issues, as it uses permanent dipoles and quadrupoles to better represent the fixed charge distribution, as well as including an induced dipolar term which accounts for polarisation, where the dipoles are updated in a self-consistent manner.[134–138] However, the increased computational cost of this force field presents

a barrier to its widespread use.

The van der Waals term in Eq. 2.3 is used to model two effects: the short-range repulsive interaction between particles (owing to nuclear repulsion and the Pauli exclusion principle) and the attractive dispersion interaction (which arises from instantaneous multipoles). These effects are typically modelled using the Lennard-Jones 12-6 equation:[139]

$$U_{vdW} = \sum_{i=1}^{N} \sum_{j>i}^{N} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.8}$$

where ε_{ij} is the minimum potential energy for the interaction between particles i and j, and σ_{ij} is the finite distance at which the interaction is zero — the interaction is repulsive at distances shorter than σ_{ij} and attractive at distances longer than σ_{ij}. The r^{-6} term arises from the dipole-dipole contribution, which is the leading term in the dispersion interaction (the higher order terms decay much more rapidly with respect to distance, and are therefore neglected by the Lennard-Jones equation). The r^{-12} term, however, is a more approximate representation. At very short distances, the repulsion decays as r^{-1} (owing to the unshielded electrostatic repulsion between the nuclei) and the decay then becomes exponential at less short distances, owing to the overlap of the molecular wavefunctions — both of these effects are approximated as r^{-12}, the reasons for which are pragmatic in nature. Firstly, the Lennard-Jones equation requires only two parameters per interaction, which makes the parameter fitting easier and less susceptible to overfitting. Secondly, in the early days of computational chemistry, when computers were much less powerful than they are now, it was of significant practical benefit that r^{-12} can be calculated directly from r^{-6}, and that both can be calculated using the square distance, r^2, which was more efficiently calculated than the distance, r.[116] The σ and ε parameters can be fitted to either QM data, or to reproduce bulk experimental measurements (such as the density of a fluid). These parameters can be transferred for different combinations of atom types, using the Lorentz-Berthelot combining rules:[140]

$$\sigma_{ij} = \frac{1}{2} \left(\sigma_{ii} + \sigma_{jj} \right) \tag{2.9a}$$

$$\varepsilon_{ij} = \left(\varepsilon_{ii} \varepsilon_{jj} \right)^{\frac{1}{2}} \tag{2.9b}$$

Other combining rules exist, but those above are the most common.

It should be noted that it is common practice to ignore non-bonded interactions for pairs of particles separated by one or two bonds, as their close distance would otherwise cause repulsions which might disturb the molecular geometry. Similarly, the non-bonded interactions for those which are separated by three bonds (known as 1-4 interactions) are often scaled, such that they do not fully interact, where the scaling factor often varies by force field.

2.1.2.1 Long-range interactions

As the non-bonded interactions must be calculated over (almost) all pairs of particles in the system, the cost of calculating the sum of these terms scales as $\mathcal{O}(N^2)$, and is one of the primary contributors to the computational cost of molecular simulations. Therefore, in order to reduce the impact of this, it is common to ignore non-bonded interactions beyond some cutoff distance, r_c, where they are assumed to be negligible. However, the process of deciding which interactions to ignore would require calculating the distances for all pairs of particles, which still scales as $\mathcal{O}(N^2)$. Therefore, modern simulation engines make use of neighbour lists, where, for each particle, a list of 'neighbour' particles which lie within some distance (which must be larger than the interaction cutoff) is maintained. Each interaction calculation therefore only involves the calculation of distances for particles within neighbour lists of each other. The neighbour list must be updated sufficiently often to ensure that no interactions are missed.[116,117] Additionally, if interactions beyond the cutoff distance are simply truncated, there is a risk that discontinuities will be introduced (especially in molecular dynamics simulations, where the derivative of the potential energy must be smooth — see section 2.5), therefore, switching functions are often employed to ensure that the interaction energies go smoothly to zero at the cutoff distance. The switching function is employed between some distance, r_s, and the cutoff, such that interactions at distances less than or equal to r_s are not scaled, those with distances greater than, or equal to, the cutoff evaluate to zero, and those in between are scaled smoothly. The switching distance must be sufficiently close to the cutoff distance to minimise the disturbance to the interactions, but not so close that the interactions are scaled too quickly. An additional requirement is that the first and second derivatives of the switching function must be equal to zero at both r_s and r_c, in order to avoid the introduction of additional discontinuities.[116,117]

In order to limit finite size effects, and to maximise the relevance of simulated systems to their macroscopic equivalents, most modern simulations employ periodic boundary conditions (PBCs). This involves considering the simulated system to be periodically surrounded by an infinite number of copies of itself, known as images. The calculation of non-bonded interactions employs the *minimum image convention*, where the interaction between a pair of particles is taken as that between their closest pair of respective images. These periodic images need not be explicitly simulated, as the calculated interparticle distances can be simply corrected to account for the periodic boundaries, based on the dimensions of the simulation. An additional requirement for simulations employing periodic boundaries is that the value of the interaction cutoff must be less than half of the shortest simulation box dimension, in order to ensure that each particle can only interact with a maximum of one image of any other particle.[116,117]

However, the choice of the cutoff distance may have an impact on the simulation behaviour, if the sum of the interactions beyond the cutoff is non-negligible. The long-range correction for a given potential energy function, $U(r)$, can be written as:[117]

$$U_{r>r_c} = \frac{N \langle \rho \rangle}{2} \int_{r_c}^{\infty} 4\pi r^2 U(r) \mathrm{d}r \tag{2.10}$$

where $\langle \rho \rangle$ is the average number density (it is assumed that the number density is equal to this value at long distances). The above therefore dictates that the potential energy term must decay at a rate greater than, or equal to r^{-3} in order to guarantee that the potential energy correction converges with increasing distance. This is true for the Lennard-Jones equation, for which the long-range correction can be analytically calculated:[117]

$$\begin{aligned} U_{r>r_c}^{vdW} &= 8\pi \langle \rho \rangle \varepsilon \int_{r_c}^{\infty} r^2 \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^{6} \right] \mathrm{d}r \\ &= \frac{8}{3} \pi \langle \rho \rangle \varepsilon \sigma^3 \left[\frac{1}{3} \left(\frac{\sigma}{r} \right)^{9} - \left(\frac{\sigma}{r} \right)^{3} \right] \end{aligned} \tag{2.11}$$

This therefore allows the effect of not explicitly calculating long-range Lennard-Jones interactions to be rather easily corrected.

However, as the electrostatic interactions between point charges decay as r^{-1}, the potential energy correction is not guaranteed to converge to a finite value at infinitely large distances — additionally, the contributions from positive and negative charges both diverge.[116,117] Therefore, the long-range correction for electrostatic interactions cannot be calculated analytically, and more elaborate methods, such as Ewald summation,[141] are therefore required. In order to make the sum of electrostatic interactions converge more rapidly, the Ewald method screens the charges by adding a Gaussian charge distribution of the opposite sign to the position of each point charge. This ensures that the direct interactions decay very quickly with respect to distance, and all interactions will be negligible at the cutoff distance. In order to correct for the neutralising charge distribution added, a second set of Gaussian charges is added, which exactly cancel the neutralising distribution — problematically, the interactions between these Gaussians and the point charges decay very slowly with respect to distance. However, provided that the simulated system is periodic, a Fourier transform can be applied, and it transpires that the interactions involving this second set of Gaussians converge very quickly in Fourier space. When using Ewald summation, Eq. 2.7 is therefore replaced

with the following:[141]

$$U_{ele}^{Ewald} = \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{q_i q_j \,\mathrm{erfc}\left(\sqrt{\alpha} r_{ij}\right)}{r_{ij}} + \frac{1}{2V}\sum_{\mathbf{k}\neq\mathbf{0}} \frac{4\pi}{k^2} |\rho\left(\mathbf{k}\right)|^2 \exp\left(-\frac{k^2}{4\alpha}\right) - \left(\frac{\alpha}{\pi}\right)^{\frac{1}{2}} \sum_{i=1}^{N} q_i^2 \tag{2.12}$$

where:

$$\mathrm{erfc}\left(x\right) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-x^2} \mathrm{d}x \tag{2.13}$$

$$\rho\left(\mathbf{k}\right) = \sum_{i=1}^{N} q_i e^{\mathrm{i}\mathbf{k}\cdot\mathbf{r_i}} \tag{2.14}$$

α is a parameter of the method (which defines the width of the Gaussians), V is the simulation volume, $\mathbf{k}$ is a vector in Fourier space, $\mathrm{i} = \sqrt{-1}$, and $\mathbf{r_i}$ is the position vector of atom i — note that the above is written in Gaussian units (for notational simplicity), hence the absence of the $(4\pi\varepsilon_0)^{-1}$ factor found in Eq. 2.7. The first term in Eq. 2.12 is the summation of all direct interactions between screened charges, which is calculated in real space. The second term is the interaction of all point charges with the second set of Gaussians (those which cancel the neutralising distribution), which is calculated in Fourier space. The calculation of the second term includes interactions between each point charge and the Gaussian distribution (of the same sign) centred on the same position — which is not correct — and the third term is a correction to account for this. The real space sum converges more rapidly for large values of α, whereas the Fourier space sum converges more rapidly for small values. When suitably optimised, the cost of Ewald summation can scale as $\mathcal{O}(N^{\frac{3}{2}})$.[116,117]

However, for very large systems, even the improved scaling of $\mathcal{O}(N^{\frac{3}{2}})$ offered by Ewald summation is prohibitively expensive. A more efficient approach to the inclusion of long-range electrostatic interactions is the Particle Mesh Ewald (PME) method.[142] This involves the resolution of the charge distribution onto a lattice (or mesh), which then allows the utilisation of the Fast Fourier Transform (FFT) technique, which greatly increases the efficiency of the Fourier space sum. This further improves the scaling of the computational cost to $\mathcal{O}(N \ln N)$, which allows PME to be routinely applied to simulations of large systems.

2.2 Classical Statistical Mechanics

Statistical mechanics is essential to computational chemistry, as this allows the particle configurations simulated at the nano- or microscopic scale to be related to macroscopic observables.[116,117,143] Here, we are primarily concerned with the equilibrium probabilities of microstates in different ensembles, as these allow us to calculate correctly weighted averages of some property of interest. For example, for some hypothetical property, A, the ensemble average is calculated as a weighted average of the value of A over all possible microstates of a system:

$$\langle A \rangle = \int \int A(\mathbf{r}^N, \mathbf{p}^N) \rho(\mathbf{r}^N, \mathbf{p}^N) \mathrm{d}\mathbf{r}^N \mathrm{d}\mathbf{p}^N \tag{2.15}$$

where $\langle \dots \rangle$ represents an ensemble average, $A(\mathbf{r}^N, \mathbf{p}^N)$ is the value of A for microstate $(\mathbf{r}^N, \mathbf{p}^N)$, and $\rho(\mathbf{r}^N, \mathbf{p}^N)$ is the equilibrium probability density of the microstate.

An ensemble represents a collection of related microstates, all of which obey several conditions specific to that ensemble. These commonly involve a set of parameters which are held constant — different ensemble types are defined by the constraints that they set. For example, the canonical ensemble is defined by constant particle number (N), volume (V) and temperature (T), and as such, is referred to as the NVT ensemble. In this section, some of the common ensemble choices are described, as well as the grand canonical ensemble, which is central to this work.

2.2.1 Canonical Ensemble

The canonical (NVT) ensemble is defined with constant particle number, volume and temperature, and is the simplest ensemble discussed here. The system is considered to be in contact with a thermal reservoir of constant temperature, with which the system exchanges energy in the form of heat, in order to maintain constant temperature. The partition function of an ensemble represents the number of microstates accessible under the prescribed conditions, and also serves as a normalisation constant for the probabilities of individual microstates. The canonical partition function, Q_{NVT}, is related to the Helmholtz free energy via:[116,117,143]

$$F = -k_B T \ln Q_{NVT} \tag{2.16}$$

where k_B is Boltzmann's constant. The following partial derivatives relate the Helmholtz free energy to several thermodynamic properties of the ensemble:[143]

$$\left(\frac{\partial F}{\partial N} \right)_{V,T} = \mu \tag{2.17}$$

$$\left(\frac{\partial F}{\partial V}\right)_{N,T} = -P \tag{2.18}$$

$$\left(\frac{\partial F}{\partial T}\right)_{N,V} = -S \tag{2.19}$$

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

For a system which can only occupy a finite set of discrete microstates, the canonical partition function is simply calculated from the following sum over all microstates:

$$Q_{NVT} = \sum_i e^{-\beta E_i} \tag{2.20}$$

where E_i denotes the total energy of the i^{th} ensemble member, and $\beta = (k_B T)^{-1}$ is the thermodynamic beta. However, for the classical systems simulated in this work, the total energy does not occupy a finite set of discrete values, but is rather a function of particle positions and momenta, both of which are continuous. Therefore, the canonical partition function must be calculated by integrating over all positions and momenta:

$$Q_{NVT} = \frac{1}{h^{3N} N!} \int \int e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} \mathrm{d}\mathbf{r}^N \mathrm{d}\mathbf{p}^N \tag{2.21}$$

where h is Planck's constant. The h^{3N} term serves to ensure that the partition function is unitless, and the $(N!)^{-1}$ term arises from the fact that, if the particles are identical, there are $N!$ possible arrangements which would appear to be the same configuration — if the particles are not identical, then this $N!$ term is not necessary. Given that the total energy is composed of potential and kinetic terms (Eq. 2.2), and that the positions and momenta are separable, the two integrals can be calculated separately. When exponentiated, the kinetic term yields a Gaussian function of momentum and the integral over the momenta can therefore be carried out analytically:

$$\int_{-\infty}^{+\infty} \exp\left\{ -\sum_{i=1}^{N} \frac{|\mathbf{p_i}|^2}{2mk_B T} \right\} \mathrm{d}\mathbf{p}^N = (2\pi m k_B T)^{\frac{3N}{2}} \tag{2.22}$$

This separation allows the canonical partition function to be simplified:

$$\begin{aligned} Q_{NVT} &= \frac{1}{h^{3N} N!} \int \int e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} \mathrm{d}\mathbf{r}^N \mathrm{d}\mathbf{p}^N \\ &= \frac{1}{h^{3N} N!} \int_{-\infty}^{+\infty} \exp\left\{ -\sum_{i=1}^{N} \frac{|\mathbf{p_i}|^2}{2mk_B T} \right\} \mathrm{d}\mathbf{p}^N \int e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N \\ &= \frac{1}{\Lambda^{3N} N!} \int e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N \end{aligned} \tag{2.23}$$

where Λ is the thermal wavelength of a particle, defined as:

$$\Lambda = \left(\frac{h^2}{2\pi m k_B T} \right)^{\frac{1}{2}} \tag{2.24}$$

Additionally, the particle positions can be re-written in terms of scaled coordinates, $\mathbf{s}^N$, where all coordinates are scaled to lie between 0 and 1:

$$Q_{NVT} = \frac{V^N}{\Lambda^{3N} N!} \int_0^1 e^{-\beta U(\mathbf{s}^N;\, V)} \mathrm{d}\mathbf{s}^N \tag{2.25}$$

where $U(\mathbf{s}^N;\ V)$ indicates that the potential energy is calculated as a function of the real coordinates, not the scaled values. Here, the partition function can be viewed as a product of ideal and excess contributions:

$$Q_{NVT}^{id} = \frac{V^N}{\Lambda^{3N} N!} \tag{2.26}$$

$$Q_{NVT}^{ex} = \int_0^1 e^{-\beta U(\mathbf{s}^N;\, V)} \mathrm{d}\mathbf{s}^N \tag{2.27}$$

where Q_{NVT}^{id} is the partition function of an equivalent ideal gas, and Q_{NVT}^{ex} is the contribution from particle interactions — if there are no intermolecular interactions in the system, then $Q_{NVT}^{ex} = 1$. The Helmholtz free energy can therefore also be separated into ideal and excess components:

$$\begin{aligned} F &= -k_B T \ln Q_{NVT}^{id} - k_B T \ln Q_{NVT}^{ex} \\ &= F^{id} + F^{ex} \end{aligned} \tag{2.28}$$

The probability density of a microstate in the canonical ensemble is:[116,117,143]

$$\rho_{NVT}(\mathbf{r}^N, \mathbf{p}^N) = Q_{NVT}^{-1} \frac{1}{h^{3N} N!} e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} \tag{2.29}$$

We are often only interested in the probability of observing a particular particle configuration, and as such, the momentum contribution can be integrated out:

$$\begin{aligned} \rho_{NVT}(\mathbf{r}^N) &= Q_{NVT}^{-1} \frac{1}{h^{3N} N!} e^{-\beta U(\mathbf{r}^N)} \int_{-\infty}^{+\infty} \exp\left\{ -\sum_{i=1}^{N} \frac{|\mathbf{p_i}|^2}{2 m k_B T} \right\} \mathrm{d}\mathbf{p}^N \\ &= Q_{NVT}^{-1} \frac{1}{\Lambda^{3N} N!} e^{-\beta U(\mathbf{r}^N)} \\ &= \frac{e^{-\beta U(\mathbf{r}^N)}}{\int e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N} \end{aligned} \tag{2.30}$$

where the probability density of each configuration is therefore dependent only on the

potential energy, normalised by the configurational integral. The ensemble average of a property, A, in the canonical ensemble can therefore be calculated as:

$$\langle A \rangle_{NVT} = \frac{\int A(\mathbf{r}^N) e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N}{\int e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N} \tag{2.31}$$

2.2.2 Isothermal-Isobaric Ensemble

The isothermal-isobaric ensemble allows simulations to be carried out at constant pressure (P), by considering the system to be in equilibrium with an ideal gas at fixed pressure.[117] The system is assumed to be in contact with this ideal gas, such that the pressure is maintained by a hypothetical piston which increases or decreases the volume of the system, in order to maintain the pressure equivalent to that of the gas. The probability of each microstate is weighted by a factor of $\exp(-\beta PV)$, in order to account for the probability of the current volume, given the system pressure.[143] The partition function for this ensemble must therefore also include an integration over all possible microstate volumes, as well as particle positions and momenta:

$$Z_{NPT} = \frac{\beta P}{h^{3N} N!} \int \int \int e^{-\beta PV} e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} \mathrm{d}\mathbf{r}^N \mathrm{d}\mathbf{p}^N \mathrm{d}V \tag{2.32}$$

where the βP term is included to ensure that the partition function is unitless.[117] The above can be simplified by integrating the momenta, and scaling the coordinates:

$$Z_{NPT} = \frac{\beta P}{\Lambda^{3N} N!} \int V^N e^{-\beta PV} \mathrm{d}V \int e^{-\beta U(\mathbf{s}^N;\, V)} \mathrm{d}\mathbf{s}^N \tag{2.33}$$

The isothermal-isobaric partition function is related to the Gibbs free energy of the system as:

$$G = -k_B T \ln Z_{NPT} \tag{2.34}$$

The Gibbs free energy can be related to other thermodynamic properties, via the following partial derivatives:[143]

$$\left(\frac{\partial G}{\partial N} \right)_{P,T} = \mu \tag{2.35}$$

$$\left(\frac{\partial G}{\partial P} \right)_{N,T} = \langle V \rangle_{N,T} \tag{2.36}$$

$$\left(\frac{\partial G}{\partial T} \right)_{N,P} = -S \tag{2.37}$$

The probability density of a given set of positions, $\mathbf{r}^N$, in a volume of V in the isothermal-isobaric ensemble can be calculated as:

$$\begin{aligned}\rho_{NPT}(\mathbf{r}^N, V) &= Z_{NPT}^{-1}\frac{\beta P}{\Lambda^{3N}N!}e^{-\beta PV}e^{-\beta U(\mathbf{r}^N)} \\ &= Z_{NPT}^{-1}\frac{\beta PV^N}{\Lambda^{3N}N!}e^{-\beta PV}e^{-\beta U(\mathbf{s}^N;V)}\end{aligned} \tag{2.38}$$

Therefore, the ensemble average of a quantity, A, can be written as:

$$\langle A\rangle_{NPT} = Z_{NPT}^{-1}\frac{\beta P}{\Lambda^{3N}N!}\int\int A(\mathbf{s}^N, V)V^N e^{-\beta PV}e^{-\beta U(\mathbf{s}^N;\,V)}\mathrm{d}V\mathrm{d}\mathbf{s}^N \tag{2.39}$$

2.2.3 Grand Canonical Ensemble

The grand canonical (μVT) ensemble is distinctive in comparison to the NVT and NPT ensembles, in that the particle number need not be constant. Instead, the chemical potential, μ, of the system is held constant, and the particle number is free to vary accordingly — this ensemble can therefore be considered a sum of canonical ensembles.[144] Here, the system is considered to be in equilibrium with an ideal gas reservoir, where particles can be exchanged between the system and the reservoir.[117] The characteristic state function (analogous to the Helmholtz and Gibbs free energies) of the grand canonical ensemble is the grand potential:

$$\Omega = -k_B T\ln\Xi_{\mu VT} \tag{2.40}$$

The following partial derivatives hold for the grand potential:[143]

$$\left(\frac{\partial\Omega}{\partial\mu}\right)_{V,T} = -\langle N\rangle_{V,T} \tag{2.41}$$

$$\left(\frac{\partial\Omega}{\partial V}\right)_{\mu,T} = -P \tag{2.42}$$

$$\left(\frac{\partial\Omega}{\partial T}\right)_{\mu,V} = -S \tag{2.43}$$

Owing to the importance of the grand canonical ensemble in this work, the partition function for this ensemble is derived below.

As previously mentioned, here we consider the system and ideal gas as a large, canonical ensemble, containing M particles within a volume of V — the system contains N particles in a volume of V_s, leaving the ideal gas with $M-N$ particles in a volume of $V-V_s$. First, we consider one specific arrangement of N and $M-N$ particles across the two volumes, where no exchange of particles is possible. The partition function for

this arrangement is:

$$\begin{aligned} Q_{MVT} &= \frac{1}{h^{3M}M!}\int\int e^{-\beta E(\mathbf{r}^M,\mathbf{p}^M)}\mathrm{d}\mathbf{r}^M\mathrm{d}\mathbf{p}^M \\ &= \frac{N!(M-N)!}{M!}Q_{NV_sT}Q_{(M-N)(V-V_s)T} \end{aligned} \tag{2.44}$$

where:

$$Q_{NV_sT} = \frac{1}{h^{3N}N!}\int\int e^{-\beta E_s(\mathbf{r}^N,\mathbf{p}^N)}\mathrm{d}\mathbf{r}^N\mathrm{d}\mathbf{p}^N \tag{2.45}$$

$$Q_{(M-N)(V-V_s)T} = \frac{1}{h^{3(M-N)}(M-N)!}\int\int e^{-\beta E_i(\mathbf{r}^{M-N},\mathbf{p}^{M-N})}\mathrm{d}\mathbf{r}^{M-N}\mathrm{d}\mathbf{p}^{M-N} \tag{2.46}$$

We can now construct the full partition function by considering all possible distributions of the M particles over the two volumes:

$$Q_{MVT} = \sum_{N=0}^{M} g(N, M-N)\frac{N!(M-N)!}{M!}Q_{NV_sT}Q_{(M-N)(V-V_s)T} \tag{2.47}$$

where $g(N, M-N)$ represents the number of possible ways in which the M particles can be separated into two groups of size N and $M-N$. This degeneracy can be calculated by the binomial coefficient:

$$g(N, M-N) = \frac{M!}{N!(M-N)!} \tag{2.48}$$

Substituting this into the above, we obtain:

$$Q_{MVT} = \sum_{N=0}^{M} Q_{NV_sT}Q_{(M-N)(V-V_s)T} \tag{2.49}$$

The partition function of the ideal gas (for a given value of $M-N$) is related to the Helmholtz free energy of the reservoir, and when $M \gg N$ and $V \gg V_s$, this can be calculated by expanding about the free energy of the combined system, F_{MVT}:[143]

$$\begin{aligned} F_{(M-N)(V-V_s)T} &\approx F_{MVT} - N\frac{\partial F}{\partial N} - V_s\frac{\partial F}{\partial V} \\ &\approx F_{MVT} - \mu N + PV_s \end{aligned} \tag{2.50}$$

This expression for the free energy of the reservoir can be used to replace the corresponding partition function:

$$\begin{aligned} Q_{MVT} &= \sum_{N=0}^{M} e^{-\beta(F_{MVT}-\mu N+PV_s)}Q_{NV_sT} \\ &= e^{-\beta PV_s}Q_{MVT}\sum_{N=0}^{M} e^{\beta\mu N}Q_{NV_sT} \end{aligned} \tag{2.51}$$

At this point, Q_{MVT} appears on both sides of the equation and therefore cancels, leaving the following relationship:

$$e^{\beta P V_s} = \sum_{N=0}^{M} e^{\beta \mu N} Q_{N V_s T} \tag{2.52}$$

It should be noted that there is now no explicit reference to the ideal gas reservoir (except for M, which is taken to be infinitely large). We can therefore consider this relationship only in terms of the grand canonical system, replacing V_s with V, which is the volume of this grand canonical ensemble:

$$e^{\beta P V} = \sum_{N=0}^{\infty} e^{\beta \mu N} Q_{NVT} \tag{2.53}$$

Here, it is convenient to write the grand potential as:[143]

$$\begin{aligned} \Omega &= V \frac{\partial \Omega}{\partial V} \\ &= -PV \end{aligned} \tag{2.54}$$

which provides a direct relationship between PV and the grand canonical partition function:

$$k_B T \ln \Xi_{\mu VT} = PV \tag{2.55}$$

Therefore, substituting this into Eq. 2.53 yields an expression for the partition function:

$$\Xi_{\mu VT} = \sum_{N=0}^{\infty} e^{\beta \mu N} Q_{NVT} \tag{2.56}$$

As previously mentioned, the above shows how the grand canonical ensemble can be considered a weighted sum of canonical ensembles with different numbers of particles (referred to as petite ensembles by Gibbs[144]).

For a microstate in the grand canonical ensemble containing N particles at positions, $\mathbf{r}^N$, with momenta, $\mathbf{p}^N$, the probability density is given by:

$$\rho_{\mu VT}(\mathbf{r}^N, \mathbf{p}^N) = \Xi_{\mu VT}^{-1} \frac{e^{\beta \mu N}}{h^{3N} N!} e^{-\beta E(\mathbf{r}^N, \mathbf{p}^N)} \tag{2.57}$$

As before, in order to obtain the probability density of a particular configuration, the momenta can be integrated, except that these terms no longer cancel:

$$\rho_{\mu VT}(\mathbf{r}^N) = \Xi_{\mu VT}^{-1} \frac{e^{\beta \mu N}}{\Lambda^{3N} N!} e^{-\beta U(\mathbf{r}^N)} \tag{2.58}$$

Therefore, the ensemble average of a property, A, in the grand canonical ensemble can be calculated as:

$$\langle A \rangle_{\mu VT} = \Xi_{\mu VT}^{-1} \sum_{N=0}^{\infty} \frac{e^{\beta \mu N}}{\Lambda^{3N} N!} \int A(\mathbf{r}^N) e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N \tag{2.59}$$

If we are interested in the probability density of observing a particular configuration, regardless of particle labels (which are physically meaningless), the factorial term is sometimes not included — this essentially sums the probability densities of the $N!$ sets of particle labels which can be arbitrarily assigned to each configuration of N particles.

2.3 Chemical Potential

The chemical potential defines the direction in which particles flow, where areas of low chemical potential are more favourable. Under conditions of constant volume and temperature, the chemical potential can be determined as the derivative of the Helmholtz free energy with respect to the number of particles[116,117,143] (Eq. 2.17), or, under conditions of constant pressure and temperature, as the derivative of the Gibbs free energy with respect to the particle number (Eq. 2.35). For simplicity, the discussion in this section focuses on the former. The chemical potential therefore indicates the extent to which adding an additional particle to a molecular system is favourable.

As previously mentioned, the Helmholtz free energy can be written in terms of ideal and excess contributions (Eq. 2.28). The chemical potential can therefore also be rewritten in such a way:

$$\begin{aligned} \mu &= \frac{\partial}{\partial N}(F^{id} + F^{ex}) \\ &= \frac{\partial F^{id}}{\partial N} + \frac{\partial F^{ex}}{\partial N} \\ &= \mu^{id} + \mu^{ex} \end{aligned} \tag{2.60}$$

where μ^{id} is the ideal chemical potential, which is the chemical potential of an analogous ideal gas, and μ^{ex} is the excess chemical potential, which is the difference made to the chemical potential by the inclusion of potential energy interactions between particles. The calculations of these two contributions are discussed in this section.

2.3.1 Ideal Chemical Potential

For a system containing N particles within a volume of V, the ideal component of the canonical partition function can be calculated analytically (Eq. 2.26). Therefore, the

corresponding Helmholtz free energy can also be determined analytically:[117]

$$\begin{aligned} F^{id}(N) &= -k_B T \ln Q^{id}_{NVT} \\ &= -k_B T \ln \left(\frac{V^N}{\Lambda^{3N} N!} \right) \end{aligned} \tag{2.61}$$

For a very large ideal gas, we can use Stirling's approximation (which is valid for very large numbers) to rewrite the above:

$$F^{id}(N) \approx -k_B T \ln \left(N \ln \left(\frac{V}{\Lambda^3} \right) - N \ln N + N \right) \tag{2.62}$$

The derivative of this free energy, and therefore, the ideal chemical potential, can be calculated as

$$\mu^{id} = \frac{\partial F^{id}}{\partial N} = -k_B T \ln \left(\frac{V}{N \Lambda^3} \right) \tag{2.63}$$

which can be rewritten in terms of the number density of the ideal gas, ρ_{ideal}, as:

$$\mu^{id} = k_B T \ln \left(\rho_{ideal} \Lambda^3 \right) \tag{2.64}$$

2.3.2 Excess Chemical Potential

The excess Helmholtz free energy of a system can, in principle, also be calculated from the excess contribution to the canonical partition function:[117]

$$\begin{aligned} F^{ex}(N) &= -k_B T \ln Q^{ex}_{NVT} \\ &\approx -k_B T \ln \left\{ \int_0^1 e^{-\beta U(\mathbf{s}^N;\, V)} \mathrm{d}\mathbf{s}^N \right\} \end{aligned} \tag{2.65}$$

Note that this integral is over scaled coordinates, such that if the interaction energy is always zero, the integral evaluates to one, and the excess free energy is therefore zero. However, the configurational integral above can only be solved analytically for the simplest possible systems, which are of little interest here. For the large, biomolecular systems of interest in this work, the excess free energy (and therefore, the excess chemical potential) cannot be calculated analytically, so numerical methods must be used. These methods involve rewriting the derivative of the excess free energy from Eq. 2.60 as a finite difference derivative, where $\Delta N = 1$:

$$\begin{aligned} \mu^{ex} &\approx \frac{\Delta F^{ex}}{\Delta N} \\ &= F^{ex}(N+1) - F^{ex}(N) \end{aligned} \tag{2.66}$$

Thus enabling the excess chemical potential to be calculated as the excess free energy of adding an additional particle.

The Widom particle insertion method is a relatively simple approach to calculate the excess free energy of adding an additional particle to a system.[145]

$$\begin{aligned} F^{ex}(N+1) - F^{ex}(N) &= -k_B T \ln \frac{Q^{ex}_{(N+1)VT}}{Q^{ex}_{NVT}} \\ &= -k_B T \ln \left\{ \frac{\int e^{-\beta U(\mathbf{s}^{N+1};\, V)} \mathrm{d}\mathbf{s}^{N+1}}{\int e^{-\beta U(\mathbf{s}^{N};\, V)} \mathrm{d}\mathbf{s}^{N}} \right\} \end{aligned} \tag{2.67}$$

Where we now have a ratio of two integrals which we cannot easily solve. This problem is avoided by rewriting the potential energy of the $N+1$ state as:

$$U(\mathbf{s}^{N+1};\, V) = U(\mathbf{s}^{N};\, V) + \Delta U_{N+1} \tag{2.68}$$

where ΔU_{N+1} refers to the interaction energy of the $(N+1)^{\text{th}}$ particle with all other particles in the system. This allows the excess chemical potential to be written in terms of this energy change:

$$\mu^{ex} = -k_B T \ln \left\{ \int \left\langle e^{-\beta \Delta U_{N+1}} \right\rangle_N \mathrm{d}\mathbf{s}_{N+1} \right\} \tag{2.69}$$

where the integral is carried out over all positions of the $(N+1)^{\text{th}}$ particle, and $\langle \dots \rangle_N$ indicates an ensemble average over the system containing N particles. In practice, the integral is solved using the Monte Carlo method, which involves generating configurations according to the equilibrium distribution of the N-particle system, and at regular intervals, the additional particle is assigned to a random position and the exponential term in Eq. 2.69 is calculated. This term is then averaged over all samples, and then the natural logarithm is taken of this value, from which the excess chemical potential is then calculated.[117]

However, a problem with the Widom method is that it requires a significant degree of phase space overlap between the systems with N and $N+1$ particles — that is, the microstates sampled from the equilibrium distribution of the N-particle system must be such that inserting a particle at random locations generates microstates which are reasonably populated in the equilibrium distribution of the $N+1$ system.[117] It should be noted that this is not a theoretical requirement of the method, but that the convergence of the result obtained will be very slow when the phase space overlap is poor — theoretically, the result will eventually converge in the limit of infinite sampling. For this reason, the Widom method is better suited to low density systems, where there is plenty of free space for the additional particle to favourably occupy. For condensed phase systems, such as the biomolecular systems of interest in this work, this becomes increasingly unlikely, as the vast majority of random insertions will be met with a steric clash, resulting in a large, positive difference in potential energy. This will cause the excess chemical potential to be calculated as significantly more positive than its true

value. For this reason, free energy perturbation methods (discussed in section 2.4) are better suited to the calculation of the excess chemical potential for condensed systems, as they make use of intermediate states which reduce the problem of phase space overlap.

2.4 Free Energy Calculations

Thus far, free energies have primarily been discussed in terms of their relationship with the chemical potential via the number of particles. In modern computational chemistry, it is often of interest to calculate the difference in free energy between two distinct states — denoted A and B — which are often related by some chemical change of interest. For example, relative binding free energy calculations seek to determine the affinity difference between a pair of compounds, by calculating the difference in free energy between them in both the protein binding site and bulk water, under identical conditions. A free energy difference can be written in terms of the partition functions (this is shown for the canonical ensemble, and can be written similarly for other ensembles):[116,117]

$$\begin{aligned} \Delta F_{AB} &= F_B - F_A \\ &= -k_B T \ln \frac{Q_B}{Q_A} \\ &= -k_B T \ln \left\{ \frac{\int e^{-\beta U_B(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N}{\int e^{-\beta U_A(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N} \right\} \end{aligned} \tag{2.70}$$

where U_A and U_B denote the potential energy functions associated with states A and B, respectively. Note that in the above, the states are assumed to have the same number of particles — this is often ensured by using 'dummy' particles which do not contribute to the potential energy. As seen with the Widom particle insertion method, the free energy difference therefore depends on the ratio of a pair of integrals which cannot be solved in practice, for all but the simplest cases.

One approach to the calculation of the ratio of integrals given in Eq. 2.70 is the exponential averaging method. This method makes use of multiplying the integral ratio by a factor of $e^x e^{-x} = 1$ to calculate the ratio of partition functions as follows:

$$\begin{aligned} \frac{Q_B}{Q_A} &= \frac{\int e^{-\beta U_B(\mathbf{r}^N)} e^{\beta U_A(\mathbf{r}^N)} e^{-\beta U_A(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N}{\int e^{-\beta U_A(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N} \\ &= \frac{\int e^{-\beta \Delta U_{AB}(\mathbf{r}^N)} e^{-\beta U_A(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N}{\int e^{-\beta U_A(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N} \\ &= \left\langle e^{-\beta \Delta U_{AB}} \right\rangle_A \end{aligned} \tag{2.71}$$

where $\Delta U_{AB} = U_B - U_A$, and the ensemble average is carried out over state A. Therefore, the ratio of integrals can be much more easily calculated via this ensemble average. Combining the above with Eq. 2.70 yields the Zwanzig equation:[146]

$$\Delta F_{AB} = -k_B T \ln \left\langle e^{-\beta \Delta U_{AB}} \right\rangle_A \tag{2.72}$$

However, as with the Widom particle insertion method,[145] the calculation of free energies using the Zwanzig equation is severely limited, owing to the requirement that the phase spaces of states A and B overlap significantly. If this is not the case, then high probability microstates from state B will be sampled infrequently (or perhaps never) during a finite simulation of state A, and therefore biasing the free energy result obtained. A solution to the issue of phase space overlap is the use of alchemical perturbations, via a scaling coordinate, λ, where the potential energy can be calculated as:[116,117]

$$U(\lambda) = \lambda U_B + (1 - \lambda) U_A \tag{2.73}$$

where $U(\lambda = 0) = U_A$ and $U(\lambda = 1) = U_B$ — note that λ values between 0 and 1 correspond to non-physical states (hence the use of the term alchemical). Note that other methods for combining the potential energy function via λ exist.

When calculating the free energy change associated with adding or removing a particle from a system (as is needed in the calculation of the excess chemical potential, for example), practical issues can occur for λ values where the molecule is largely non-interacting.[116,117] Notably, the widely-used Lennard-Jones potential (Eq. 2.8) at short distances does not go smoothly to zero as λ goes to zero (using the convention that $\lambda = 0$ corresponds to the non-interacting state). The constant presence of short-range repulsions limits phase space overlap with the non-interacting state, and therefore, the efficiency of the calculation. A solution is to 'soften' the interactions, using so-called softcore potentials,[147] such as:

$$U_{ij}^{vdW}(r_{ij}, \lambda) = 4\lambda\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}^{eff}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}^{eff}} \right)^{6} \right] \tag{2.74}$$

where the effective distance is calculated as:

$$r_{ij}^{eff}(\lambda) = \sigma_{ij} \left(\frac{1 - \lambda}{2} + \left(\frac{r_{ij}}{\sigma_{ij}} \right)^6 \right)^{\frac{1}{6}} \tag{2.75}$$

It should be noted that many forms of softcore potential exist, the above is just one example. However, the presence of finite Lennard-Jones interactions at $r_{ij} = 0$ introduces the possibility that attractive electrostatic interactions may cause particle centres to converge as they overcome the finite steric repulsions, resulting in numerical instabilities. For this reason, a common approach is to separate λ into electrostatic and van der Waals components, λ_{ele} and λ_{vdW}, allowing these interactions to be decoupled separately. When decoupling a molecule (or vice versa for coupling), first the electrostatic interactions are decoupled, and then the van der Waals interactions are decoupled, removing the risk of electrostatic singularities. When this is the case, the electrostatic interactions can simply be scaled by λ, but if the electrostatic and van der Waals scaling is not separated, then the electrostatic interactions must also be softened.

Using alchemical perturbation, the free energy calculation between A and B can be separated into $M-1$ smaller free energy calculations between adjacent λ values, where the phase space overlap will be much better (note that here, M refers to the number of λ values used):[116,117]

$$\Delta F_{AB} = -k_B T \sum_{i=1}^{M-1} \ln \left\langle e^{-\beta(U_{i+1}-U_i)} \right\rangle_{\lambda_i} \tag{2.76}$$

where $U_i = U(\lambda_i)$, $\lambda_1 = 0$ and $\lambda_M = 1$. This offers a significant improvement on the direct free energy calculation from A to B; however, this method is still rarely used in practice, owing to the availability of more efficient methods, some of which are discussed below.

2.4.1 Thermodynamic Integration

The thermodynamic integration method makes use of the fact that the partition function (and therefore the free energy) can be treated as a function of λ. Therefore, the free energy difference between states A and B can be written as the integral of the derivative of the free energy with respect to λ:[116,117]

$$\Delta F_{AB} = \int_0^1 \frac{\partial F(\lambda)}{\partial \lambda} \mathrm{d}\lambda \tag{2.77}$$

This derivative can be rewritten as:

$$\begin{aligned}\frac{\partial F(\lambda)}{\partial \lambda} &= -k_B T \frac{\partial}{\partial \lambda} \ln Q(\lambda) \\ &= \frac{-k_B T}{Q(\lambda)} \frac{\partial Q(\lambda)}{\partial \lambda} \\ &= \frac{1}{Q(\lambda)} \frac{1}{\Lambda^{3N} N!} \int \frac{\partial U(\mathbf{r}^N, \lambda)}{\partial \lambda} e^{-\beta U(\mathbf{r}^N, \lambda)} \mathrm{d}\mathbf{r}^N \\ &= \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda\end{aligned} \tag{2.78}$$

where the derivative of the free energy can be equated to the ensemble average of the derivative of the potential energy with respect to λ, allowing the free energy difference to be rewritten in terms of the latter:

$$\Delta F_{AB} = \int_0^1 \left\langle \frac{\partial U(\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{2.79}$$

where this derivative is typically calculated using numerical methods, such as:

$$\frac{\partial U(\lambda)}{\partial \lambda} \approx \frac{U(\lambda + \delta\lambda) - U(\lambda - \delta\lambda)}{2\delta\lambda} \tag{2.80}$$

where $\delta\lambda$ is suitably small.

An advantage of this method is that the derivatives at different values of λ are calculated independently. Therefore, if the number of λ values is deemed insufficient after running a set of simulations, additional λ-states can be added without requiring that the initial simulations be discarded.

2.4.2 Bennett Acceptance Ratio

The Bennett acceptance ratio (BAR) method is based on the fact that the ratio of partition functions is equivalent to a ratio of ensemble averages:[148]

$$\frac{Q_A}{Q_B} = \frac{\left\langle w e^{-\beta U_A} \right\rangle_B}{\left\langle w e^{-\beta U_B} \right\rangle_A} \tag{2.81}$$

which holds for any $w(\mathbf{r}^N)$. The free energy difference can therefore be written in terms of these ensemble averages as:

$$\beta \Delta F_{AB} = \ln \left\langle w e^{-\beta U_A} \right\rangle_B - \ln \left\langle w e^{-\beta U_B} \right\rangle_A \tag{2.82}$$

Bennett discovered that the statistical uncertainty of the free energy is minimised for the following choice of $w(\mathbf{r}^N)$:[148]

$$w(\mathbf{r}^N) = \frac{\text{constant}}{\frac{Q_A}{n_A} e^{-\beta U_B(\mathbf{r}^N)} + \frac{Q_B}{n_B} e^{-\beta U_A(\mathbf{r}^N)}} \tag{2.83}$$

where n_A and n_B are the number of statistically independent samples collected for states A and B, respectively. Substituting this into the above, we obtain:

$$e^{\beta \Delta F_{AB}} = \frac{\langle f\left(\beta(U_A - U_B + C)\right)\rangle_B}{\langle f\left(\beta(U_B - U_A - C)\right)\rangle_A} e^{\beta C} \tag{2.84}$$

where the above holds for any function f such that $f(x)/f(-x) = e^{-x}$, and any constant C with units of energy. However, statistically optimal results are obtained for the following choices:[148]

$$f(x) = \frac{1}{1 + e^x} \tag{2.85}$$

$$C = \beta^{-1} \ln \frac{Q_A n_B}{Q_B n_A} \tag{2.86}$$

It can be seen here that the optimal value of C depends on the ratio of partition functions, which is the unknown quantity of interest. However, one can begin with an initial guess and then optimise C in a self-consistent manner, based on the following equations:

$$\beta \Delta F_{AB} = \ln \left\{ \frac{\sum_B f\left(\beta(U_A - U_B + C)\right)}{\sum_A f\left(\beta(U_B - U_A - C)\right)} \right\} - \ln \frac{n_B}{n_A} + \beta C \tag{2.87a}$$

$$\beta \Delta F_{AB} = \beta C - \ln \frac{n_B}{n_A} \tag{2.87b}$$

where self-consistency is achieved when the following is true:

$$\sum_A f\left(\beta(U_B - U_A - C)\right) = \sum_B f\left(\beta(U_A - U_B + C)\right) \tag{2.88}$$

The BAR method is a significant improvement over the Zwanzig equation, as it makes use of sampling from both end states, rather than just one. However, this still requires (albeit to a lesser extent) significant phase space overlap between the two states. One approach to alleviate this issue is to, again, break the calculation into a series of smaller free energy calculations, which can be summed. However, the multistate Bennett acceptance ratio (MBAR) method[149] is a superior alternative to this. This method is an extension of BAR, such that data from an arbitrary number of λ values can be analysed, to provide free energy estimates between all pairs of λ values — in the case where only two values of λ are considered, MBAR reduces to the BAR method as described by

Bennett. MBAR has been shown to offer significantly better statistical performance than other free energy estimators.[149]

2.4.3 Nonequilibrium Methods

The free energy methods discussed thus far are known as equilibrium methods. This is because at each value of λ, the simulation is carried out at equilibrium. However, there also exists another class of methods, known as nonequilibrium free energy methods.[117] Rather than running equilibrium simulations at a range of λ values, these methods involve gradually switching the value of λ from 0 to 1, whilst calculating the work required to do so. This makes use of the fact that the free energy difference between two states is equal to the reversible work required to transform one state into the other:[117]

$$\Delta F = W_{\infty} \tag{2.89}$$

where the ∞ subscript indicates that the transformation from A to B is carried out infinitely slowly. In practice, the value of λ is increased incrementally, and the nonequilibrium work is calculated as the sum of the potential energy changes associated with each of these λ increments.

As simulated nonequilibrium protocols cannot be infinitely long, they are not strictly reversible and therefore Eq. 2.89 cannot be used to calculate free energy changes. The Jarzynski estimator calculates the free energy from a set of nonequilibrium work values using the following relationship:[150]

$$\Delta F_{AB} = -k_B T \ln \left\langle e^{-\beta W_{AB}} \right\rangle_A \tag{2.90}$$

where W_{AB} is the work required to perturb A into B, and this holds in the limit of infinite sampling — this can be either an infinite number of work samples, or a work calculated from an infinitely long protocol. In practice, this means that shorter switches will require a larger number of work values to produce an accurate estimate of the free energy. Additionally, the BAR method can be used to estimate the free energy, using work values for both the forward and reverse transformations.[151] This works just as described above for BAR analysis of equilibrium simulations, except that the potential energy differences are replaced with work values. It has been found empirically that nonequilibrium free energy differences calculated using BAR are more consistent than those calculated using the Jarzynski estimator.[152]

2.5 Molecular Dynamics Simulation

Thus far in this chapter, extensive reference has been made to the sampling of microstates from a given ensemble, with no mention as to how this is done. Molecular dynamics (MD) is a simulation technique which seeks to generate samples by propagating the system through time.[116,117] The time evolution of a set of particle positions can be classically modelled using Newton's second law of motion:

$$\begin{aligned} \mathbf{F} &= m\mathbf{a} \\ &= m\frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} \end{aligned} \tag{2.91}$$

The force on each particle in a molecular simulation can be calculated as the negative derivative of the potential energy:

$$\mathbf{F_i} = -\nabla_\mathbf{i} U(\mathbf{r}^N) \tag{2.92}$$

where $\nabla_\mathbf{i}$ indicates differentiation with respect to the positions of the i^{th} particle — for this reason, most force fields use potential energy expressions for which these derivatives can be readily calculated (section 2.1). The equations above can therefore be combined to directly relate the motion of the particles to the potential energy:

$$\frac{\mathrm{d}^2\mathbf{r_i}}{\mathrm{d}t^2} = -\frac{1}{m}\nabla_\mathbf{i} U(\mathbf{r}^N) \tag{2.93}$$

thus yielding the concept which underpins molecular dynamics simulations. The above can be re-written as the following system of ordinary differential equations (ODEs):

$$\frac{\mathrm{d}\mathbf{r_i}}{\mathrm{d}t} = \mathbf{v_i} \tag{2.94a}$$

$$\frac{\mathrm{d}\mathbf{v_i}}{\mathrm{d}t} = -\frac{1}{m}\nabla_\mathbf{i} U(\mathbf{r}^N) \tag{2.94b}$$

where $\mathbf{v_i}$ is the velocity of the i^{th} particle.

However, for systems containing three or more particles, Eq. 2.94 becomes a many body problem, as the motions of all particles are coupled. Therefore, the time evolution of a many particle system cannot be solved analytically, and numerical integration techniques must be employed, as discussed below.

2.5.1 Integration

Given an initial set of positions and velocities, the system of ODEs in Eq. 2.94 can be treated as an initial value problem — the positions are often obtained from experimental data or a structural model, and velocities can be directly drawn from their equilibrium distribution for a given temperature. Therefore, starting from a given microstate, the positions and velocities can be integrated by making small, discrete jumps in time (known as a timestep, δt) — the smaller the timestep, the more accurately the true solution is represented, at increased computational cost per unit time.[116,117] There are a number of approaches to evolve a system in this way, with some of the more common algorithms described in this section.

The simplest approach for an initial value problem is Euler's method:

$$f(x + \delta x) = f(x) + \delta x \frac{\mathrm{d}f}{\mathrm{d}x} \tag{2.95}$$

When applied to Eq. 2.94, this yields the following algorithm:

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) \tag{2.96a}$$

$$\mathbf{v}(t + \delta t) = \mathbf{v}(t) + \delta t \mathbf{a}(t) \tag{2.96b}$$

However, the error of this method scales as $\mathcal{O}(\delta t)$ when δt is small (known as a first order method). This error is quickly accumulated, leading to unstable dynamics, and ultimately, incorrect results.

A more accurate algorithm is the Verlet integrator,[153] which makes use of Taylor series expansions of the positions about a point in time:

$$\mathbf{r}(t - \delta t) = \mathbf{r}(t) - \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) \tag{2.97a}$$

$$\mathbf{r}(t + \delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + \frac{1}{2}\delta t^2 \mathbf{a}(t) \tag{2.97b}$$

These two expressions can be combined to given an expression for $\mathbf{r}(t + \delta t)$, based on $\mathbf{r}(t)$ and $\mathbf{r}(t - \delta t)$:

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \delta t^2 \mathbf{a}(t) \tag{2.98}$$

The Verlet integrator is a second order method, and therefore much more stable. However, there are two issues with this method. The first, relatively minor issue is that two initial values are required: a set of positions at $t = t$, and another at $t = t - \delta t$ — a potential solution is to use Euler's method for the very first step, to generate the second microstate, and then continue the simulation with the Verlet integrator. The second,

more serious issue is that this integration algorithm has no dependence on the velocities, and therefore, no dependence on the temperature, meaning that it cannot be used to sample canonical ensembles[116] (see the following section for details on temperature control). A further improvement is the velocity Verlet integrator, which makes use of a half step for the velocities:[154]

$$\mathbf{v}(t+\frac{1}{2}\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t\mathbf{a}(t) \tag{2.99a}$$

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t+\frac{1}{2}\delta t) \tag{2.99b}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t+\frac{1}{2}\delta t) + \frac{1}{2}\delta t\mathbf{a}(t+\delta t) \tag{2.99c}$$

This method is also second order, but has the advantage that the temperature can be controlled via the velocities, allowing the simulation of canonical ensembles.[116] It should be noted that higher order integrators exist, but are not widely used.

2.5.2 Temperature Control

By default, the integrators discussed above should conserve the total energy of the simulation, and therefore sample the microcanonical (NVE) ensemble. In order to sample the canonical ensemble (or any other ensemble defined at constant temperature), molecular dynamics simulations make use of thermostats, which modify the velocities of the particles throughout the simulation to regulate the temperature. This makes use of the following relation, where the temperature can be related to the ensemble average of the kinetic energy:[116]

$$\left\langle \sum_{i=1}^{N} \frac{m_i|\mathbf{v_i}|^2}{2} \right\rangle = \frac{3}{2}Nk_BT \tag{2.100}$$

where, on average, each degree of freedom contributes $\frac{1}{2}k_BT$ to the kinetic energy — it should be noted that the application of constraints on a simulation reduces the number of degrees of freedom, and therefore has an impact on the above. There are a number of methods which aim to impose constant temperature on a simulation, some of which are described below.

The simplest approach to maintain a constant temperature is to simply rescale the velocities at regular intervals by a factor of $\sqrt{T/T(t)}$, where T is the desired temperature and $T(t)$ is the temperature calculated from the velocities at time, t.[154] However, this method will maintain the temperature exactly constant, with zero fluctuation, which is incorrect.[116] Additionally, this approach can result in 'hot' and 'cold' regions of a simulation, causing strange and incorrect dynamics.[116] It should be noted that algorithms also exist which apply velocity rescaling in a stochastic fashion which correctly samples the canonical ensemble.[155] A method which allows some fluctuation in the temperature

is the Berendsen thermostat, which, by considering the system to be in thermal equilibrium with a heat bath, causes the temperature to converge exponentially towards the desired value, rescaling the velocities by the following factor:[156]

$$\lambda^2 = 1 + \frac{\delta t}{\tau_T}\left(\frac{T}{T(t)} - 1\right) \tag{2.101}$$

where τ_T is a coupling constant, which controls the rate of convergence — if $\tau_T = \delta t$, this method reduces to the simple rescaling approach. However, whilst the Berendsen thermostat is an improvement over simple velocity rescaling, it still does not sample the correct canonical ensemble.[116]

One more rigorous approach is the Andersen thermostat.[157] Here it is assumed that the heat bath periodically emits a hypothetical 'thermal particle', which collides with atoms with the system, transferring kinetic energy. In practice, this involves selecting a random particle (or group of particles), and replacing the velocity vector with one drawn randomly from the Maxwell-Boltzmann distribution. The rigour with which the canonical ensemble is sampled is therefore dependent on the frequency of these stochastic collisions. A potential concern is that if these collisions are too frequent, discontinuous dynamics can result, which may be undesirable for some applications.[116]

A different approach for constant temperature simulation is the use of Langevin dynamics, where the heat bath can be considered as a medium through which the particles move and 'collide', exchanging kinetic energy.[116] Here, Eq. 2.92 is replaced with the following:

$$\mathbf{F_i} = -\nabla_{\mathbf{i}} U(\mathbf{r}^N) - \gamma m \mathbf{v_i} + (2\gamma m k_B T)^{\frac{1}{2}} \mathcal{N} \tag{2.102}$$

where γ is the friction coefficient and $\mathcal{N}$ is a random vector drawn from a normal distribution with a mean of 0 and a variance of 1 (also known as a Wiener process). When $\gamma = 0$, the dynamics will be completely deterministic, and the above is reduced to Eq. 2.92, and in the high friction limit, the above gives Brownian dynamics. Langevin integrators can be used to simulate at constant temperature, where one such example

is the BAOAB Langevin integrator (also denoted VRORV):[158,159]

$$\mathbf{v}(t+\frac{1}{2}\delta t) = \mathbf{v}(t) - \frac{1}{2m}\delta t \nabla U(\mathbf{r}(t)) \tag{2.103a}$$

$$\mathbf{r}(t+\frac{1}{2}\delta t) = \mathbf{r}(t) + \frac{1}{2}\delta t \mathbf{v}(t+\frac{1}{2}\delta t) \tag{2.103b}$$

$$\mathbf{v}'(t+\frac{1}{2}\delta t) = e^{-\gamma\delta t}\mathbf{v}(t+\frac{1}{2}\delta t) + \left(\frac{k_B T}{m}(1-e^{-2\gamma\delta t})\right)^{\frac{1}{2}} \mathcal{N}(t) \tag{2.103c}$$

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t+\frac{1}{2}\delta t) + \frac{1}{2}\delta t \mathbf{v}'(t+\frac{1}{2}\delta t) \tag{2.103d}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}'(t+\frac{1}{2}\delta t) - \frac{1}{2m}\delta t \nabla U(\mathbf{r}(t+\delta t)) \tag{2.103e}$$

The first and fifth steps are deterministic velocity updates (denoted 'B' or 'V'), the second and fourth steps are deterministic position updates (denoted 'A' or 'R'), and the third is a stochastic velocity update (denoted 'O'). These steps can be combined in any order (ABOBA instead of BAOAB, for example), but the behaviour of different Langevin integrators is not equivalent.[158,159] It has been empirically observed that the BAOAB variant is of particularly high quality, showing second order accuracy at low values of γ and fourth order accuracy at high values.[158] Additionally, it transpires that the samples generated using the BAOAB integrator obey the canonical equilibrium distribution very well.[159]

Other, more complex methods of temperature regulation also exist, such as the Nosé-Hoover thermostat.[160,161] This method rigorously treats the heat bath as an intrinsic part of the simulated system, which is counted as an additional degree of freedom. Such a system is referred to as an 'extended system', and a description of this method is beyond the scope of this section.[116]

2.5.3 Pressure Control

Just as thermostats are required to carry out simulations at constant temperature, barostats are required for constant pressure simulations. These algorithms maintain the pressure of the system by scaling the simulation volume (and particle coordinates). This can be thought of as connecting the simulated system to a pressure bath via a piston, which expands or compresses the volume of the system, in order to balance the pressure of the simulation with that of the bath.[116]

Analogously to temperature control, there also exist barostat implementations of the rescaling and Berendsen thermostats, except that these methods involve scaling of the simulation volume, instead of velocities.[156] However, as before, these methods do not produce the correct fluctuations in the system pressure/volume, and therefore, do not

sample the NPT ensemble correctly.[116] The Andersen barostat is an extended system method in which the hypothetical piston is considered an additional degree of freedom, and is able to rigorously sample the NPT ensemble.[157] Another rigorous approach is the Monte Carlo barostat, described in section 2.6.1.[117]

2.6 Monte Carlo Simulation

Monte Carlo (MC) is a statistical simulation technique which seeks to generate a set of samples of a given system, according to their equilibrium probabilities. Unlike molecular dynamics, Monte Carlo simulations have no time-dependence, and there is no requirement to generate a smooth or continuous transition between microstates. These simulations employ 'moves' which involve proposing random changes to the system, which are then accepted or rejected, according to the equilibrium likelihood of the proposed state. Whilst these simulations offer no dynamic information about a system of interest, they are able to sample the equilibrium distribution very accurately,[116,117] and offer the ability to sample degrees of freedom beyond particle rearrangements, as discussed later in this section.

In order to ensure that a simulation remains at equilibrium, the probability distribution must be stationary. This is typically imposed via the *detailed balance condition*, which ensures that, for any pair of microstates, x and y, there is no net flux of probability between them:

$$\pi(x)P(y|x)A(y|x) = \pi(y)P(x|y)A(x|y) \tag{2.104}$$

where $\pi(x)$ is the equilibrium probability of microstate x, $P(y|x)$ is the probability of proposing a move to y from x, and $A(y|x)$ is the probability of accepting that move. Therefore, we can use this condition to determine the probability of accepting a given move proposal, whilst ensuring that the equilibrium distribution is preserved. Eq. 2.104 can be rearranged into the following ratio of acceptance probabilities (also known as an acceptance ratio):

$$\frac{A(y|x)}{A(x|y)} = \frac{P(x|y)\pi(y)}{P(y|x)\pi(x)} \tag{2.105}$$

However, when deciding whether or not a specific move should be accepted or rejected, we need to be able to determine the probability of accepting this move, $A(y|x)$, which we cannot calculate directly. Therefore, we need some expression for the acceptance probability which satisfies the acceptance ratio above. A solution to this problem is the Metropolis-Hastings criterion:[162,163]

$$A(y|x) = \min\left[1,\ \frac{A(y|x)}{A(x|y)}\right] \tag{2.106}$$

This therefore allows us to compute the probability of accepting a move, without requiring the unknown probability of accepting the reverse move.

In order to ensure that a given move is accepted with the correct probability, the calculated acceptance probability is compared to a random number drawn from a uniform distribution between 0 and 1. If the acceptance probability is greater than the random number, the move is accepted, and the proposed microstate is added to the ensemble and used as the starting point for the next move. Otherwise, the move is rejected and an additional copy of the initial microstate is added to the ensemble. This process generates a Markov chain, in which the probability of generating any microstate in the chain is dependent only on the previous microstate, with no dependence on the history of the chain. In the ensemble generated, the microstates are represented according to their equilibrium probabilities, so the ensemble average of a property, A, can be calculated as a simple mean over all microstates:

$$\langle A \rangle \approx \frac{1}{M} \sum_{i=1}^{M} A_i \tag{2.107}$$

where M is the number of samples generated, and A_i is the value of the property for the i^{th} sample. The accuracy of this average increases as the number of samples increases, and is exact in the limit of infinite sampling.

As an example of how Monte Carlo simulation can be applied in practice, we consider a simple, canonical system of spherical particles. We can simulate this system by proposing a change in the particle coordinates from $\mathbf{r}^N$ to $\mathbf{r}^N_{new}$, which is then accepted or rejected. This can be done by selecting a particle at random, and then applying a randomly generated translation vector, $\delta\mathbf{r}$:

$$\delta\mathbf{r} = \Delta_{max} \begin{pmatrix} 2\zeta_1 - 1 \\ 2\zeta_2 - 1 \\ 2\zeta_3 - 1 \end{pmatrix} \tag{2.108}$$

where Δ_{max} is the maximum possible translation in any dimension (this is determined prior to simulation), and ζ_n are distinct random numbers drawn from a uniform distribution between 0 and 1. Here, the probability of proposing the reverse move (selecting the same particle at random, and then displacing it by $-\delta\mathbf{r}$) is equal to that of the forward move, and as such, the acceptance ratio depends only on the probability ratio of

the two configurations:

$$\begin{aligned}\frac{A(\mathbf{r}^N_{new}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^N_{new})} &= \frac{\pi_{NVT}(\mathbf{r}^N_{new})}{\pi_{NVT}(\mathbf{r}^N)} \\ &= \frac{Q^{-1}_{NVT}\Lambda^{-3N}(N!)^{-1}e^{-\beta U(\mathbf{r}^N_{new})}}{Q^{-1}_{NVT}\Lambda^{-3N}(N!)^{-1}e^{-\beta U(\mathbf{r}^N)}} \\ &= e^{-\beta\Delta U}\end{aligned} \tag{2.109}$$

where ΔU is the potential energy change associated with the particle displacement. We then determine the probability of accepting the move, by calculating $A(\mathbf{r}^N_{new}|\mathbf{r}^N)$ using Eqs. 2.106 and 2.109, and comparing this to a random number drawn from a uniform distribution between 0 and 1. It should be noted that the proportion of moves which are accepted (known as the acceptance rate) will be dependent on the value chosen for Δ_{max}. For larger values of this parameter, it becomes increasingly likely that the move proposals will generate steric clashes between particles, causing a larger percentage of moves to be rejected (especially for simulations of condensed phases). Conversely, if the value chosen for Δ_{max} is very small, the acceptance rate will be high, but a very large number of moves will be needed in order to observe significant changes in the system. In practice, the value of Δ_{max} should be chosen to optimise this risk-reward balance.

2.6.1 Pressure Control

Monte Carlo sampling can be used to maintain the pressure of a simulation by proposing random changes in the volume, and then accepting or rejecting these changes, as appropriate. These moves can be applied at regular intervals during a simulation, to allow the volumes to correctly sample the NPT ensemble. Given a microstate containing N particles with scaled coordinates, $\mathbf{s}^N$, within a volume of V, we can propose a change in volume of ΔV, generating a new volume of $V_{new} = V + \Delta V$. It should be noted that here it is assumed that ΔV is drawn from a uniform distribution over a suitable range, which is symmetric about zero — this ensures that the probabilities of proposing the forward and reverse moves are equal. The acceptance ratio for such a move can therefore be written as a ratio of the equilibrium probabilities for the two microstates:[117]

$$\begin{aligned}\frac{A(V_{new}|V)}{A(V|V_{new})} &= \frac{\pi_{NPT}(\mathbf{s}^N; V_{new})}{\pi_{NPT}(\mathbf{s}^N; V)} \\ &= \frac{Z^{-1}_{NPT}\beta P\Lambda^{-3N}(N!)^{-1}V^N_{new}e^{-\beta PV_{new}}e^{-\beta U(\mathbf{s}^N;V_{new})}\mathrm{d}\mathbf{s}^N\mathrm{d}V}{Z^{-1}_{NPT}\beta P\Lambda^{-3N}(N!)^{-1}V^N e^{-\beta PV}e^{-\beta U(\mathbf{s}^N;V)}\mathrm{d}\mathbf{s}^N\mathrm{d}V} \\ &= \left(\frac{V_{new}}{V}\right)^N e^{-\beta P(V_{new}-V)}e^{-\beta\Delta U} \\ &= \left(\frac{V_{new}}{V}\right)^N e^{-\beta P\Delta V}e^{-\beta\Delta U}\end{aligned} \tag{2.110}$$

where ΔU is the difference in potential energy caused by the proposed change in volume. It should be noted that a Monte Carlo barostat could also be implemented by proposing random changes in the simulation box lengths, or by applying random changes to the logarithm of the volume, but the acceptance ratio would have to be modified accordingly.[117]

2.6.2 Nonequilibrium Candidate Monte Carlo

As previously mentioned, Monte Carlo sampling is often very limited for the simulation of condensed phase systems. The high density of these systems makes it overwhelmingly likely that, for all but the smallest move proposals, a steric clash will be created, resulting in rejection. For this reason, only relatively small proposals are typically attempted, to ensure a significant overlap between the initial and proposed states, in order to give a reasonable chance of move acceptance. This approach is highly problematic for the sampling of transitions between microstates with high probabilities that are separated by potential energy barriers, as the intermediate microstates will likely never be sampled within a finite simulation time. Additionally, if a concerted motion is required, it becomes overwhelmingly unlikely that the sequence of proposals required are all randomly attempted (and accepted) in order. Nonequilibrium candidate Monte Carlo (NCMC) is a method which attempts to resolve this issue by increasing the acceptance probabilities of large transitions between high probability microstates.[115]

The core concept of this method is that a large move is proposed (such as rotation of a restricted dihedral, for example), and this move is broken into a series of smaller perturbations, from which the combined effect gives the large proposal.[115] The perturbation steps are separated by relaxation steps, which allows the system to relax in response to the perturbation. This resolves the issues typically presented by large move proposals in MC simulations, by allowing the environment to adapt to the proposal, thereby greatly reducing the likelihood that the proposal results in steric clashes. NCMC has been applied to improve the sampling of ligand binding modes,[164–166] rotation about restricted dihedrals,[167] fluctuations in salt concentration,[168] water binding[169] and changes in protonation states.[170,171] In some of these cases, the use of NCMC has been found to improve the acceptance rates of large moves by many orders of magnitude, relative to traditional Monte Carlo moves.[164,168]

Each NCMC move is separated into a series of perturbation steps (denoted a_n), in which work is done on the system to drive a change in a particular direction, and propagation steps (denoted K_n), where the system releases heat as it relaxes in response to the perturbation. The order in which these steps are applied is referred to as the move protocol, $\Lambda_p = \{a_1, K_1, \ldots, a_T, K_T\}$. When this protocol is applied to a state, x_0, (where x

represents all positions, momenta and any other simulation parameters which are subject to change), a sequence of microstates (or path) is generated, $X = \{x_0,\ x_1, \ldots,\ x_T\}$. The forward move can therefore be represented as:

$$x_0 \xrightarrow{a_1} x_1^* \xrightarrow{K_1} x_1 \rightarrow \cdots \rightarrow x_{T-1} \xrightarrow{a_T} x_T^* \xrightarrow{K_T} x_T \tag{2.111}$$

where x_n^* is generated by perturbing x_{n-1}. The detailed balance condition requires that there must be a non-zero probability of selecting the reverse protocol, $\tilde{\Lambda}_p$, in which the sequence of perturbation and propagation steps is the reverse of Λ_p, such that when applied to $\tilde{x}_T$ (where the tilde indicates that the momenta have been reversed) the reverse sequence of microstates, $\tilde{X}$, is generated, returning the system to x_0. In order to preserve the equilibrium distribution, the momenta must be reversed upon either acceptance or rejection.[172,173]

The probabilities of each step in the protocol must all be accounted for, arriving at the following, highly generalised, acceptance ratio:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = \frac{P(\tilde{\Lambda}_p|\tilde{x}_T)}{P(\Lambda_p|x_0)} \frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)} \frac{\pi(\tilde{x}_T)}{\pi(x_0)} e^{-\Delta\mathcal{S}(X|\Lambda_p)} \tag{2.112}$$

where $P(\Lambda_p|x_0)$ is the probability of selecting protocol Λ_p from x_0, $\alpha(X|\Lambda_p)$ is the cumulative probability of each perturbation step from the forward move, and $\Delta\mathcal{S}(X|\Lambda_p)$ is the conditional path action difference (where the conditional path action is the negative natural logarithm of the conditional path probability[174]). The latter two terms are related to the probabilities of the individual perturbation and propagation steps as follows:

$$\frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)} = \prod_{t=1}^{T} \frac{a_t(\tilde{x}_t^*,\ \tilde{x}_{t-1})}{a_t(x_{t-1},\ x_t^*)} \tag{2.113}$$

$$e^{-\Delta\mathcal{S}(X|\Lambda_p)} = \prod_{t=1}^{T} \frac{K_t(\tilde{x}_t,\ \tilde{x}_t^*)}{K_t(x_t^*,\ x_t)} \tag{2.114}$$

where $a_t(x_{t-1},\ x_t^*)$ is the probability of generating x_t^* by applying perturbation a_t to x_{t-1}, and, similarly, $K_t(x_t^*,\ x_t)$ is the probability of generating x_t by applying propagation K_t to x_{t-1}. It should be noted that the acceptance ratio above, derived by Nilmeier *et al.*, is based on a strict, pathwise form of detailed balance, which reduces to the traditional detailed balance condition if all possible trajectories/paths between x_0 and $\tilde{x}_T$ are accounted for.[115]

It should be noted that Eq. 2.112 is a highly generalised form of the acceptance ratio, and, when applied in practice, is typically much less complex, owing to a number

of simplifications which can typically be made, depending on the specific implementation. For example, if the forward and reverse protocols are selected with equal probability, then $P(\Lambda_p|x_0) = P(\tilde{\Lambda}_p|\tilde{x}_T)$, and these terms thus cancel. If the perturbation kernels are applied in a deterministic fashion, it typically transpires that $\alpha(X|\Lambda_p) = \alpha(\tilde{X}|\tilde{\Lambda}_p)$, though this is not necessarily the case if there is a stochastic element to the perturbations.

The method used for propagating the system during the relaxation phases can also have a significant impact on the acceptance criterion. If the propagation steps are deterministic (as they would be if using velocity Verlet integration, for example), then $K_t(x_t^*,\ x_t) = K_t(\tilde{x}_t,\ \tilde{x}_t^*)$, and the conditional path action difference evaluates to zero, so the exponential of this term can be safely omitted. However, if the propagation steps are carried out using an equilibrium-preserving method (such as Monte Carlo sampling, for example), then the conditional path action difference is related to the heat, $Q(X|\Lambda_p)$, released over the course of the forward move, as $\Delta\mathcal{S}(X|\Lambda_p) = -\beta Q(X|\Lambda_p)$. The heat and work associated with the forward move are calculated as:[115,175]

$$W(X|\Lambda_p) = \sum_{t=1}^{T} [E(x_t^*) - E(x_{t-1})] \tag{2.115}$$

$$Q(X|\Lambda_p) = \sum_{t=1}^{T} [E(x_t) - E(x_t^*)] \tag{2.116}$$

When these various simplifications are combined, the acceptance ratio is typically much more compact than Eq. 2.112. For the full derivation of these terms and further details on the underlying theory of NCMC, the 2011 publication by Nilmeier *et al.* is highly recommended.[115]

2.6.3 Grand Canonical Monte Carlo

Here, the theory underpinning grand canonical Monte Carlo (GCMC) is described, along with some of the free energy analyses that can be performed using GCMC simulations.[104,105]

2.6.3.1 Acceptance Criteria

Monte Carlo simulations are the primary method of simulating the grand canonical ensemble,[21,99–105,107–109] as they provide a convenient way to allow the number of particles in a simulation to vary according to an imposed chemical potential. This is implemented via the inclusion of Monte Carlo moves which insert and delete particles to/from the system, in a theoretically rigorous manner. This section shows how the

acceptance criteria for these moves can be derived.[117]

We begin with the derivation of a particle insertion move, in which we increase the number of particles in the system from N to $N+1$, by translating a particle from the ideal gas reservoir. We again make use of the treatment of the combined system and ideal gas reservoir as a large canonical ensemble, as described in section 2.2.3. The particles in the system and the reservoir are identical, with the only difference being that particles exhibit intermolecular interactions when in the system, but not when in the reservoir — when an ideal particle is inserted into the system, it becomes 'real' (i.e. exhibits interactions with other particles), and conversely, when a 'real' particle is moved to the ideal gas reservoir, it becomes ideal. Here, the equilibrium probability of a microstate containing N particles in the system and $M-N$ particles in the ideal gas, with position vectors, $\mathbf{r}^N$ and $\mathbf{r}^{M-N}$, respectively, is given by:

$$\pi(\mathbf{r}^N, \mathbf{r}^{M-N}) = Q_{MVT}^{-1} \Lambda^{-3N} \Lambda^{-3(M-N)} e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N \mathrm{d}\mathbf{r}^{M-N} \tag{2.117}$$

At any point in time, the particles in the ideal gas are indistinguishable from each other (as are those in the system), but the $M-N$ identical ideal particles can be distinguished from the N identical particles in the system. In the above, the factorial terms have been dropped, as this gives the probability of observing any arrangement of identical particles at those positions, and is therefore independent of arbitrary particle labelling schemes — note that this step is not necessary for the derivation, but makes the rest of this section more intuitive. The forward move involves translating one of the $M-N$ ideal gas particles to a random point (with infinitesimal volume, $\mathrm{d}\mathbf{r}$) in the simulated system, with the following probability of proposal:

$$P(\mathbf{r}^{N+1}|\mathbf{r}^N) = \frac{1}{2} \frac{1}{M-N} \frac{\mathrm{d}\mathbf{r}}{V_{sys}} \tag{2.118}$$

where the first term is because insertion and deletion moves are equally likely, the second arises from selecting one particle at random from $M-N$ particles, and the third is owing to the probability of selecting a particular position being inversely proportional to the accessible volume. Conversely, the reverse move involves selecting one of the $N+1$ particles from the proposed microstate, and translating it to a random point in the ideal gas, with the following probability of proposing the initial microstate:

$$P(\mathbf{r}^N|\mathbf{r}^{N+1}) = \frac{1}{2} \frac{1}{N+1} \frac{\mathrm{d}\mathbf{r}}{V_{ideal}} \tag{2.119}$$

These terms can be combined with the equilibrium probabilities of the initial and proposed microstates to calculate the acceptance ratio:

$$\begin{aligned}\frac{A(\mathbf{r}^{N+1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} &= \frac{P(\mathbf{r}^N|\mathbf{r}^{N+1})}{P(\mathbf{r}^{N+1}|\mathbf{r}^N)}\frac{\pi(\mathbf{r}^{N+1},\mathbf{r}^{M-N-1})}{\pi(\mathbf{r}^N,\mathbf{r}^{M-N})} \\ &= \frac{\mathrm{d}\mathbf{r}V_{ideal}^{-1}(N+1)^{-1}}{\mathrm{d}\mathbf{r}V_{sys}^{-1}(M-N)^{-1}}\frac{Q_{MVT}^{-1}\Lambda^{-3(N+1)}\Lambda^{-3(M-N-1)}e^{-\beta U(\mathbf{r}^{N+1})}\mathrm{d}\mathbf{r}^{N+1}\mathrm{d}\mathbf{r}^{M-N-1}}{Q_{MVT}^{-1}\Lambda^{-3N}\Lambda^{-3(M-N)}e^{-\beta U(\mathbf{r}^N)}\mathrm{d}\mathbf{r}^N\mathrm{d}\mathbf{r}^{M-N}} \\ &= \frac{M-N}{V_{ideal}}\frac{V_{sys}}{N+1}e^{-\beta\Delta U}\end{aligned} \tag{2.120}$$

where ΔU is the potential energy change associated with the particle insertion. As M goes to infinity, the number density of the ideal gas can be substituted as:

$$\lim_{M\to\infty}\frac{M-N}{V_{ideal}} = \rho_{ideal} \tag{2.121}$$

The number density is related to the chemical potential of the ideal gas as shown in Eq. 2.64, which allows the acceptance ratio to be re-written in terms of the chemical potential:

$$\frac{A(\mathbf{r}^{N+1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} = \frac{1}{N+1}\frac{V_{sys}}{\Lambda^3}e^{\beta\mu}e^{-\beta\Delta U} \tag{2.122}$$

We can then introduce the Adams parameter:[99,100]

$$B = \beta\mu + \ln\left(\frac{V_{sys}}{\Lambda^3}\right) \tag{2.123}$$

which allows the acceptance ratio to be written more simply, as:

$$\frac{A(\mathbf{r}^{N+1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} = \frac{1}{N+1}e^{B}e^{-\beta\Delta U} \tag{2.124}$$

Similarly, the acceptance ratio for a deletion move, which decreases the number of particles in the system from N to $N-1$, can be derived to be the following (the derivation is not given here, to avoid repetition of many of the steps above):

$$\frac{A(\mathbf{r}^{N-1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N-1})} = Ne^{-B}e^{-\beta\Delta U} \tag{2.125}$$

where ΔU here is the potential energy associated with the deletion of a particle. Note that in both cases, the acceptance ratio has no explicit dependence on the ideal gas reservoir, meaning that the ideal gas need not be simulated.[117] In order to allow this, any non-translational degrees of freedom must be accounted for. For example, water

molecules in the ideal gas will show uniformly distributed orientations, and when inserting a water molecule in a GCMC simulation, a random orientation must therefore be generated, in order to replicate the orientation that the water would have in the ideal gas. It should also be noted that the derivation presented in this section can be carried out differently, whilst arriving at the same result.[117] The derivation presented here is that thought to be most intuitive.

2.6.3.2 Calculation of Water Network Binding Free Energies

The chemical potential at which water sites are observed in a GCMC simulation is related to their stability.[104] For example, a very tightly bound water molecule would require a significant reduction in the chemical potential of the simulation in order to make the deletion favourable. This section describes how a set of simulations at a range of different chemical potentials can be used to perform a free energy analysis on the waters present in a particular region of interest — these calculations are known as titration calculations. An advantage of GCMC in this regard, is that cooperative effects between waters are captured implicitly.

The calculation of the free energy change associated with transferring waters from the ideal gas to the system was first presented by Ross *et al.*, and can be separated into the following:[104]

$$\Delta F_{trans} = \Delta F_{sys} - \Delta F_{ideal} \tag{2.126}$$

where ΔF_{sys} is the Helmholtz free energy difference associated with changing the number of waters in the system, and ΔF_{ideal} is the difference of changing the number of waters in the ideal gas. In the thermodynamic limit, the Helmholtz free energy is related to the grand potential via:[143]

$$F_{NVT} = \Omega_{\mu VT} + N\mu \tag{2.127}$$

Which allows the free energy change of the system to be written in terms of the grand potential and the chemical potential:

$$\Delta F_{sys}(N_i \rightarrow N_f) = \Delta\Omega_{sys}(\mu_i \rightarrow \mu_f) + N_f\mu_f - N_i\mu_i \tag{2.128}$$

where μ_f is the chemical potential at which N_f waters are observed, on average. The partial derivative of the grand potential with respect to the chemical potential is the negative of the particle number (Eq. 2.41), meaning that the change in the grand potential can therefore be calculated by the following integral:

$$\Delta\Omega_{sys}(\mu_i \rightarrow \mu_f) = -\int_{\mu_i}^{\mu_f} N(\mu)\mathrm{d}\mu \tag{2.129}$$

Substituting this into the above gives an expression for the change in free energy of the system:

$$\begin{aligned}\Delta F_{sys}(N_i \to N_f) &= N_f\mu_f - N_i\mu_i - \int_{\mu_i}^{\mu_f} N(\mu)\mathrm{d}\mu \\ &= k_BT\left\{N_fB_f - N_iB_i - (N_f - N_i)\ln\left(\frac{V_{sys}}{\Lambda^3}\right) - \int_{B_i}^{B_f} N(B)\mathrm{d}B\right\}\end{aligned} \tag{2.130}$$

The free energy of an ideal gas containing N particles is given in Eq. 2.61, from which the difference in ideal gas free energy between N_f and N_i waters can be written as:

$$\Delta F_{ideal}(N_i \to N_f) = k_BT\ln\left(\frac{N_f!}{N_i!}\right) - k_BT(N_f - N_i)\ln\left(\frac{V_{ideal}}{\Lambda^3}\right) \tag{2.131}$$

This can be combined with the change in system free energy to give the transfer free energy:

$$\beta\Delta F_{trans}(N_i \to N_f) = N_fB_f - N_iB_i + \ln\left(\frac{N_i!}{N_f!}\right) - (N_f - N_i)\ln\left(\frac{V_{sys}}{V_{ideal}}\right) - \int_{B_i}^{B_f} N(B)\mathrm{d}B \tag{2.132}$$

Note that this is not exactly the same expression as given for $\beta\Delta F_{trans}$ by Ross *et al.*,[104] as here we have not assumed the volumes of the system and ideal gas to be equal.

However, we are more interested in the difference in binding free energy, ΔG_{bind}, between two water networks containing N_f and N_i water molecules, respectively. Using the thermodynamic cycle shown in Fig. 2.1, this quantity can be calculated from the transfer free energy as:[105]

$$\Delta G_{bind} = \Delta F_{trans} + \Delta F_{ideal} - \Delta G_{sol} \tag{2.133}$$

where ΔG_{sol} is the free energy change associated with changing the number of waters in bulk solution — note that the differences between Helmholtz and Gibbs free energies are assumed to be negligible. The free energy change in the bulk solvent is given by:

$$\begin{aligned}\Delta G_{sol}(N_i \to N_f) &= (N_f - N_i)\mu_{sol} \\ &= (N_f - N_i)\left(\mu_{sol}^{ex} + k_BT\ln(\rho_{sol}\Lambda^3)\right)\end{aligned} \tag{2.134}$$

where μ_{sol} and ρ_{sol} are the chemical potential and number density of bulk water, respectively. Eqs. 2.131, 2.132 and 2.134 can therefore be substituted into Eq. 2.133:

$$\beta\Delta G_{bind}(N_i \to N_f) = N_fB_f - N_iB_i - (N_f - N_i)\left\{\beta\mu_{sol}^{ex} + \ln\left(\rho_{sol}V_{sys}\right)\right\} - \int_{B_i}^{B_f} N(B)\mathrm{d}B \tag{2.135}$$

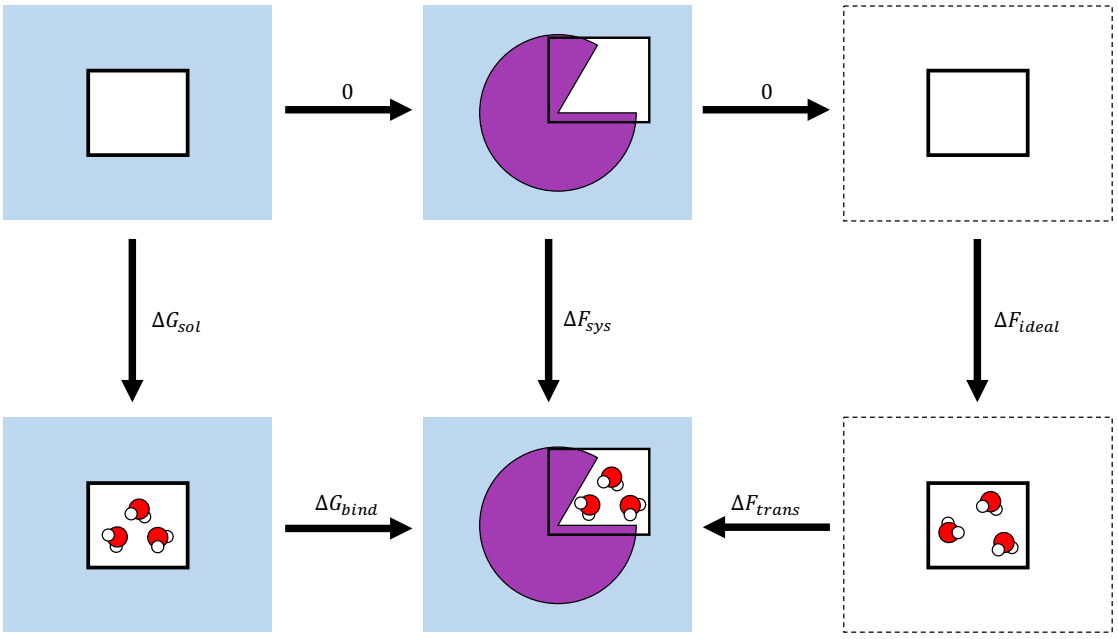


FIGURE 2.1: Thermodynamic cycle for the calculation of the binding free energy of waters to a GCMC region. The columns represent different environments, with bulk water on the left, the simulated system (a protein binding site, for example) in the centre, and an ideal gas on the right. The top row represents fully dehydrated GCMC regions, and the bottom row represents hydrated GCMC regions. Image based on Fig. 1.5 in the thesis of Dr Hannah Bruce Macdonald.[176]

Under standard state conditions, we take $\rho^{\circ}_{sol} = 1/V^{\circ}$, where V° is the standard state volume of water. The above can therefore be re-written as a difference in standard state binding free energies:[105]

$$\beta\Delta G^{\circ}_{bind}(N_i \rightarrow N_f) = N_f B_f - N_i B_i - (N_f - N_i)\left\{\beta\mu^{ex}_{sol} + \ln\left(\frac{V_{sys}}{V^{\circ}}\right)\right\} - \int_{B_i}^{B_f} N(B)\mathrm{d}B \tag{2.136}$$

Free energies calculated using Eq. 2.136 have been found to be in excellent agreement with those calculated using double decoupling, at reduced computational cost (the increase in efficiency is greater for larger water networks).[105] Whilst multiple references have been made to the thermodynamic limit in this derivation, it was verified by Ross *et al.* that these relationships are also valid for very small values of N.[104,105] Further details on the derivations presented in this section can be found in the relevant publications by Ross *et al.*[104,105]

It should be noted that multiple references have been made above to the integral of the number of waters observed as a function of the Adams value. Given that the data obtained from a set of GCMC titration data is not a simple function of B, we must make some approximations to calculate this integral from numerical data. In the limit

of infinite sampling, $N(B)$ increases monotonically with B, and it transpires that this data is well represented by a sum of sigmoid functions:[104]

$$N(B) \approx \sum_{i=1}^{m} \frac{n_i}{1 + \exp(w_{0,i} - w_i B)} \tag{2.137}$$

where m is the number of sigmoids, and n_i, $w_{0,i}$ and w_i are the parameters of the i^{th} sigmoid. The value of m is chosen with some degree of user judgement, but a convenient rule of thumb is to set m to the number of 'steps' observed in a plot of $N(B)$ against B. A benefit of fitting this set of sigmoids, is that the integral of the fitted curve can then be calculated analytically:[104]

$$\int_{B_i}^{B_f} N(B)\mathrm{d}B \approx \sum_{i=1}^{m} \frac{n_i}{w_i} \ln\left(\frac{e^{w_i B_f} + e^{w_{0,i}}}{e^{w_i B_i} + e^{w_{0,i}}}\right) \tag{2.138}$$

This therefore allows the integral to be calculated with relative ease.

2.6.3.3 GCMC Simulations at Equilibrium

The number of waters observed in a particular region, when in equilibrium with bulk water, could be determined by running a full titration calculation, and then determining the value of N which produces the lowest value of $\Delta G^{\circ}_{bind}(N)$. However, this would require simulations at a large number of chemical potentials, over a range which is not known *a priori*. For this reason, it is of great use to be able to determine the Adams value at which equilibrium behaviour will be observed, B_{equil}.

As mentioned above, the point of equilibrium is defined by a minimum in $\Delta G^{\circ}_{bind}(N)$, with respect to N, where the derivative will be zero:[105]

$$\frac{\partial \Delta G^{\circ}_{bind}(N)}{\partial N} = 0 \tag{2.139}$$

Here, it is convenient to write the standard binding free energy as the following (see Fig. 2.1):

$$\Delta G^{\circ}_{bind}(N) = \Delta F_{sys}(N) - \Delta G^{\circ}_{sol}(N) \tag{2.140}$$

where the superscripts indicate that the free energies are referenced to the standard state. We can also write ΔF_{sys} as:

$$\Delta F_{sys}(N) = \int_0^N \mu_{sys}(N)\ \mathrm{d}N \tag{2.141}$$

where the above holds in the thermodynamic limit.[104,105] This can then be substituted back into the expression for the standard binding free energy:

$$\Delta G^{\circ}_{bind}(N) = \int_0^N \mu_{sys}(N)\, \mathrm{d}N - N\mu^{\circ}_{sol} \tag{2.142}$$

where μ°_{sol} is the standard state chemical potential of bulk water. The above can now be easily differentiated with respect to the number of waters:

$$\frac{\partial \Delta G^{\circ}_{bind}(N)}{\partial N} = \mu_{sys} - \mu^{\circ}_{sol} \tag{2.143}$$

which shows that at the free energy minimum, the chemical potential of the system must be equal to the standard state chemical potential of bulk water, thereby establishing this as a condition for equilibrium.[105] Therefore, the value of the Adams parameter (Eq. 2.123) at equilibrium can be determined from the standard state chemical potential of bulk water:

$$\begin{aligned} B_{equil} &= \beta\mu^{\circ}_{sol} + \ln\left(\frac{V_{sys}}{\Lambda^3}\right) \\ &= \beta\left(\mu^{ex}_{sol} + k_B T \ln\left(\frac{\Lambda^3}{V^{\circ}}\right)\right) + \ln\left(\frac{V_{sys}}{\Lambda^3}\right) \\ &= \beta\mu^{ex}_{sol} + \ln\left(\frac{V_{sys}}{V^{\circ}}\right) \end{aligned} \tag{2.144}$$

This value of the Adams parameter can then be used to run a single simulation in which the GCMC region will demonstrate equilibrium behaviour with bulk water.[105] The excess chemical potential and standard state volume of bulk water are therefore required parameters for an equilibrium GCMC simulation.

It is important to note that Eq. 2.141 is only exact in the thermodynamic limit, and it has not yet been theoretically demonstrated that the resulting expression for B_{equil} (Eq. 2.144) holds for small values of N. In practice, however, the average number of waters observed in a single simulation at B_{equil} is in agreement with the value of N which corresponds to a minimum in the water network binding free energy (see data in chapter 4), indicating that this result also applies away from the thermodynamic limit.

2.6.3.4 Accounting for the Non-Spherical Nature of Water

All discussion of partition functions and chemical potentials in this chapter have been with respect to a system of spherical particles, for simplicity. However, for systems of non-spherical molecules, terms must be included to account for the additional, non-translational degrees of freedom of the molecules. The ideal partition function for a

system containing N molecules can be written as:[177]

$$Q_{NVT}^{id} = \frac{(q^{id})^N}{N!} \tag{2.145}$$

where q^{id} is the ideal partition function for a single molecule under identical conditions, and can be written as a product of the following terms:[177]

$$q^{id} = q^{trans} q^{rot} q^{vib} \tag{2.146}$$

where these are the partition functions associated with integrating all translational, rotational, and vibrational degrees of freedom, respectively, for an ideal molecule. The translational component of the molecular partition function has been presented already:

$$q^{trans} = \frac{V}{\Lambda^3} \tag{2.147}$$

The vibrational partition function can be safely ignored here, as in the vast majority of modern simulations, water molecules are constrained, and therefore have no vibrational degrees of freedom, giving $q^{vib} = 1$. However, the rotational partition function, q^{rot} cannot be neglected for a molecule such as water. This therefore affects the chemical potential of water:[178]

$$\begin{aligned} \mu^{id} &= \frac{\partial F^{id}}{\partial N} \\ &= -k_B T \frac{\partial}{\partial N} \ln\left(\frac{(q^{id})^N}{N!}\right) \\ &= -k_B T \ln\left(\frac{q^{id}}{N}\right) \\ &= -k_B T \ln\left(\frac{V q^{rot}}{N\Lambda^3}\right) \\ &= k_B T \ln\left(\frac{\rho_{ideal}\Lambda^3}{q^{rot}}\right) \end{aligned} \tag{2.148}$$

For molecules more complex than water (or a flexible model of water), the q^{vib} term would be included alongside q^{rot}. Note that the effects of intermolecular interactions on each of these degrees of freedom are absorbed into the excess chemical potential, which must be calculated numerically (section 2.3.2).

Whilst the rotational partition function should strictly be accounted for in GCMC, it transpires that this term can be absorbed into the Adams value, and when determining the value of B_{equil}, q^{rot} is cancelled exactly. Additionally, this term also cancels in the derivation of the water network binding free energies. Therefore, when GCMC is carried out via the Adams value (as is standard practice for many) the neglect of the rotational partition function has no impact. The proof for the cancellation of q^{rot} is given

in appendix A.

2.7 Summary

This chapter has provided a theoretical foundation for the methods which were used in the work presented in the rest of this thesis. The focus of this thesis is simulation of the grand canonical ensemble, and as such, the theory underlying both this ensemble and how it can be simulated and analysed in a Monte Carlo context, has been discussed in detail. Chapter 3 demonstrates how GCMC can be combined with MD sampling, yielding simulations referred to as GCMC/MD. In chapter 4, GCMC/MD titrations are carried out, in order to determine binding free energies of water networks to a protein of interest, using Eq. 2.136. The theory relating to NCMC is referenced extensively in chapter 5, where this technique is used to enhance the acceptance rates of GCMC moves by allowing water insertions and deletions to be applied in an adaptive fashion — this technique is referred to as grand canonical nonequilibrium candidate Monte Carlo (GCNCMC). This proves very useful in chapter 6, where GCNCMC allows the insertion and deletion of benzene molecules, which is extremely difficult using conventional GCMC moves.

Chapter 3

Development of the *grand* Python Module

3.1 Introduction

Molecular dynamics simulations can be carried out in the isothermal-isobaric (NPT) ensemble by periodically updating the volume of the simulation according to the imposed pressure, via the use of a barostat. The volume is fixed between these updates, so the simulation trajectory is strictly a concatenation of many short canonical trajectories with different volumes. If the time between volume updates is short, relative to the total length of the simulation, the simulated equilibrium distribution should be indistinguishable from that of the NPT ensemble. A similar concept can be used to carry out MD simulations in the grand canonical ensemble, where, instead of updating the volume according to the pressure, the number of particles is updated according to the chemical potential. Grand canonical Monte Carlo (GCMC) moves (as described in detail in section 2.6.3) are a convenient way to allow the particle number to vary according to the equilibrium distribution.[143] Therefore, simulations can sample the grand canonical ensemble by interspersing canonical molecular dynamics with GCMC particle insertion and deletion moves — here, such simulations are referred to as GCMC/MD. However, it should be noted that this work is not the first application to combine molecular dynamics with GCMC sampling in this way.[107]

As previously mentioned, during this work, the aims were not only to apply GCMC/MD simulations to systems of interest, but also to expand upon the existing theory and methodology, in order to extend the applicability of this method. Therefore, it was determined that this work would be best served by the development of a new software package which could be used as a foundation for any such advances. For this reason, a Python module, named *grand*,[113] was developed, which serves as a 'bolt-on' tool

to carry out GCMC sampling of water molecules during molecular simulations using the OpenMM simulation engine.[114,179] OpenMM was chosen for this purpose as it has a growing user base and, more importantly, the highly versatile Python applied programming interface (API) makes it very well suited to the development of prototype code for the incorporation of novel ideas into molecular simulations.[114,179]

In this chapter, the implementation of GCMC sampling in the *grand* module is described, along with the results of a series of tests performed to validate this implementation. A number of tests were carried out using simulations of bulk water, as this serves as a simple, homogeneous test case. It should be noted, however, that the bulk water tests performed are not trivial, as they are much larger than the protein binding sites for which the GCMC implementation in *grand* is intended, and therefore require significantly more sampling in order to observe sufficient fluctuations in the particle number. This increased size means that any theoretical or methodological errors in *grand* should be exacerbated by the size of these systems. Finally, *grand* was applied to a protein system, in order to verify that stable waters within a binding site can be correctly inserted.

3.2 Implementation

The implementation of GCMC sampling of water molecules in *grand* is largely based on that of the ProtoMS software package.[180] However, practical considerations have dictated a number of changes in implementation, which are described in this section. These differences primarily arise from those between MC and MD sampling of protein systems.

In ProtoMS, the GCMC region sampled is defined as a cuboid centred on a particular point in Cartesian space.[104,180] This is feasible because (as previously stated) large-scale translations and rotations of the protein are highly unlikely, and as such, the position and orientation of the GCMC region, relative to the protein, are practically constant. However, in MD simulations, large-scale protein motions are a common occurrence, so the definition of the GCMC region must be adapted accordingly. Therefore, in *grand*, the GCMC region is defined as a sphere, centred on the mean coordinate of a subset of protein atoms (chosen by the user). The use of a sphere ensures that the region of the protein covered is independent of protein rotation, and the use of reference atoms attaches the region to the protein, rather than to a fixed point in space.

In the ProtoMS implementation of GCMC, water molecules can only enter or leave the GCMC region via insertion/deletion moves, as translation is blocked by so-called 'hard

wall' constraints — this involves applying a large, sudden energy change as soon as a water crosses the boundary, resulting in the rejection of such moves. These constraints are not easily implemented in MD simulations, as discontinuous energy surfaces are problematic for the calculation of forces. An alternative approach is the use of 'soft wall' restraints, in which a half-harmonic function could be used to repel waters which attempt to move in or out of the GCMC region via translation. However, Newton's third law dictates that for every force applied, there must also exist an equal and opposite force. This opposite force, if resolved onto the protein, could result in artificial dynamics which would distort the simulation. Therefore, it was deemed least objectionable to not prevent waters from translating across the boundaries of the GCMC region in *grand*. This is not expected to present a problem for simulations at B_{equil}, but it could be problematic for GCMC titrations. For example, at very low B values, where deletion moves are favoured, the deleted waters may be continually replaced by those diffusing from bulk water, preventing the GCMC region from being dehydrated.

It should also be noted that, by default, long-range corrections to the Lennard-Jones interactions are disabled in *grand*. This is because the Lennard-Jones interactions require softcore potentials to prevent singularities from occurring when 'ghost' waters (those which are non-interacting, owing to their having been deleted or not yet inserted) overlap with other particles. This is implemented via a *CustomNonbondedForce* object in OpenMM, which are not able to utilise the analytical calculation of the long-range correction, and must therefore calculate this numerically. Whilst this numerical solution has been written very efficiently to prevent unnecessary repeated calculations, these efficiency measures are upset whenever the simulation parameters are changed (such as the attempted insertion or deletion of a water molecule), which triggers a recalculation of the entire solution. Therefore, the use of long-range Lennard-Jones corrections becomes very expensive with GCMC/MD as implemented in *grand*, and so the decision was made to neglect this contribution.

Another key difference between *grand* and ProtoMS is that, unlike MC simulations, MD requires that all particles in the system have an associated momentum vector. After a batch of GCMC moves during a GCMC/MD simulation, the system momenta are unchanged, except for any water molecules introduced by GCMC insertions, whose momenta should be random (as they would have been in the ideal gas, prior to insertion). In the latest development version of *grand* (which has not yet been publicly released, as of the time of writing), random velocities (drawn from the Maxwell-Boltzmann distribution, using the native OpenMM functionality) are explicitly assigned to waters upon insertion. In earlier versions of *grand*, inserted waters would retain the velocities which they possessed prior to insertion — as the waters would have been 'ghosts' prior to insertion, the velocities sampled would be effectively random, owing to the lack of

intermolecular interactions. In any case, it is expected that the velocities of inserted particles would have minimal effect on the simulated behaviour, as they would likely be quickly decorrelated from these values by the effect of interactions with their environment, and the stochastic updates made by the Langevin integrator (section 2.5.2).

3.2.1 Software Details

One of the primary benefits of the OpenMM simulation engine is the control over the simulation afforded to the user by the Python API. It was intended that *grand* would serve as a Python module which could be imported into a Python script describing an OpenMM simulation with minimal disruption, in order to preserve as much user control as possible. It was also decided that the amount of additional knowledge required to use *grand* should be minimal for a user who is already familiar with OpenMM. For this reason, the vast majority of the underlying GCMC operations (such as generating microstates and determining whether or not they should be accepted) and tracking of appropriate variables are handled 'under the hood' by a set of *Sampler* Python objects, with which the user can interact at a relatively high level, without requiring any specialist knowledge or mathematical understanding of GCMC.

A *Sampler* object must first be initialised with the information needed to carry out GCMC moves. This includes GCMC-specific data, such as a definition of the GCMC region (a sphere can be defined via a radius and a set of reference atoms, or the entire simulation volume can be used), as well as the values of the excess chemical potential and standard state volume of water, from which the value of B_{equil} is determined — alternatively, the user can override the calculation of B_{equil} with a chosen value for the Adams parameter. A number of more general parameters, including the temperature and system-specific information (such as the topology) must also be supplied. Once the *Sampler* object has been created, the particle coordinates must be supplied, along with the identities of non-interacting ('ghost') waters (*grand* includes a function to add ghost waters to a system), so that their interactions are not counted. From this point, GCMC sampling can be easily carried out at any point in the simulation, by running the *Sampler.move()* function, which will execute a prescribed number of GCMC moves on the system, from which molecular dynamics (or other OpenMM functionalities) can be applied to the new system configuration. This allows the user to still use OpenMM as they normally would, but with the additional ability to carry out GCMC sampling on their system, with the flexibility of applying any amount of GCMC in combination with any amount of molecular dynamics. Additionally, *grand* also includes a number of functions to support analysis of GCMC/MD simulations, including simple trajectory processing, and clustering of water sites. In order to facilitate usage of *grand*, a number of example scripts demonstrating different ways in which the module can be used are

provided on the GitHub page (see below).

The *grand* module[113] is released under the MIT licence and is freely available for download (along with example scripts) at https://github.com/essex-lab/grand. Alternatively, the module can be downloaded and installed using conda via the following command:

```
conda install -c omnia -c anaconda -c conda-forge -c essexlab grand
```

3.3 Simulation Details

The following simulation conditions were used for all simulations presented in this chapter. The AMBER ff14SB and TIP3P force fields were used to model the protein[181] and water,[182] respectively, with the Joung-Cheatham parameters used for neutralising ions.[183,184] Nonbonded interactions were subjected to a cutoff of 12 Å, with a switching function applied between 10 and 12 Å, and PME was used to calculate the effect of long-range electrostatic interactions.[142] All simulations were carried out at 298 K, using the BAOAB Langevin integrator ($\gamma = 1$ ps^{-1}, $\delta t = 2$ fs) to integrate the dynamics and control the temperature.[158] All bonds involving hydrogen were constrained to their equilibrium values, using the SHAKE algorithm for the protein,[185,186] and the SETTLE algorithm for water.[187] All simulations were performed at constant volume, unless explicitly stated otherwise — where constant pressure simulations were performed, the pressure was set to 1 bar and regulated using a Monte Carlo barostat, with volume changes attempted every 25 timesteps. All GCMC/MD simulations were carried out at the appropriate B_{equil} value, calculated using $\mu_{sol}^{ex} = -6.09$ kcal mol^{-1} and $V^{\circ} = 30.345$ Å^3 (as determined in section 3.3.1), unless explicitly stated otherwise. All simulations used version 7.2.2 of OpenMM[114,179] and version 1.0.0 of *grand*.[113]

3.3.1 Thermodynamic Parameters

The accuracy of GCMC/MD simulations is dependent on the values of the excess chemical potential, μ_{sol}^{ex}, and the standard state volume of water, V°, as these parameters define the reference state with which the simulated system is in equilibrium — a change in these parameters would correspond to a change in reference state. The values of the excess chemical potential and standard state volume can be calculated as the hydration free energy of water and the volume per water molecule, respectively. Experimentally, these values have been reported as -6.324 kcal mol^{-1} and 30.003 Å^3,[188,189] where the latter is calculated from the density of bulk water. In order to ensure self-consistency in

the event of discrepancies between simulated and experimental behaviour, these values were also determined computationally.

The hydration free energy of water was calculated by alchemically decoupling a water molecule from bulk, over 30 equally-spaced λ values from 1 to 0. For each independent repeat, 1000 potential energy samples were collected for each value of λ, with 10 ps of constant pressure MD carried out between samples — resulting in a total of 300 ns of simulation time per repeat. The data were processed to remove correlated samples[190] and the free energy was then calculated via the MBAR method,[149] using the functions provided in the *pymbar* Python module.[191] The calculation of the standard state volume of water is significantly simpler. Each independent repeat was simulated for 50 ns of constant pressure MD, with the average volume per water molecule sampled every 5 ps.

For both sets of simulations, a box of pre-equilibrated water, containing 2094 water molecules at a density of 0.978 g mL^{-1}, was used as the starting structure. In each case, the number of independent repeats carried out was increased until the standard error in the mean was judged to be sufficiently small.

3.3.2 Bulk Water Density

In order to assess the accuracy of the GCMC implementation, simulations were carried out to calculate the density of bulk water, where fluctuations in the density arise from variation in the number of waters in the system, at fixed volume. These simulations were compared to results from constant pressure simulations, in which fluctuations in the density are caused by variations in the volume, for a fixed number of particles. The NPT simulations were carried out for 100 ns, with the density sampled every 0.5 ns. The GCMC/MD simulations were also carried out for 100 ns, with 125 GCMC moves attempted every 250 fs, and densities recorded every 0.5 ns — in these simulations, the entire system volume was subjected to GCMC sampling. GCMC/MD simulations were carried out using the experimental values of μ_{sol}^{ex} and V° ($B_{equil} = -3.039$), and also with those calculated in this chapter ($B_{equil} = -2.655$).

Additional, shorter simulations were carried out to further test the GCMC/MD implementation. In order to test the ability of GCMC/MD to equilibrate the density of a water box, simulations were run with initial densities of 0.804 and 1.199 g mL^{-1}. The sensitivity of the density to the value of the excess chemical potential was also tested by setting this parameter to -6.15, -6.20 and -6.25 kcal mol^{-1}. Each of these simulations were run for 25 ns, with densities reported every 0.1 ns, with an additional set of

equivalent NPT simulations run for comparison with these results.

In each case, three independent runs were carried out, each starting from a pre-equilibrated water box containing 2094 water molecules at a density of 1.004 g mL^{-1} — except the GCMC/MD simulations where the equilibration of the density was tested, in which the initial number of water molecules was changed to give the densities stated.

3.3.3 Ensemble Validation

An additional, more rigorous test can be carried out to verify that the grand canonical ensemble is sampled correctly, using the method reported by Shirts.[192] This method involves comparing the probability distributions of some observable sampled under two different sets of thermodynamic conditions. For the grand canonical ensemble, the probability distribution in which we are interested is that of the particle number, which, for a given value of μ, is distributed as follows:

$$\pi(N|\mu) = \Xi^{-1} \frac{\int e^{-\beta U(\mathbf{r}^N)} \mathrm{d}\mathbf{r}^N}{\Lambda^{3N} N!} e^{\beta \mu N} \tag{3.1}$$

where the above is obtained by integrating the probability density (Eq. 2.58) over all configurations containing N particles. For two ensembles which differ only in the value of the chemical potential, the dependence on the unknown configurational integrals can be removed by taking the ratio of the probability distributions:

$$\frac{\pi(N|\mu_2)}{\pi(N|\mu_1)} = \frac{\Xi_2^{-1}}{\Xi_1^{-1}} e^{\beta(\mu_2 - \mu_1)N} \tag{3.2}$$

Taking the natural logarithm of this ratio yields a linear relationship with respect to N:

$$\ln\left(\frac{\pi(N|\mu_2)}{\pi(N|\mu_1)}\right) = \ln\left(\frac{\Xi_2^{-1}}{\Xi_1^{-1}}\right) + \beta(\mu_2 - \mu_1)N \tag{3.3}$$

where the gradient of the line is equal to $\beta(\mu_2 - \mu_1)$. Therefore, when two histograms of N are plotted from two simulations at different chemical potentials, if the natural logarithm of the probability ratio is plotted for the values of N in the overlapping region, then the gradient should yield this result, if the ensemble is sampled correctly.[192] The chemical potential values should be chosen to be sufficiently close that their distributions overlap well, but also sufficiently different that the distributions are distinct. As the samples of N collected are integers, there is no error introduced to the analysis from the choice of histogram bin width, provided that the width is set to 1.

For this analysis, GCMC/MD simulations were carried out on a bulk water system containing 2094 water molecules in a volume of (40.004 Å)^3, under the same conditions as those described in section 3.3.2, except that data were saved every 100 ps, instead of every 500 ps. Simulations were performed at $B_1 = B_{equil} = -2.630$ and $B_2 = -2.672$ (both rounded to 3 decimal places), corresponding to a difference in chemical potential of 0.025 kcal mol^{-1}. For each of these B values, 10 independent repeats were carried out from the same starting structure.

3.3.4 Bovine Pancreatic Trypsin Inhibitor

As the ultimate aim of the *grand* module is to support the sampling of water molecules buried within protein structures, bovine pancreatic trypsin inhibitor (BPTI) was also simulated, based on a buried pocket containing three waters, which has previously been used to validate GCMC results.[104,105] The three water sites have been found to be very stable,[105] and should be well reproduced.

The starting structure for the protein was taken from the Protein Data Bank,[31,32] with structure ID 5PTI.[193] Side chain issues were resolved by selecting the more occupied conformation, and all hydrogen atoms were removed, then re-added using the *Modeller* tool in OpenMM[114] — the terminal residues were charged. The protein was then solvated in a water box extending at least 8 Å from the protein in each dimension, and chloride ions were added to ensure that the system was neutrally charged, using the tleap program in AmberTools.[194]

The GCMC sphere was centred on the mean coordinate of the C_α atoms of the Tyr10 and Asn43 residues and had a radius of 4.2 Å, corresponding to $B_{equil} = -7.959$. All waters present in the sphere were deleted prior to equilibration, which took place over three stages — GCMC/MD to equilibrate the water sites, NPT MD to equilibrate the system volume, then GCMC/MD to further equilibrate the waters at the new volume. The first stage involved 10k initial GCMC moves, followed by 1 ps of GCMC/MD (1000 GCMC moves every 1 fs) — a high ratio of GCMC moves to MD was used, as the priority in the early stages of the equilibration is to insert waters before the dry pocket collapses. The NPT stage lasted 500 ps, and was followed by 500 ps of GCMC/MD (200 moves every 1 ps). The equilibration was followed by a production run of 10 ns of GCMC/MD (50 moves every 1 ps). For comparison, an identical set of simulations was carried out, except with the GCMC moves removed, leaving the GCMC/MD stages replaced with NVT MD.

3.4 Results

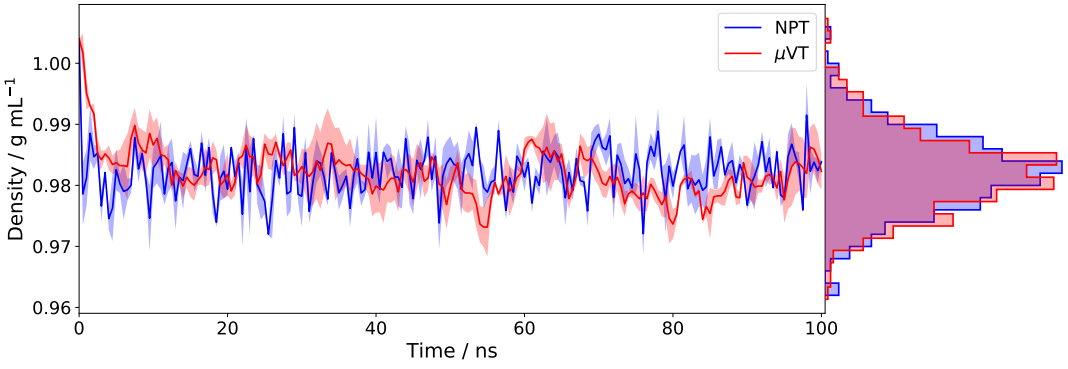
3.4.1 Thermodynamic Parameters

The hydration free energy calculation of water (in water) was carried out for a total of 50 independent runs, giving a mean value (with standard error) of -6.087 ± 0.005 kcal mol^{-1}. From this, the excess chemical potential was taken as $\mu_{sol}^{ex} = -6.09$ kcal mol^{-1}. Only 10 independent runs were necessary for the calculation of the standard state volume, giving a value of 30.3454 ± 0.0006 Å^3, from which the standard state volume is taken as $V^\circ = 30.345$ Å^3. These values are fairly close to the experimental values of -6.324 kcal mol^{-1} and 30.003 Å^3. The excess chemical potential is in good agreement with recently reported simulated values of -6.09 ± 0.04,[107] -6.18 ± 0.02,[195] and -6.05[196] kcal mol^{-1} for the excess chemical potential of TIP3P water (this is by no means intended to be an extensive list) — values reported in older publications vary more significantly.[197–199] Values for the standard state volume of TIP3P water of 29.855 Å^3 and 30.525 Å^3 have been reported (both values have been converted from the reported mass densities),[200,201] where the value calculated here is in better agreement with the latter.

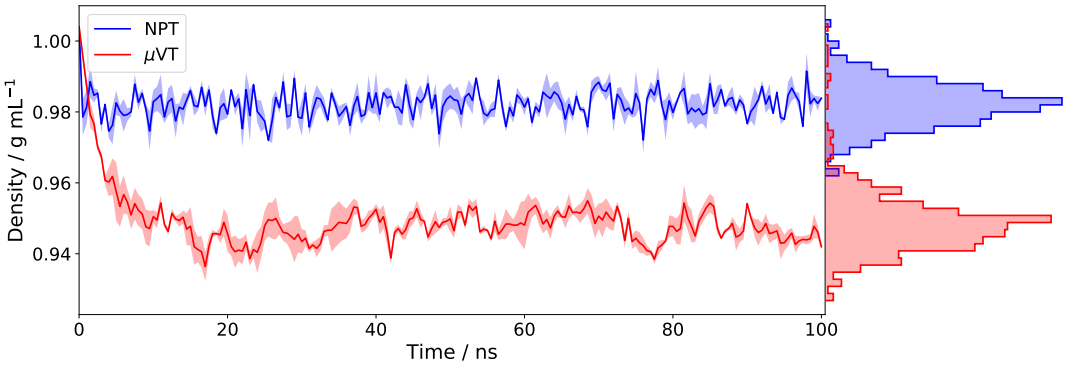
As previously mentioned, the values of the excess chemical potential and standard state volume of water can be interpreted as the definition of the reference state with which a simulated system is in equilibrium (in terms of water transfer). Here, that reference state has been taken as pure water, but it may be more appropriate for protein simulations to determine these parameters for more realistic, physiologically relevant mixtures. The use of more biologically relevant reference states may be of benefit to the simulation of waters in protein binding sites, particularly for those waters with binding affinities close to zero, where a slight shift in these parameters might make the difference between a water site appearing favourable or unfavourable. However, this more realistic approach brings additional complications. First, the composition of this solution (i.e. the relevant components and their concentrations) which is most appropriate will likely be dependent on the specific application, and also may be difficult to define. Secondly, the concentration of each component would be expected to fluctuate significantly about their macroscopic values (owing to the relatively small volume of simulated systems). This second issue could be resolved using the osmostat method presented by Ross *et al.*, although this method also requires running concentration-specific parameterisations of the chemical potential (or differences in chemical potential) for each component of the solution,[168] which would make this a rather complicated solution. Therefore, pure water is used as a reference state throughout this work, largely for convenience and simplicity.

3.4.2 Bulk Water Density

The densities of the water box sampled from GCMC/MD simulations using both the experimental and computational values of μ_{sol}^{ex} and V° are compared against those from the NPT simulation in Fig. 3.1. As can be seen, the results of the comparison are very dependent on the values of these parameters. When the simulated values of $\mu_{sol}^{ex} = -6.09$ kcal mol^{-1} and $V^{\circ} = 30.345$ Å^3 are used, the agreement with the NPT results is very good, with the two distributions showing excellent overlap. However, when using the slightly different, experimental values of $\mu_{sol}^{ex} = -6.324$ kcal mol^{-1} and $V^{\circ} = 30.003$ Å^3, the densities from the GCMC/MD simulations are notably lower, and the distribution overlap with the NPT densities becomes very poor. These observations demonstrate the importance of self-consistency within a simulation, as obtained by determining these parameters under the simulation conditions of interest — if GCMC/MD were carried out under conditions different to these, it would be advisable to recalculate these values. This self-consistency ensures that a simulated system is at



(A) $\mu_{sol}^{ex} = -6.09$ kcal mol^{-1} and $V^{\circ} = 30.345$ Å^3.

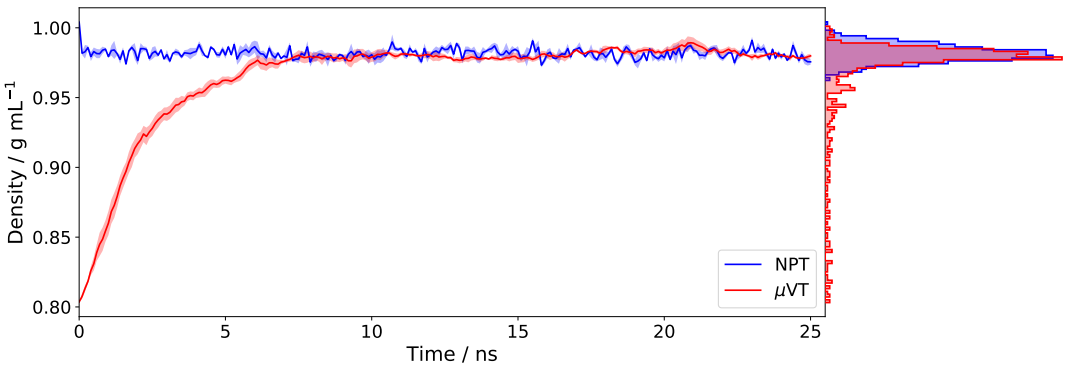


(B) $\mu_{sol}^{ex} = -6.324$ kcal mol^{-1} and $V^{\circ} = 30.003$ Å^3.

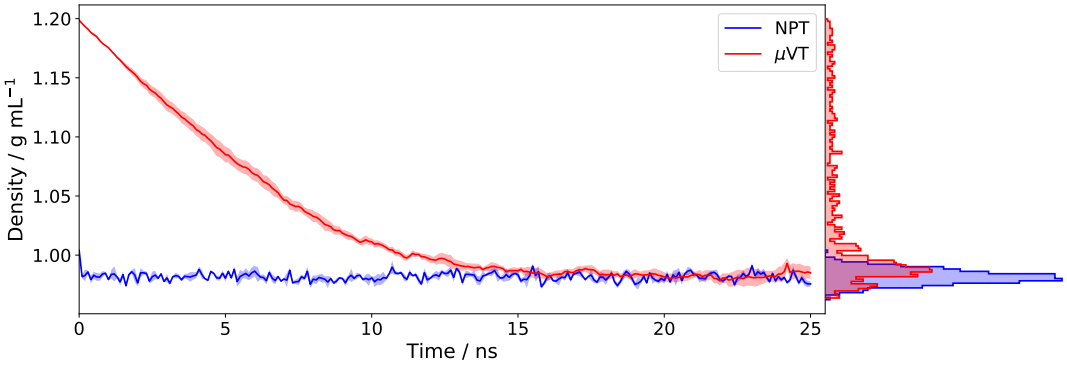
FIGURE 3.1: Comparison of the water densities observed using both constant pressure MD (blue) and GCMC/MD (red). The GCMC/MD densities were collected using both the calibrated values of μ_{sol}^{ex} and V°, as well as the experimental values. In each case, the solid line represents the mean density from the three repeats, and the shaded region covers the values within one standard error of the mean.

equilibrium with a simulated water box.

The results shown in Fig. 3.2 show the densities sampled using GCMC/MD when the starting density is far from the equilibrium value (approximately 20 % too high or too low). As can be seen, in both cases, the GCMC/MD densities converge within 10-15 ns, and then agree well with the NPT results. This is a reassuring result, which further evidences that the GCMC implementation and the accompanying thermodynamic parameters are physically correct. It should be noted that the equilibration times are fairly long, but this is because sampling density fluctuations by adding and removing one water molecule at a time is not an efficient approach. It would not normally be recommended to run GCMC sampling of water over such a large volume, but in this case, it provides a robust test for the methodology, where any errors in theory or implementation would likely be amplified.

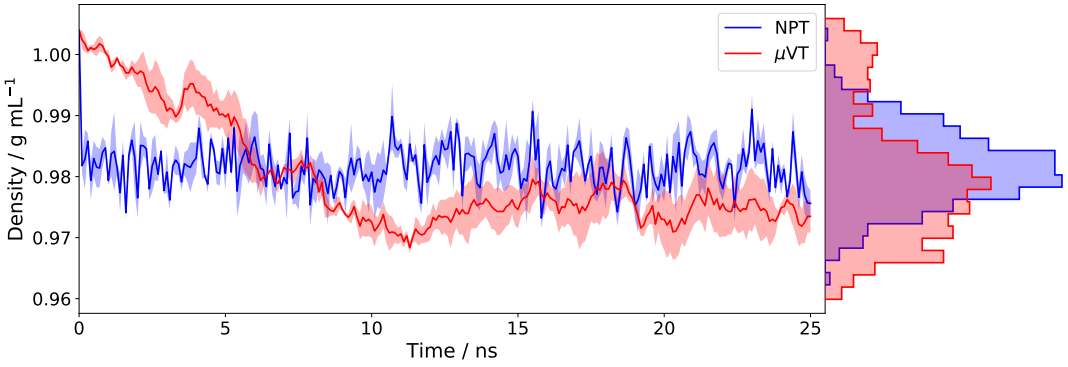


(A) Initial density of 0.804 g mL^{-1}.

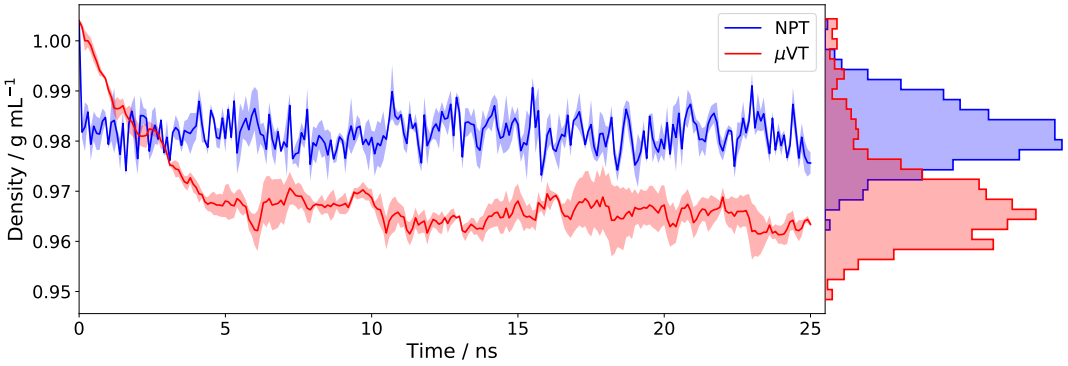


(B) Initial density of 1.199 g mL^{-1}.

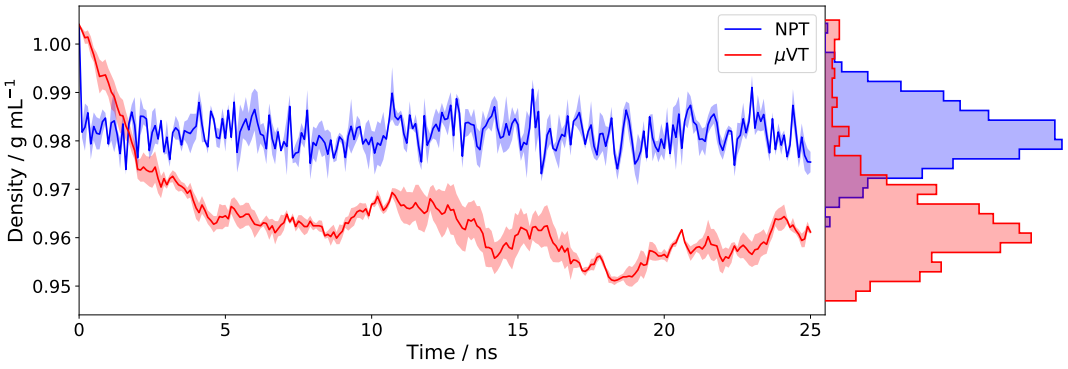
FIGURE 3.2: Comparison of the water densities observed using both constant pressure MD (blue) and GCMC/MD (red). The GCMC/MD simulations were started using densities of 0.804 and 1.199 g mL^{-1}. In each case, the solid line represents the mean density from the three repeats, and the shaded region covers the values within one standard error of the mean.



(A) $\mu_{sol}^{ex} = -6.15$ kcal mol^{-1} and $V^{\circ} = 30.345$ Å^3.



(B) $\mu_{sol}^{ex} = -6.20$ kcal mol^{-1} and $V^{\circ} = 30.345$ Å^3.



(C) $\mu_{sol}^{ex} = -6.25$ kcal mol^{-1} and $V^{\circ} = 30.345$ Å^3.

FIGURE 3.3: Comparison of the water densities observed using both constant pressure MD (blue) and GCMC/MD (red). The GCMC/MD densities were collected using values of -6.15, -6.20 and -6.25 kcal mol^{-1} for the excess chemical potential. In each case, the solid line represents the mean density from the three repeats, and the shaded region covers the values within one standard error of the mean.

The results showing the sensitivity of GCMC/MD to the value of the excess chemical potential are shown in Fig. 3.3. As can be seen, the density values sampled using GCMC/MD are highly sensitive to the value of this parameter, as even with $\mu_{sol}^{ex} = -6.15$ kcal mol^{-1} (a difference of -0.06 kcal mol^{-1} from the calibrated value) a clear difference is seen from the densities observed using constant pressure MD. This sensitivity appears somewhat alarming, but it should be noted that (as previously mentioned) for a GCMC volume which is so large, with so many water molecules, any errors will be magnified. It is therefore likely that when applying GCMC to smaller regions, such as protein binding sites, this sensitivity would be less pronounced. In any case, this further reinforces the recommendation that the values of μ_{sol}^{ex} and V° should be reparameterised when simulating under conditions significantly different to those described in this chapter.

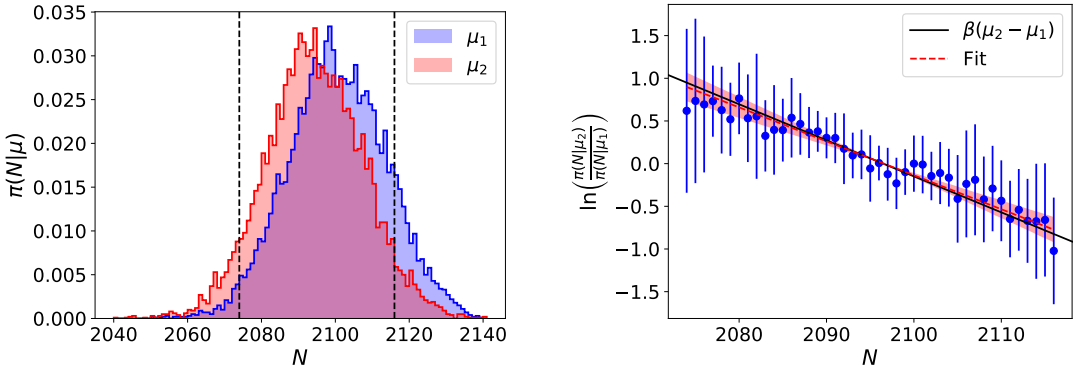
3.4.3 Ensemble Validation

As previously described, the validation method described by Shirts (as carried out in this work) involves calculating the probabilities of observing each value of N at two chemical potentials, and then plotting the natural logarithm of the probability ratio against N, from which the gradient can be compared to the theoretical value (Eq. 3.3).[192] The distributions of N observed at each chemical potential (or B value) are shown in Fig. 3.4a. As can be seen visually, there is a significant overlap between the two histograms, which provides plenty of data for comparison. The overlap region considered for analysis was taken as $2074 \leq N \leq 2116$ — this was chosen as every value of N in this range was sampled at least once by all independent repeats. The probability of observing each value of N in this range was calculated separately for each independent repeat, then probability ratios were calculated for all values of N by comparing the probabilities from all ten independent repeats at μ_1 against all ten repeats at μ_2, yielding 100 values of the logarithm of the probability ratio (Eq. 3.3) for each value of N.

The fitting procedure was carried out using a bootstrapping analysis. For each bootstrap, a subset of the data was collected by selecting one value of the logarithm of the probability ratio (corresponding to one random comparison between independent repeats) for each value of N in the overlap region, and a first order polynomial was then fitted to this data. This was repeated for 50,000 bootstraps, yielding the data shown in Table 3.1 for the gradient and intercept. The mean gradient observed over all bootstraps was -3.978×10^{-2}, with a standard deviation of 0.754×10^{-2}. This is in good agreement with the theoretical value of $\beta(\mu_2 - \mu_1) = -4.222\times10^{-2}$, when considering the uncertainty in the fitted gradient — the difference between the theoretical value and the mean fitted value is 0.323σ, where σ is the standard deviation of the fitted gradients. The agreement between the two gradients is shown visually in Fig. 3.4b, though there

Parameter	Theoretical value	Mean (fit)	Std. dev. (fit)
Gradient	-4.222×10^{-2}	-3.978×10^{-2}	7.538×10^{-3}
Intercept	Unknown	83.41	15.80

TABLE 3.1: Mean values of the gradient and intercept observed over all bootstrap samples carried out, including the standard deviations of these values. The theoretical value of the gradient (Eq. 3.3) is also included, for reference — note that the theoretical value of the intercept is not known. All values are given to four significant figures.



(A) Distributions of N observed for the two simulated values of μ.

(B) Comparison of the raw and fitted data against the theoretical gradient.

FIGURE 3.4: Results obtained from the analysis of the grand canonical sampling carried out by *grand*. (A) Distributions of N observed at the two different chemical potential values. The overlap region considered for analysis is indicated with dashed vertical lines. (B) The fitted gradient is shown in comparison to both the raw data and the theoretical gradient, where the red shaded region corresponds to one standard deviation either side of the mean fit. For the raw data, the natural logarithm of the probability ratio is calculated for all 100 comparisons of individual repeats at μ_1 against those of μ_2, with the mean value plotted, and the error bars represent the standard deviation.

is significant noise in the simulated data, owing to the slow convergence of GCMC sampling for such a large system. This analysis provides further evidence that the implementation of GCMC in *grand* appears to be correct.

3.4.4 Bovine Pancreatic Trypsin Inhibitor

The water locations observed in both the GCMC/MD and NVT production runs were clustered, in order to compare the simulated positions to the crystallographic water sites. This analysis was carried out using average-linkage hierarchical clustering (via SciPy[202]) with a cutoff of 2.4 Å — distances between water observations from the same simulation frame were set to an arbitrarily high value, in order to prevent the merging of distinct water sites — the location of each cluster was then taken as the position of the water snapshot closest to the average coordinate. The cluster positions from the two sets of simulations are shown alongside the experimental positions in Fig. 3.5,

where it can be seen that the GCMC/MD results agree much better with the experimental data than the NVT results. From the GCMC/MD results, all three experimental water sites are reproduced to within 0.6 Å, and are highly occupied (greater than 99 %), which is consistent with free energy analyses which have indicated these sites to be very stable.[105] However, from the NVT simulation, only two of these sites are reproduced within 1.0 Å, with occupancies of 57 and 51 % — for reference, a water site with a standard binding free energy of zero would be expected to show an occupancy of 50 % (though this assumes that water binding/unbinding events are well sampled). Both sets of results show additional sites, but these are all of sufficiently low occupancy that they can be safely ignored.

It should be noted that GCMC/MD as implemented in *grand* is somewhat more computationally expensive (in terms of wall time) than conventional molecular dynamics in OpenMM — by a factor of ∼4.4 for these BPTI simulations. However, if the GCMC/MD simulations were used as an equilibration tool, rather than for production sampling, this cost would be significantly reduced. For example, the three crystallographic sites in BPTI were all inserted within the first few minutes of GCMC/MD equilibration, whereas this did not occur during two hours of canonical molecular dynamics. It should be noted that the *grand* module has not been extensively optimised for efficiency.

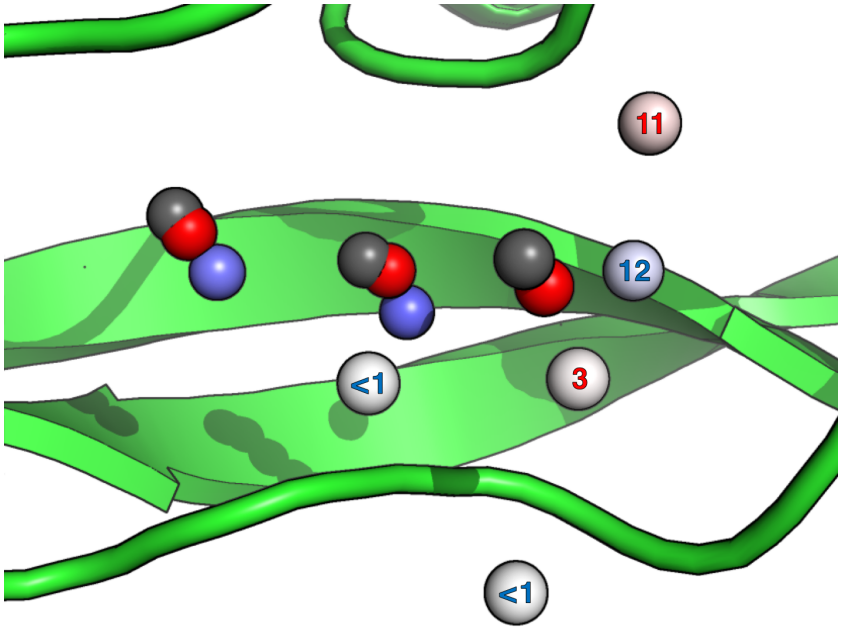


FIGURE 3.5: Comparison of the cluster sites determined from the GCMC/MD (red) and NVT MD (blue) simulations with the experimentally determined water sites (grey). The colours of the cluster sites are faded as the occupancy decreases, where white would indicate 0 % occupancy — where the colour is unclear, the site is annotated with the percentage occupancy, coloured accordingly.

3.5 Summary

This chapter presents the *grand* Python module,[113] implemented in this work, which allows the execution of GCMC sampling of water molecules within OpenMM, making use of the highly customisable simulation framework.[114,179] The accuracy of this module has been demonstrated, as the water density distribution sampled at constant chemical potential using *grand* is in very good agreement with that sampled by constant pressure simulations at the same temperature (Figs. 3.1a and 3.2). It was also verified, using the method described by Shirts,[192] that the probability distribution of the particle number correctly responds to changes in the chemical potential (Fig. 3.4). Finally, the method was applied to a protein test case, by reproducing three buried crystallographic water molecules in bovine pancreatic trypsin inhibitor (BPTI), which are not easily reached using conventional molecular dynamics (Fig. 3.5).

However, it should be noted that the density distribution sampled for bulk water is rather sensitive to the value used for the excess chemical potential of water (Figs. 3.1b and 3.3). This underlines the importance of ensuring that the thermodynamic parameters necessary to establish a correct equilibrium — the excess chemical potential and standard state volume of water — are well calibrated. Values for these parameters are provided in this chapter ($\mu^{ex}_{sol} = -6.09$ kcal mol^{-1} and $V^{\circ} = 30.345$ Å^3), but these should be recalibrated if the simulation conditions of interest differ significantly from those described in section 3.3.1. However, the sensitivities demonstrated are for bulk water, which would not normally be simulated using GCMC/MD, so those observed for protein binding sites (which will often be at least two orders of magnitude smaller than the bulk water system simulated in this chapter) will likely be less severe. Another limitation of this implementation is that the acceptance rates observed for the simulations presented in this chapter are very low (around 0.03 %), meaning that a large percentage of the computational effort spent on GCMC sampling is effectively wasted. Efforts to improve the acceptance rate, and also the efficiency, of grand canonical sampling are discussed in chapter 5.

The *grand* module will form the foundation for the work presented in later chapters of this thesis.

Chapter 4

Insights Into Drug Binding to the M2 Protein Using GCMC/MD

The work presented in this chapter was carried out as part of a larger collaboration involving researchers from multiple institutions. This chapter focuses on the aspects of these projects carried out by MLS, however, Athina Konstantinidi (AK, University of Athens) and Dimitrios Kolokouris (DK, University of Oxford) both contributed to the preparation of the protein-ligand complexes, prior to simulation — these contributions are made clear in italics, where relevant. All simulations and analyses presented in this chapter were performed by MLS.

4.1 Introduction

4.1.1 Background

The matrix 2 (M2) protein is a 97-residue, homotetrameric protein, which carries out several roles within the influenza A virus.[203–207] One of these roles is proton transport, which lowers the pH of the viral interior, leading to the unpacking of viral ribonucleoporins,[208] where the transmembrane domain (residues 23-46) is a minimally functional structure for proton transport.[209,210] Studies of the transmembrane domain are easier in practice,[211] and give results which are almost indistinguishable from those of the full protein.[210,212,213] Inhibition of proton transport across M2 has been found to prevent replication of the influenza A virus,[214] and is the motivation for M2 as a drug target.

Adamantyl-amines are a class of influenza drugs which inhibit the M2 proton channel, including amantadine and rimantadine (Fig. 4.1). Crystal structures of these inhibitors (with resolution better than 3.5 Å[212]) have only become available in recent years,[211] allowing detailed studies of the drug-target interface. These drugs are highly non-polar,

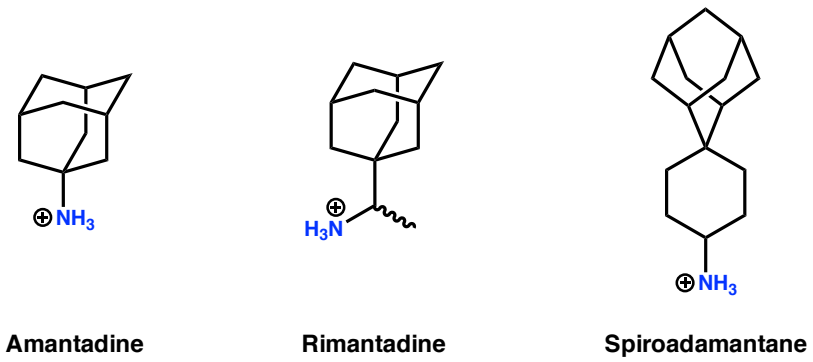


FIGURE 4.1: Structures of the adamantyl-amine drugs, amantadine and rimantadine, and also the spiroadamantane inhibitor.

and yet are very effective inhibitors of the M2 channel. There are several factors of interest here: the drugs have little to no conformational flexibility, and therefore the entropic penalty of binding is significantly reduced.[211] The charged ammonium group appears to take advantage of the fact that the protein has evolved to stabilise hydronium ions at multiple locations in the channel,[215] and the steric bulk of the ligand is thought to provide a very complementary fit to the shape of the channel pore.[216] However, clinical use of amantadine and rimantadine has been discontinued, owing to widespread resistance.[217,218] Spiroadamantane (Fig. 4.1), is a spiro-adamantyl-amine, which was rationally designed to inhibit the amantadine-resistant V27A mutant of M2, as well as the wild type (WT).[216]

Despite the fact that amantadine and rimantadine can no longer be administered, M2 remains a promising drug target, so the interactions of these potent inhibitors with the target warrant further study. Of particular interest in this work is role of water molecules in the M2 channel. Water is not only vital for the transport of protons through the channel, but has also been observed to form organised networks which interact directly with the bound drugs.[211] Fig. 4.2 shows several of the crystallographic arrangements of water in the complexes between M2 and the three ligands studied in this work. A key detail is that the waters appear to bind in 'layers', where two layers of waters are observed for both amantadine and rimantadine in complex with the WT protein.[211] Interestingly, spiroadamantane forms one layer of water in complex with the wild type (the ligand displaces the upper water layer), and two in complex with the V27A mutant — this detail is thought to be related to its ability to inhibit both forms of the protein.[219]

Owing to the importance of water in the binding of drugs to the M2 transmembrane domain, it is of interest to investigate what insights can be offered using GCMC/MD simulations. The work presented in this chapter is from two related projects, based on different aspects of drug binding to the M2 channel. These are discussed further below.

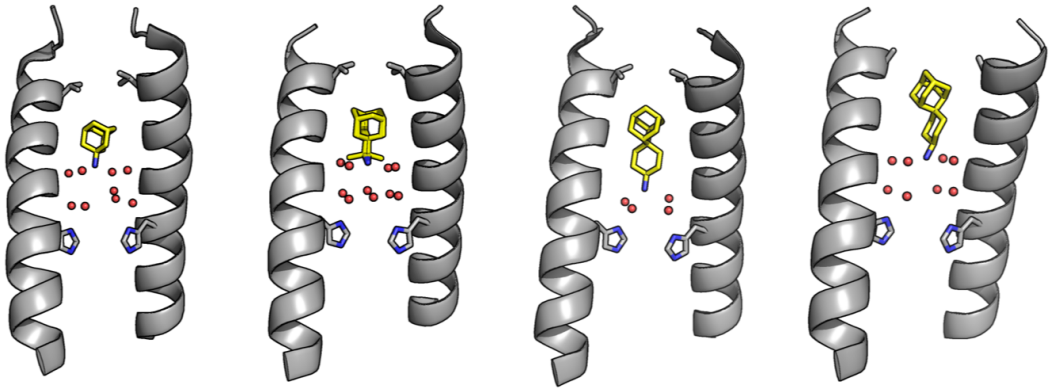


FIGURE 4.2: Crystal structures of amantadine, rimantadine and spiroadamantane in complex with the transmembrane domain of M2. The side chains of residue 27 and His37 are shown for reference, and two of the protein helices are omitted for clarity. (A) Amantadine in complex with the WT structure (PDB ID: 6BKK,[211] 2.00 Å). (B) Rimantadine (racemic) in complex with the WT structure (PDB ID: 6BKL,[211] 2.00 Å). (C) Spiroadamantane in complex with the WT structure (PDB ID: 6BMZ,[211] 2.63 Å). (D) Spiroadamantane in complex with the V27A structure (PDB ID: 6NV1,[219] 2.50 Å).

4.1.2 Rimantadine Stereoselectivity

When in clinical usage, rimantadine was administered as a racemic mixture. However, given that rimantadine is a chiral compound, it is possible that one of the enantiomers may bind more strongly to the M2 channel than the other — such enantiomeric selectivity would be of interest for the design of future inhibitors. There is some disagreement in the literature as to whether or not the two enantiomers bind differently to M2. Solid state NMR studies have claimed that the *(R)*-enantiomer binds more strongly, based on changes in chemical shifts upon drug binding.[220] However, other experiments have suggested that the inhibition of M2 by the two enantiomers is indistinguishable, using electrophysiological, isothermal titration calorimetry (ITC) and antiviral assays,[221,222] and also *in vivo* studies in mice.[223] In the work presented in this chapter, GCMC/MD titrations (described in section 2.6.3.2) are used to investigate the structure and thermodynamics of the water networks found in the M2 channel, in the presence of each enantiomer, in order to determine any atomic-level differences between the two enantiomers.

4.1.3 V27A Resistance

As previously mentioned, resistance of M2 to the adamantyl-amine drugs has become ubiquitous, leading to their withdrawal as treatments for influenza.[217,218] Resistance can be conferred by a number of mutations to the M2 protein,[224,225] but only three tend to be observed in transmissible variants of influenza: L26F, V27A and S31N.[217,226,227] Of these, the V27A mutant is of particular interest for two reasons: V27A appears to be the most selected for under drug pressure,[217,226] and this mutant offers total,

rather than partial, resistance to amantadine.[216,228] However, as previously mentioned, spiroadamantane inhibits both wild-type (WT) and the V27A mutant of M2.[216] A recent analysis of crystal structures of spiroadamantane in complex with both of these structures revealed that the ligand binds in slightly different positions to each of these protein structures, notably with one water layer in the WT structure and two in the V27A structure (Fig. 4.2).[211,219] The ability of spiroadamantane to 'shift' its position within the channel, according to the identity of residue 27 was hypothesised by Thomaston *et al.* to be the reason that spiroadamantane is able to inhibit both the WT and V27A forms of M2.[219]

Given that there may be a water-mediated aspect to the resistance (or lack thereof) of the M2 protein to inhibitors, the differences in structure and thermodynamics of the water networks in the binding site of the M2 protein were analysed using GCMC/MD titrations (section 2.6.3.2). Titrations were carried out for both the WT and V27A structures, each run separately in complex with amantadine and spiroadamantane, in order to investigate any possible differences in water binding free energy — especially as there does not appear to be a published crystal structure for amantadine in complex with the V27A channel. One hypothesis considered during this work was that, when amantadine binds to the V27A mutant, a water wire might be able to bypass amantadine, allowing proton transport to continue even in the presence of the bound ligand.

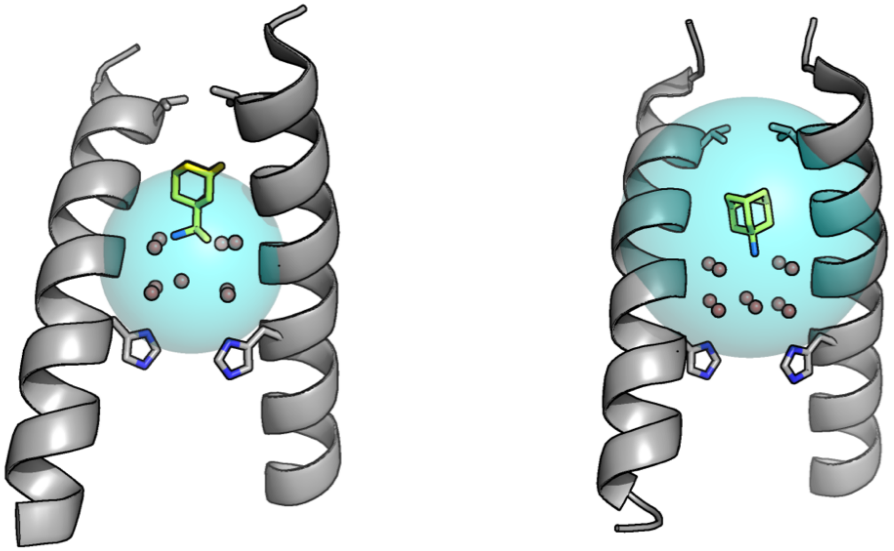
4.2 Simulation Details

Unless explicitly stated otherwise, all simulations reported in this chapter were carried out under the following conditions. The AMBER ff14SB, lipid17 and TIP3P force fields were used to model the protein,[181] lipids[229] and water,[182] respectively, and Joung-Cheatham parameters[183,184] were used for any ions present in the simulation. The ligands were modelled using the general AMBER force field[120] (GAFF) with AM1-BCC charges,[122,123] calculated using *antechamber*[230] — all ligands were simulated as positively charged. A real space interaction cutoff of 12 Å was used, with a switching function applied between 10 and 12 Å, and PME was used to calculate long-range electrostatic interactions.[142] Simulations were carried out at 300 K, using the BAOAB Langevin integrator[158] ($\gamma = 1$ ps^{-1}, $\delta t = 2$ fs). All bonds involving hydrogen atoms were constrained to their equilibrium values, using the SETTLE algorithm for water molecules,[187] and the SHAKE algorithm otherwise.[185,186] Any simulations at constant pressure employed a semi-isotropic Monte Carlo barostat with a pressure of 1 bar and surface tension of zero, with volume changes attempted every timestep. All simulations here used version 7.3.1 of OpenMM.[114,179]

4.2.1 Rimantadine Stereoselectivity

The preparation of the protein structures for both (R)- and (S)-rimantadine as described in this paragraph was carried out by AK. The structures for the *(R)*- and *(S)*-enantiomers in complex with the transmembrane domain of M2 were taken from PDB entries 6US9 and 6US8, respectively. The protein was embedded within a bilayer of approximately 200 POPC lipids, and solvated in a system containing around 18,000 water molecules. Sodium ions were added to neutralise the system charge, and then pairs of sodium and chloride ions were added to reproduce a concentration of 150 mM. For all simulations in this work, the His37 residues were simulated in their neutral, ε-protonated form (where the protons point away from the ligand), as in previously reported MD studies of ligand-bound M2.[211,219]

The rimantadine GCMC/MD simulations were carried out with version 1.0.0 of *grand*,[113] and the GCMC sphere was centred on the mean coordinate of the C_α atoms of the Gly34 tetrad, with a radius of 6 Å (see Fig. 4.3a). For each enantiomer, the following procedure was carried out once. The initial structure was first minimised for 500 steps, to remove any clashes, and all waters present in the GCMC sphere were then removed. The system was then subjected to three stages of GCMC/MD equilibration at $B_{equil} = -6.820$: the first stage involved 50,000 GCMC moves over 1 ps (500 moves per 100 fs), followed by 50,000 moves over 10 ps (1000 moves per 200 fs) and then 100,000 moves over 50 ps (500 moves per 250 fs). The volume of the system was then equilibrated over 500 ps of



(A) Sphere used for rimantadine titrations. (B) Sphere used for V27A titrations.

FIGURE 4.3: Visual representation of the GCMC spheres used for the two different sets of titration calculations. The sidechains of the Val27 and His37 residues are shown for reference. Two of the polymer chains are hidden for ease of visualisation. (A) shows *(R)*-rimantadine and (B) shows amantadine, each in complex with the WT structure.

NPT MD. The resulting structure was then used as a starting point for the independent repeats of each enantiomer.

Each independent titration of each system first involved a further 500 ps of GCMC/MD with a total of 100,000 moves (50 moves per 250 fs) at B_{equil}, to allow some divergence from the shared starting point. The resulting structure was then used for a further equilibration of the same length at each individual B value — 21 evenly spaced Adams values between -24.820 and -4.820 were used. Then, a production GCMC/MD simulation was carried out at each B value, for 100,000 moves over 2.5 ns (20 moves every 500 fs), with simulation frames saved every 5 ps. The average number of waters, $\langle N \rangle$, was recorded for each B value, and used for the calculation of water network binding free energies (Eq. 2.136). Three independent titrations were carried out for both the *(R)*- and *(S)*-enantiomers of rimantadine, bound to the transmembrane domain of the WT structure of M2.

4.2.2 V27A Resistance

The initial structures of both amantadine and spiroadamantane in complex with the WT protein were prepared by AK, and that of spiroadamantane in complex with the V27A mutant was prepared by DK. As there was no known experimental structure of amantadine in complex with the V27A mutant of M2, this structure was created by MLS, by simply mutating the Val27 residues from the WT system. The structure of amantadine in complex with the WT protein was taken from PDB entry 6BKK,[211] that for spiroadamantane was taken from 6BMZ,[211] and the structure of spiroadamantane in complex with the V27A mutant was taken from 6NV1.[219] Similarly to the rimantadine structures, the protein was embedded in a membrane of ∼200 POPC lipids, and solvated with approximately 16,000-19,000 water molecules. Sodium ions were added to neutralise the system charge, and the concentration of sodium chloride was set to 150 mM.

Apart from the differences stated here, the equilibration and titration protocols employed were identical to those used for the simulations of rimantadine in complex with WT M2, as described above. The GCMC/MD simulations used version 1.0.1 of *grand*,[113] with a GCMC sphere centred on the mean coordinate of the C_α atoms of the four Ile32 residues, with a radius of 9 Å (see Fig. 4.3b) — corresponding to a B_{equil} value of -5.604. A larger sphere was used here, in order to account for any possible water-mediated effects around the Val27/Ala27 residues. Accordingly, the Adams values for the titrations were taken as 21 evenly spaced values from -23.604 to -3.604. Three independent titrations were carried out for both amantadine and spiroadamantane, bound to both the WT and V27A structures of the transmembrane domain of M2. It should be noted that the amantadine-V27A simulation was equilibrated for slightly

longer than the other simulations as this was the only system which did not start from a native crystal structure (but instead started from the crystal structure of the WT complex). This extra GCMC/MD equilibration stage took place after the NPT equilibration (before separating the independent repeats), and consisted of 200,000 GCMC moves over 1 ns (50 moves every 250 fs).

4.3 Results

4.3.1 Rimantadine Stereoselectivity

As mentioned in the previous chapter, there is a risk when carrying out GCMC/MD titration calculations that the impact on the waters in the GCMC sphere by insertion and deletion moves will be continually offset by diffusion of waters during the MD portions of the simulations. Fortunately, this does not appear to be the case for the simulations carried out in this chapter. Fig. 4.4 shows the total flux of water observed into the GCMC sphere during the MD portions of the simulations at each B value. Whilst some degree of diffusion of waters in and out of the GCMC sphere was observed — interestingly, this effect appears to be more pronounced for *(R)*-rimantadine than the *(S)*-enantiomer — the magnitude of this effect seems to be rather small, relative to the size of the GCMC sphere and the length of the simulation. This is evidenced by fact that the simulations at low B were able to fully dehydrate the sphere (Fig. 4.5), although these diffusion effects likely account for some of the noise in the titration data. The fact that diffusion of waters does not appear to have a significant impact is likely because

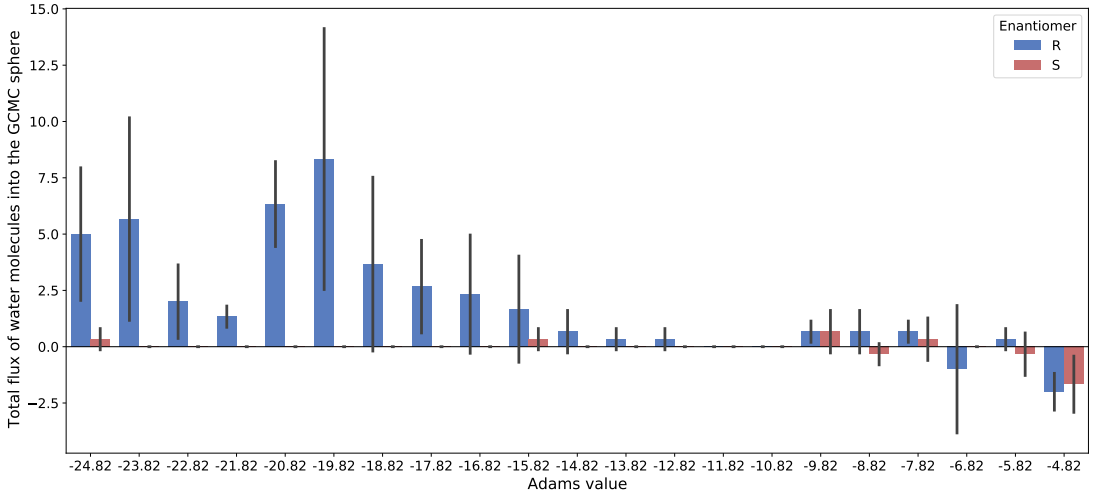
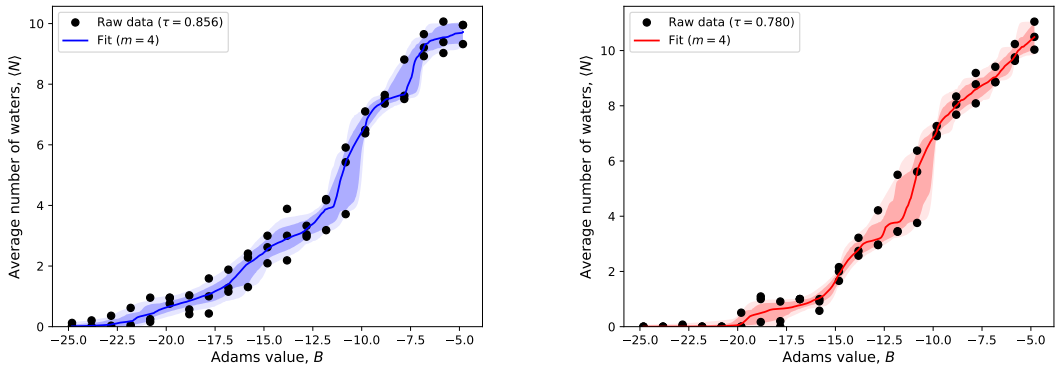


FIGURE 4.4: Bar chart showing the total diffusion of waters into the GCMC sphere during the MD portions of the GCMC/MD simulations at each of the B values. For each simulation, the change in the number of waters over each batch of MD steps was calculated and these values were summed to give the total flux over the course of the simulation. The error bars indicate the standard deviation over the three repeats.



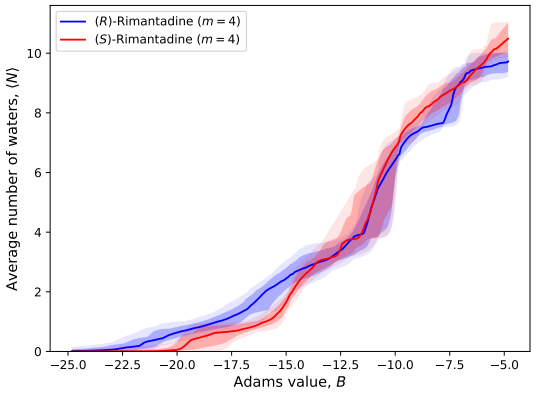
(A) *(R)*-rimantadine. (B) *(S)*-rimantadine.

FIGURE 4.5: Titration data collected for both *(R)*- and *(S)*-rimantadine, along with the fitted curves. The τ values mentioned in the legends represent the Kendall tau coefficients for the raw data (which measure the monotonicity of the data on a scale from 0 to 1). The solid line represents the median fit, and the shaded regions correspond to 68 % and 95 % of the fits — the value of 68 % was chosen in each case, as this would correspond to approximately one standard deviation, if the data were normally distributed. For reference, $B_{equil} = -6.820$.

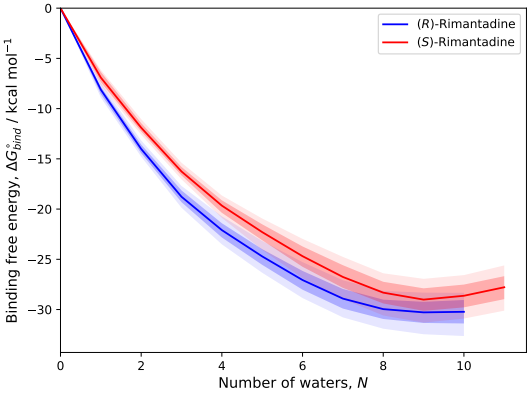
the GCMC sphere is fairly occluded from bulk water, and the ligands are able to block this diffusion to a large extent.

After collecting the average number of water molecules observed in the GCMC sphere at each simulated B value, the titration curves were fitted, using 1000 bootstrap fits. Each bootstrap was carried out by taking one random value of $\langle N \rangle$ from each value of B, and then performing a fit to this subset of the data, using four sigmoid functions (four was qualitatively determined to give the best fit). The fitted titration curves, along with the raw data, are shown in Fig. 4.5 for both enantiomers. Whilst there is some noise in the data (primarily in regions of high gradient), the fits appear good. The two fitted curves are shown alongside each other in Fig. 4.6a. Interestingly, the predicted number of waters in the GCMC sphere at $B_{equil} = -6.820$ agrees quite well between the enantiomers, but the two curves appear to show very distinct hydration profiles with respect to changes in the chemical potential.

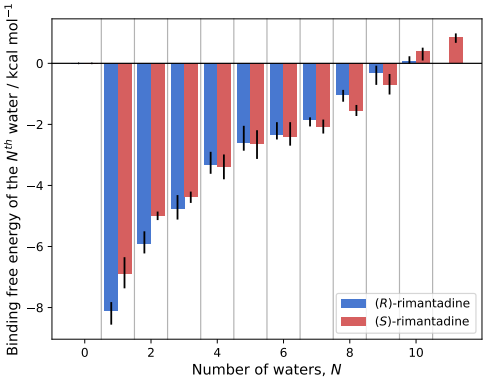
The binding free energy of the water network is plotted with respect to the number of water molecules for both enantiomers in Fig. 4.6b, where both curves show a free energy minimum for $N = 9$ water molecules — for reference, average occupancies of 9.26 $\pm$ 0.17 and 9.04 $\pm$ 0.15 were observed at B_{equil} for *(R)*- and *(S)*-rimantadine, respectively (quoted are the mean and associated standard error of the average number of waters over the three independent repeats). It should be noted that whilst the uncertainty in the free energy curves makes it appear as though there is also a lot of uncertainty in the position of the free energy minimum, this is not the case. Fig. 4.6c shows that 9 is the



(A) Fitted titrations.



(B) Water network binding free energy.



(C) Binding free energy of the N^{th} water.

FIGURE 4.6: Comparisons between the titrations and extracted free energies for the two enantiomers of rimantadine. For (A) and (B), the lines correspond to the median result, and the shaded regions correspond to 68 % and 95 % of the data. For (C), the bars are drawn at the median free energy, and the error bars correspond to 68 % of the data. For reference, $B_{equil} = -6.820$

largest value of N which introduces a negative change in binding free energy for the network (for both enantiomers) and the uncertainties in this value do not dispute $N = 9$ as the free energy minimum. The uncertainty observed in Fig. 4.6b is an accumulation of the uncertainties shown in Fig. 4.6c. This observation is in good agreement with the crystal structures, where 9 and 10 waters are observed for the *(R)*- and *(S)*-enantiomers, respectively — in the latter case, the crystallographic occupancies of the 10 waters sum to 9.23 $\pm$ 0.06 (quoted are the mean and standard error over the four protein units in the crystal structure). The titration results predict the free energy change of introducing a tenth water site to be positive, but very small ($+0.06$ [-0.02, $+0.23$] kcal mol^{-1} and $+0.38$ [$+0.09$, $+0.51$] kcal mol^{-1} for *(R)* and *(S)*, respectively), indicating that this additional water is also thermally accessible — the values quoted are the median value from the bootstrap sampling, and those in square brackets represent the range where 68 % of the fits lie (this value was chosen as it would correspond to one standard deviation, if the data were normally distributed). The values of the free energy minimum are -30.3 [-31.3, -29.2] kcal mol^{-1} and -29.0 [-29.8, -27.5] kcal mol^{-1} for the *(R)*- and

(S)-enantiomers, respectively, which would appear to indicate that the water network is slightly more stable in the presence of the *(R)*-enantiomer, however, there is a large overlap between the uncertainties (Fig. 4.6). The distributions of these free energy values are shown in Fig. 4.7, where it seems that the distributions do appear to be distinct, providing further, albeit weak, support for the water network being more stable in the presence of the *(R)*- than the *(S)*-enantiomer, although the magnitude of this difference cannot be precisely determined, owing to the aforementioned large overlap.

Of particular interest from Fig. 4.6b is that the difference in water network free energies between the two enantiomers is very large (even relative to the uncertainty) at lower levels of hydration. For this reason, structural analyses were carried out in order to identify any trends across the B values used in the titration calculations. Some of the structural trends observed are depicted via a series of representative simulation frames, shown in Fig. 4.8. Fig. 4.9 shows a series of distributions for the positions of water molecules along the channel length at different B values for both enantiomers, and similarly, Fig. 4.10 shows a similar distribution for the position of the rimantadine nitrogen atom. At high B values, one can clearly see two broad peaks in the water distribution, corresponding to the two water layers (Figs. 4.8c and 4.8f), which qualitatively appear very similar between the two enantiomers at $B_{equil} = -6.820$ (though there does seem to be a slight difference in the positions of the ligand here). As the Adams value decreases, the more weakly bound water molecules are first preferentially deleted from the upper layer, and the ligand drops further into the channel in order to interact directly with the lower layer (Figs. 4.8b and 4.8e). Finally, when the

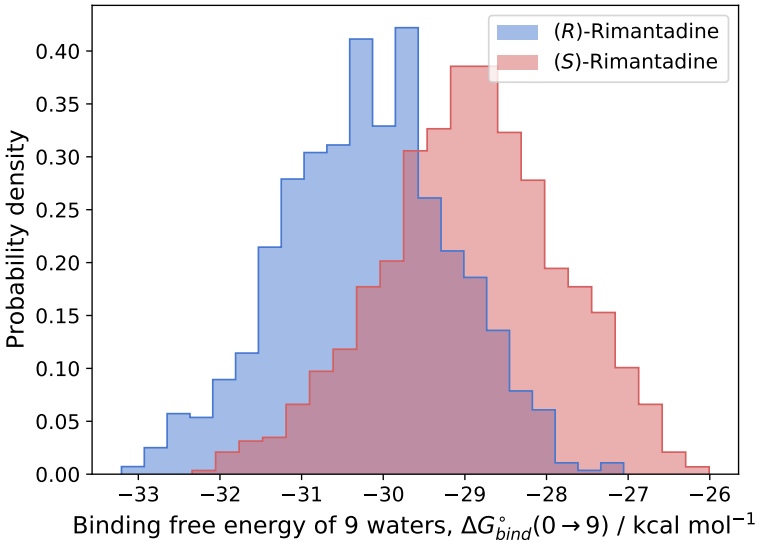
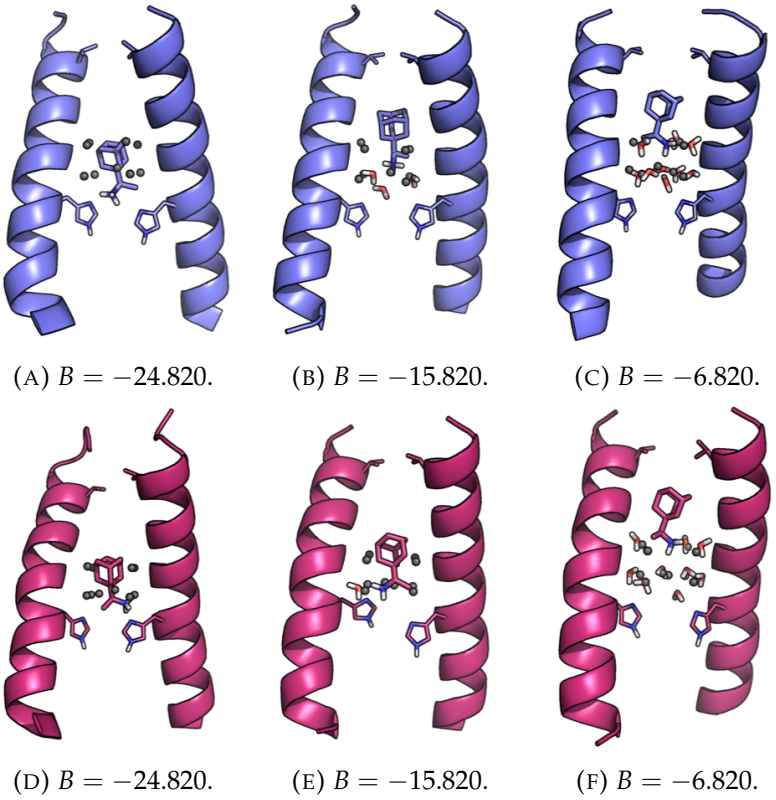


FIGURE 4.7: Distributions of the binding free energies calculated for a 9-water network from the rimantadine titration data. Each of the free energy values were calculated from individual titration bootstraps.



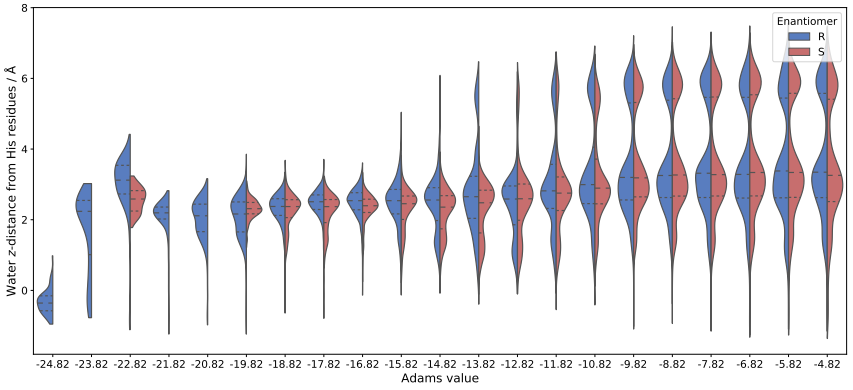
(A) $B = -24.820$. (B) $B = -15.820$. (C) $B = -6.820$.

(D) $B = -24.820$. (E) $B = -15.820$. (F) $B = -6.820$.

FIGURE 4.8: Representative snapshots of the rimantadine simulations at different B values. The top row corresponds to *(R)*-rimantadine, and the bottom to *(S)*-rimantadine. The crystallographic water sites are show as grey spheres, and two of the transmembrane helices are omitted for clarity. For reference, $B_{equil} = -6.820$.

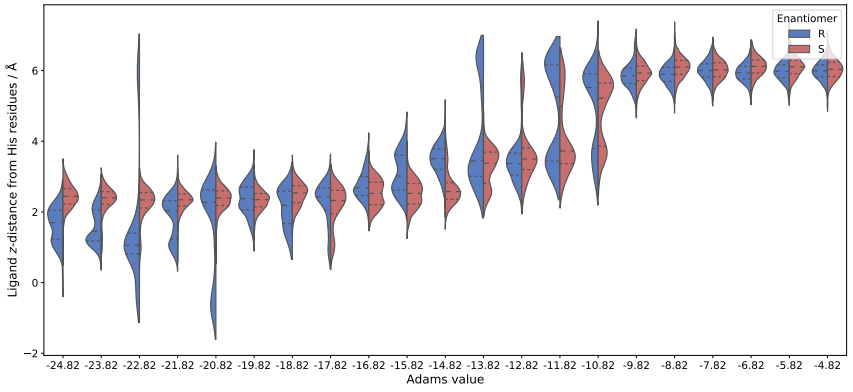
lower layer is removed, the ligand drops again to interact directly with the His37 tetrad (Figs. 4.8a and 4.8d). At very low B values, when the channel is dehydrated, it appears that *(R)*-rimantadine drops slightly further into the channel than *(S)*-rimantadine, providing possible evidence of an apparent chiral effect when the water is removed. It should be noted that a quantitative comparison between the locations of the simulated and crystallographic water sites has not been reported here, owing to the flexibility of the simulated water network. As the waters are highly mobile, they are therefore not suitable for quantitative comparison with the fixed, crystallographic water locations.

In summary, these data appear to suggest that the water network binds very slightly more favourably to the M2 channel in the presence of the *(R)*-enantiomer of rimantadine than the *(S)*-enantiomer (Fig. 4.6b). However, this difference is around 1.3 kcal mol^{-1} (with a large degree of uncertainty) — where 1 kcal mol^{-1} is widely considered to be the sensitivity limit of calculated free energy values[3] — so this free energy difference is compatible with there being no difference in binding affinity between the two enantiomers. However, it is interesting to note that when the channel is partially dehydrated, the difference in binding free energy between the water networks is much



(A) (B)

FIGURE 4.9: (A) Violin plots demonstrating the positions of the waters within the M2 pore at each Adams value for the rimantadine titrations. Each data point is taken as the z-coordinate (the z-axis is perpendicular to the plane of the membrane) of the water oxygen atom, relative to the mean z-coordinate of the C_α atoms of the four His37 residues. The data are normalised such that all violins have the same width, which means the violins at very low B values are in some cases highly distorted by the fact that there are many fewer waters observed. For reference, $B_{equil} = -6.820$. (B) Crystal structure of *(R)*-rimantadine (PDB ID: 6US9), labelled with several z-distances, for reference.



(A) (B)

FIGURE 4.10: (A) Violin plots showing the z-positions of the rimantadine nitrogen atom within the M2 pore at each Adams value, relative to the C_α atoms of the His37 tetrad. The data are normalised such that all violins have the same width. For reference, $B_{equil} = -6.820$. (B) Crystal structure of *(R)*-rimantadine (PDB ID: 6US9), labelled with several z-distances, for reference.

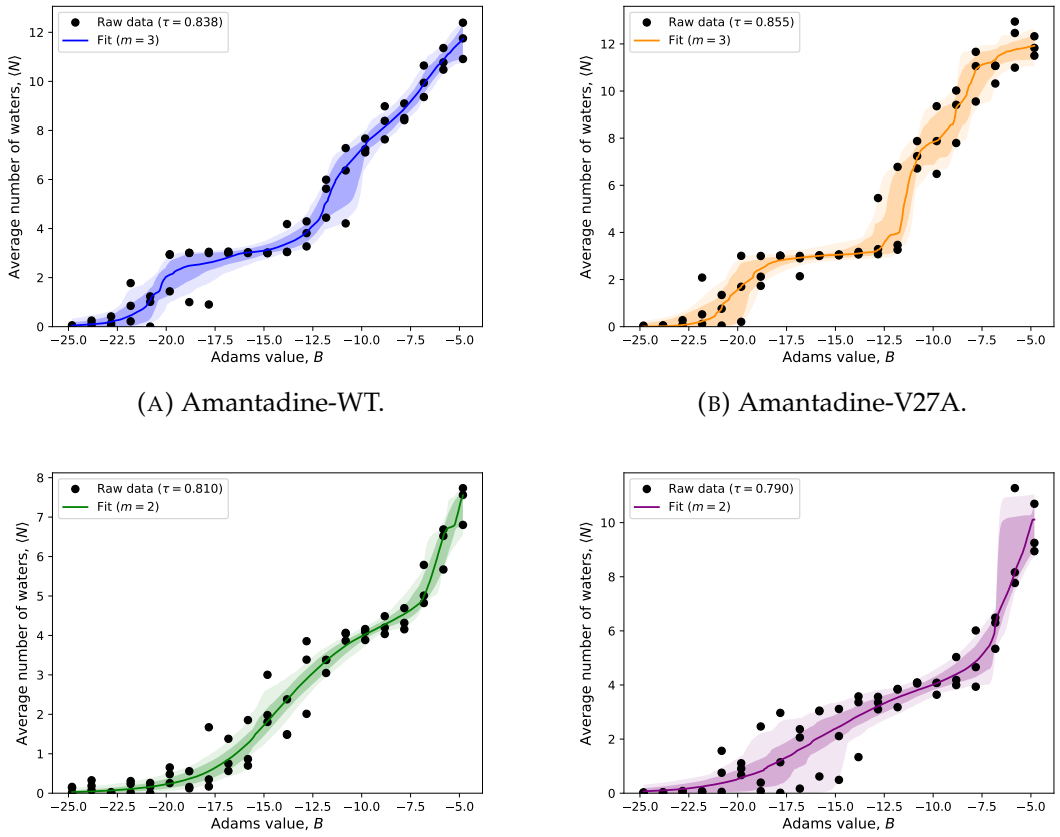
more significant, relative to the corresponding uncertainties (Fig. 4.6b). This would appear to indicate that a chiral difference becomes apparent when there are fewer waters bound to the channel, as supported by the structural trends observed at low B values (Figs. 4.8, 4.9 and 4.10). It is hypothesised that at B_{equil}, the larger water network is sufficiently flexible to adapt to the protein-ligand complex, cancelling the chirality of the protein, such that the ligand appears to interact with an achiral environment. However, at low B values, the reduced flexibility of smaller water networks would mask

the chirality of the protein less effectively, such that the stereochemistry of the ligand becomes relevant. Whilst this does not affect rimantadine, as these poorly hydrated configurations are likely not thermally accessible, this might be relevant for future M2 inhibitors, where those which displace the upper water layer may demonstrate enantiomeric selectivity.

4.3.2 V27A Resistance

As previously mentioned, a larger GCMC sphere was used for the V27A titrations, in order to capture any potential water effects around residue 27 — possibilities considered include the binding of a water layer above the amantadine position (owing to the increased space from the removal of the Val27 side chains), and the formation of a water wire in the amantadine-V27A structure, which might allow proton conductance across the channel. Visual inspection of the simulation trajectories indicated that this was not the case (see Fig. 4.13), so the titration data were post-processed using the same GCMC sphere as used for the rimantadine simulations, in order to focus the analysis on the water layers between the ligand and the His37 tetrad, with the raw data and fitted curves (using the same bootstrapping procedure as described previously) shown in Fig. 4.11. Note that the B values are shifted to reflect the fact that the volume used for the free energy analysis is smaller than that used for sampling. The quality of the data is generally very good, except in the case of spiroadamantane-V27A, where there is significant noise in the raw data at lower B values — this will increase the uncertainty in the extracted free energy data.

For both amantadine and spiroadamantane, the water titrations and free energies are compared between the WT and V27A simulations in Fig. 4.12. The fitted titration curves for amantadine bound to the WT and V27A structures (Fig. 4.12a) are remarkably similar at low-medium B values, and at higher B values, more waters are observed in the V27A structure — when the Val27 side chains are removed, the ligand has more space to occupy higher positions within the channel, thereby allowing slightly more water to bind (this is discussed further in a later section). Fig. 4.12c shows that the free energy curves are also very similar for amantadine, except that the binding of additional waters is slightly more favourable, with a slightly lower network binding free energy in the V27A structure. As shown in Fig. 4.12e, the optimal number of waters is 10 in the WT protein (in agreement with the crystallographic data), and 11 in the V27A structure (the $N = 12$ state also appears very accessible, owing to the very small change in free energy upon the binding of the 12^{th} water) — at B_{equil}, average occupancies of 9.98 ± 0.30 and 10.82 ± 0.20 are observed, respectively. Overall, there does not appear to be any large thermodynamic difference in the water network of amantadine-bound



(A) Amantadine-WT.

(B) Amantadine-V27A.

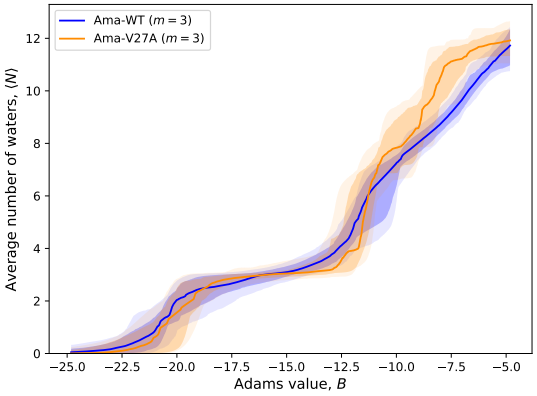
(C) Spiroadamantane-WT.

(D) Spiroadamantane-V27A.

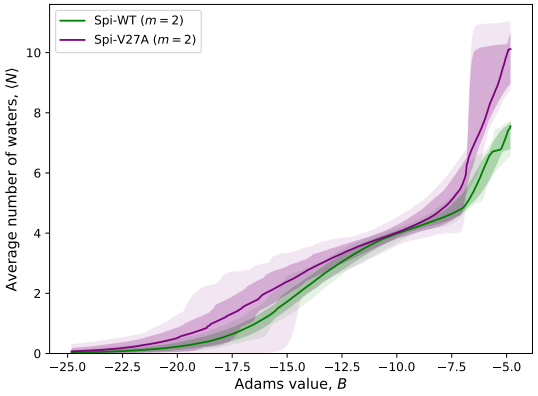
FIGURE 4.11: Titration data collected for amantadine and spiroadamantane, in complex with the transmembrane domains of both wild type M2 and the V27A mutant. The solid line represents the median fit, and the shaded regions correspond to 68 % and 95 % of the fits. Note that the B values are all shifted, relative to those at which the simulations were performed, owing to the use of a smaller volume to process the simulation data. For these analyses, B_{equil} corresponds to -6.820, owing to the focus on a smaller volume.

M2 between the WT and V27A structures.

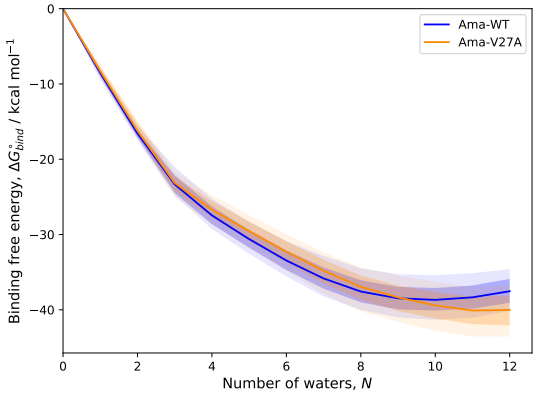
The titration curves for the spiroadamantane complexes (Fig. 4.12b) are also fairly similar between the WT and V27A structures, except for the large divergence at high B values, which corresponds to the binding of the second water layer observed in the crystal structure of spiroadamantane-V27A.[219] The free energy curves show that the water network appears to be several kcal mol^{-1} more favourable in the V27A structure than WT, but the exact difference is not clear, owing to the large uncertainty in the free energy for the V27A simulation (caused by the aforementioned noise in the corresponding titration). However, the data is clear that the optimal number of waters in the WT and V27A systems are 5 and 6, respectively, when spiroadamantane is bound (Fig. 4.12f) — these are respectively in agreement with average occupancies of 5.20 $\pm$ 0.24 and 6.04 $\pm$ 0.29 at B_{equil}. The crystal structure of spiroadamantane bound to the



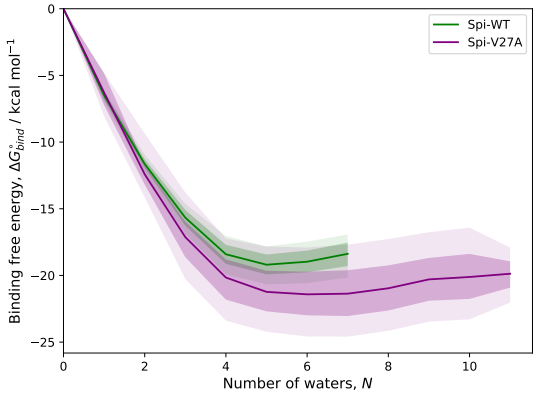
(A) Fitted titrations (amantadine).



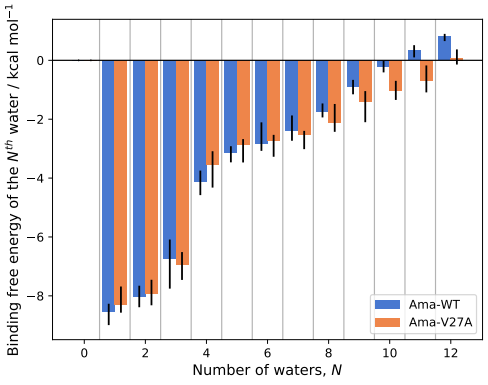
(B) Fitted titrations (spiroadamantane).



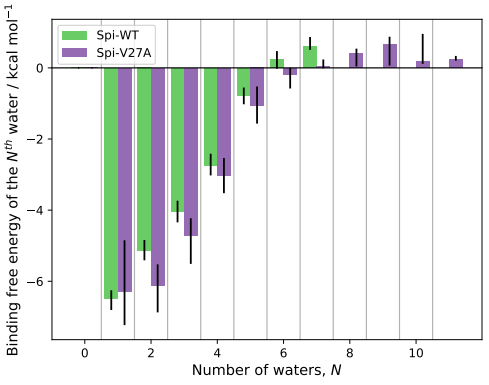
(C) Binding free energy curves (amantadine).



(D) Binding free energy curves (spiroadamantane).

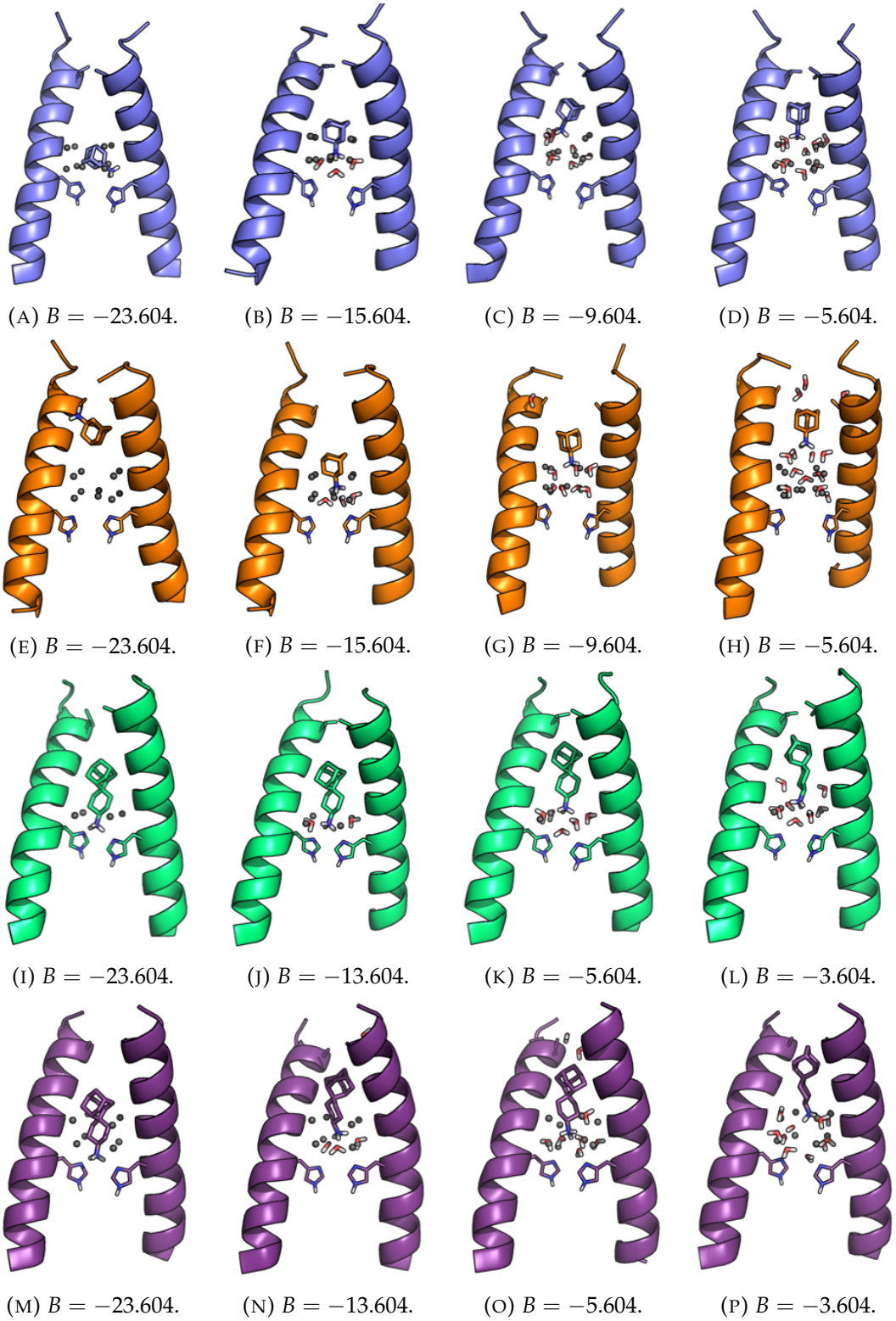


(E) Binding free energy of the N^{th} water (amantadine).



(F) Binding free energy of the N^{th} water (spiroadamantane).

FIGURE 4.12: Comparisons between the titrations and extracted free energies for the WT and V27A titrations, for both amantadine (left column) and spiroadamantane (right column). For (A-D), the lines correspond to the median result, and the shaded regions correspond to 68 % and 95 % of the fits. For (E-F), the bars are drawn at the median free energy, and the error bars correspond to 68 % of the data. For these analyses, B_{equil} corresponds to -6.820.



(A) $B = -23.604$. (B) $B = -15.604$. (C) $B = -9.604$. (D) $B = -5.604$.

(E) $B = -23.604$. (F) $B = -15.604$. (G) $B = -9.604$. (H) $B = -5.604$.

(I) $B = -23.604$. (J) $B = -13.604$. (K) $B = -5.604$. (L) $B = -3.604$.

(M) $B = -23.604$. (N) $B = -13.604$. (O) $B = -5.604$. (P) $B = -3.604$.

FIGURE 4.13: Representative snapshots from the GCMC/MD simulations of amantadine and spiroadamantane in complex with both the WT and V27A structures of M2, at different Adams values. First row: amantadine-WT, second row: amantadine-V27A, third row: spiroadamantane-WT, fourth row: spiroadamantane-V27A. Two of the protein chains are omitted for ease of visualisation, and the crystallographic water sites are shown as grey spheres, for reference — for amantadine-V27A, the crystallographic sites are taken from the WT crystal structure. B_{equil} corresponds to -5.604.

WT protein shows 4 water sites bound in a single layer — the additional site observed in the simulations sits slightly below this water layer (Fig. 4.13k). A more surprising difference is that the crystal structure of spiroadamantane with the V27A mutant contains 8-9 water molecules in two layers, whereas 6 is identified as the optimal number here. Visual inspection of the simulations shows a single water layer (as observed in the WT simulations), but with a single site (occasionally two) occupied in the upper layer (Fig. 4.13o), rather than a full second layer. This discrepancy between the simulated and experimental data for spiroadamantane-V27A is discussed further in the structural analysis below.

A series of representative snapshots were extracted from each of the simulations at different B values, and are shown in Fig. 4.13. As discussed for rimantadine, a similar set

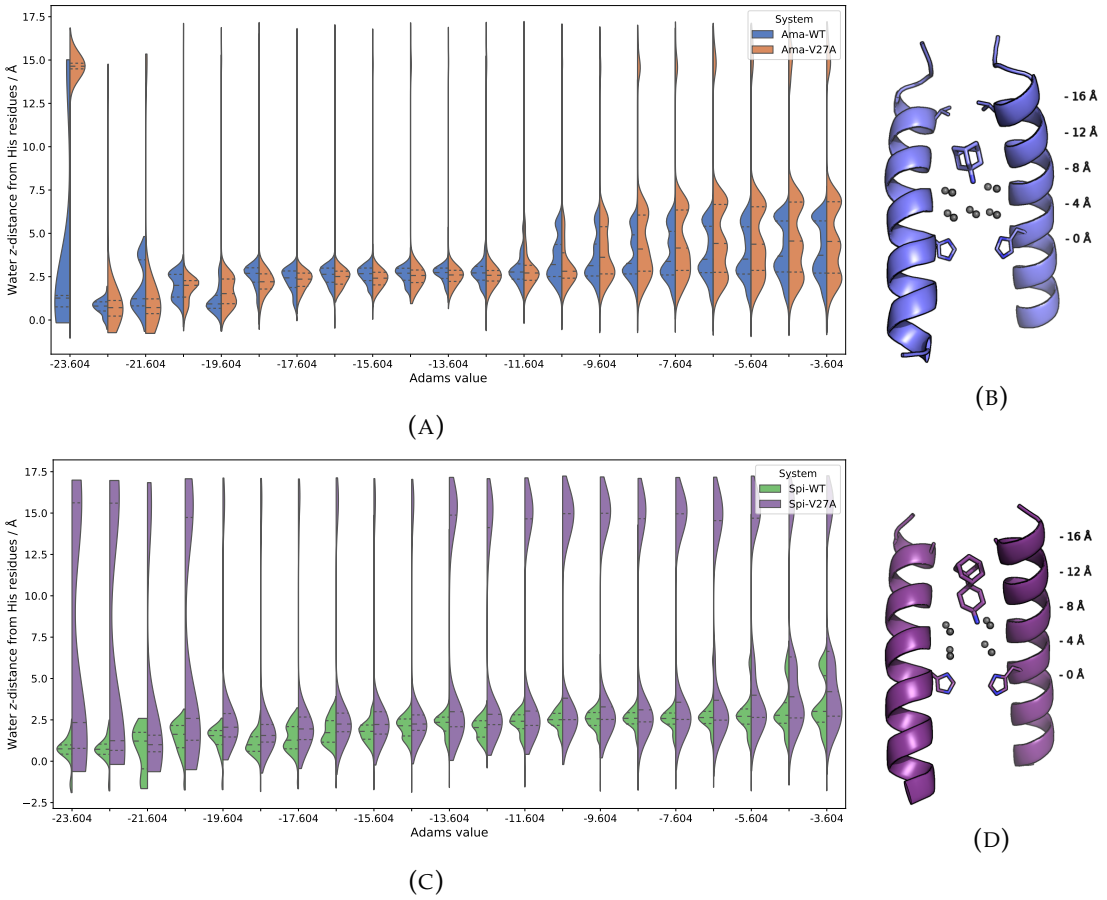


FIGURE 4.14: Violin plots demonstrating the z-positions of the water oxygen atoms within the M2 pore at each Adams value for the V27A titrations, relative to the C_α atoms of the His37 tetrad. The data are normalised such that all violins have the same width, leading to some visualisation artefacts at low B values, where many fewer waters are observed. For these analyses, B_{equil} corresponds to -5.604. (A) and (C) show the violin plots for amantadine and spiroadamantane, respectively. (B) and (D) show the crystal structures of amantadine-WT and spiroadamantane-V27A, with labelled z-distances for reference.

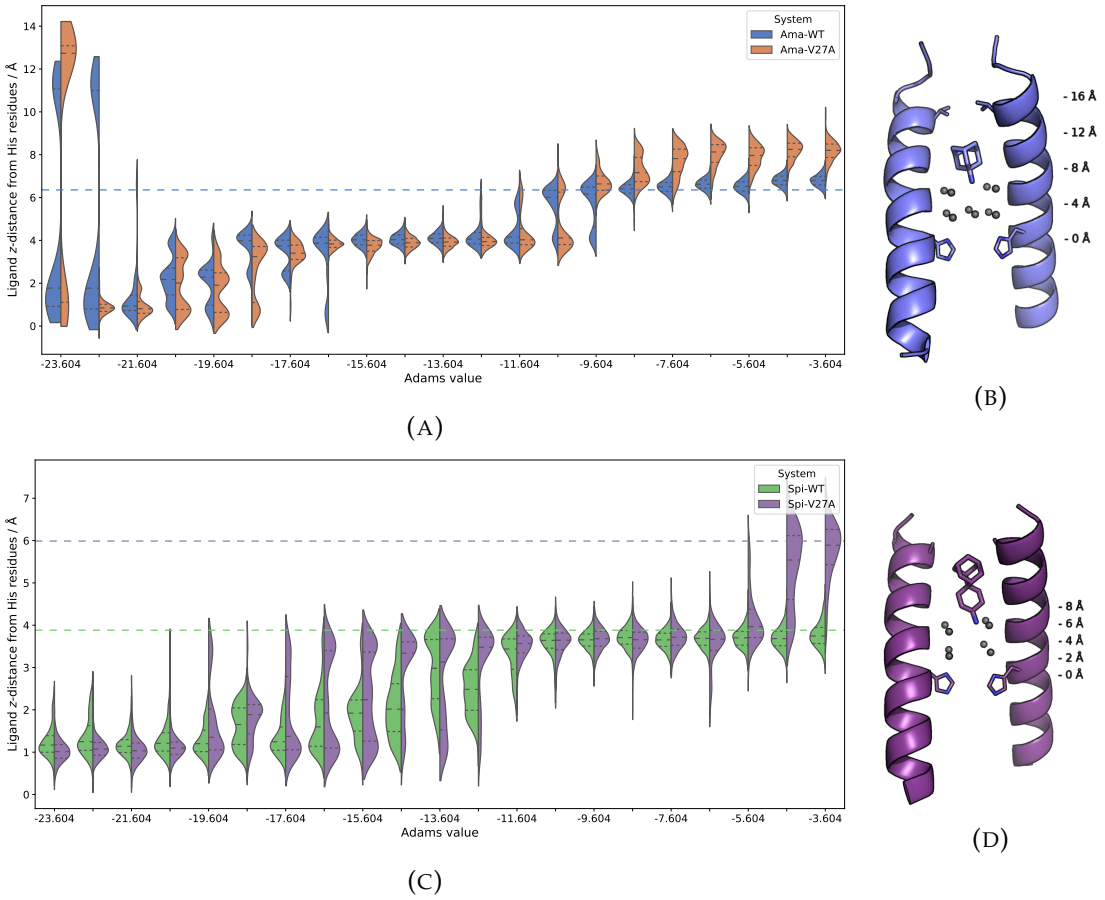


FIGURE 4.15: Violin plots showing the z-positions of the ligand nitrogen atom within the M2 pore at each Adams value for the V27A titrations, relative to the C_α atoms of the His37 tetrad. The data are normalised such that all violins have the same width. The value of this coordinate observed in the appropriate crystal structure is included as a dashed line, for comparison. For these analyses, B_{equil} corresponds to -5.604. (A) and (C) show the violin plots for amantadine and spiroadamantane, respectively. (B) and (D) show the crystal structures of amantadine-WT and spiroadamantane-V27A, with labelled z-distances for reference.

of violin plots for the distributions of the water and ligand positions in the M2 channel are provided for both amantadine and spiroadamantane in Figs. 4.14 and 4.15. The general trends observed for amantadine (Figs. 4.14a and 4.15a) are similar to those observed for rimantadine, in that the more weakly bound waters are first removed from the upper layer, and the ligand drops further into the channel accordingly. A difference between the WT and V27A simulations, is that at around B_{equil} and above, amantadine can adopt a higher position in the V27A channel than in the WT channel, owing to the removal of the Val27 'valve'[219] (Fig. 4.13h). Interestingly, the water distributions at the corresponding B values show three peaks (between 0 Å and around 7.5 Å on Fig. 4.14a), rather than two, indicating that the additional space created causes the water network to form three layers, as shown in Fig. 4.13h. At very low B values, amantadine is sometimes observed at an unusually high position within the channel, with visual inspection revealing that the ligand had turned upside-down at the lower level of hydration, as

shown in Fig. 4.13e. Initially, this was of little interest, as it did not impact the titration calculation (the number of waters observed was not affected), and was assumed to be a transient artefact of simulating at an unnaturally low chemical potential. Interestingly, however, such a 'flipped' conformation has been proposed by Llabrés *et al.* as an intermediate configuration in the binding and unbinding of amantadine to/from the M2 channel.[231] It should also be noted that water distributions in Fig. 4.14a show no sign of water density in the region blocked by amantadine, indicating no evidence of a water wire (which might allow proton conductance) bypassing the ligand in the resistant mutant — this was supported by visual inspection of the simulation trajectories.

The corresponding violin plots for spiroadamantane (Figs. 4.14c and 4.15c) show that the distributions tend to agree well between the WT and V27A structures at the majority of B values (although there are some differences), except for the very high B values, where the second layer of water binds to the V27A complex. Interestingly, the fact that the upper layer is only partially occupied in the V27A complex at B_{equil} means that spiroadamantane is situated lower in the M2 channel than is shown in the crystal structure (Fig. 4.13o). However, for $B > B_{equil}$, the upper layer of waters is more significantly occupied and the ligand position agrees much more closely with that seen in the crystal structure (Fig. 4.13p). The GCMC/MD data collected in this work therefore indicates that the upper layer of waters is not fully occupied at equilibrium, but that the free energy difference associated with completing the upper layer (and correspondingly shifting the ligand up the channel) is positive, but very small (Fig. 4.12d). Note that, owing to the noise in the binding free energy curve (Fig. 4.12d), it is difficult to assign a precise value to the size of this free energy difference. However, the second layer is observed at $B = -4.604$, which is only one unit more positive than B_{equil}, corresponding to a chemical potential which is only $k_BT = 0.596$ kcal mol^{-1} more positive than that of bulk water, indicating that only a very subtle bias is needed to favour the binding of the second layer. Interestingly, Thomaston *et al.* carried out constant pressure MD simulations of spiroadamantane in complex with the V27A mutant, starting the simulation from a homology model based on the WT structure (i.e. the Val27 residues were mutated into alanine), where they report the binding of the second water layer after 300 ns.[219] Given that the analysis above suggests that there is a very small, positive free energy difference associated with the binding of the full second water layer, it is very possible that the difference between these GCMC/MD simulations and the MD data reported by Thomaston *et al.*[219] might be caused by a difference in force field. That is, the configurations containing two water layers might correspond to the free energy minimum when using a different force field, or under different simulation conditions.

Owing to the discrepancy between the simulated and experimental data for spiroadamantane in complex with the V27A mutant — and that the simulation result appears to be

sensitive to the simulation parameters — the experimental data was inspected more closely. The electron density score for individual atoms (EDIA) is a quantitative measure of how well an atom in a crystal structure represents the raw electron density.[232,233] According to Meyder *et al.*, an EDIA score between 0.8 and 1.2 represents strong electron density evidence for an atom, a score between 0.4 and 0.8 indicates minor inconsistencies between the atom and the electron density, and a score below 0.4 indicates substantial inconsistencies.[232] The EDIA scores for the upper and lower layer waters were calculated from PDB entry 6NV1[219] using the ProteinsPlus server,[234,235] with the EDIA scores plotted in Fig. 4.16 (note that one water which lies between the two layers was ignored, owing to its ambiguity). As can be seen, a number of the upper layer waters from this structure have very poor EDIA scores, and the others are moderate, at best — this appears to support the observation from the titration data that the upper layer is likely only partially occupied at equilibrium. However, it is very curious to note that, whilst the experimental evidence for the upper water layer is weak, this does not appear to be the case for the ligand position. Fig. 4.17a shows the crystallographic binding mode of spiroadamantane to the V27A mutant in comparison with the electron density map, where the agreement appears qualitatively very good, despite the poor agreement between the upper layer waters and the electron density. In contrast, comparison of the simulated structure with the electron density (as shown for a representative frame in Fig. 4.17b) shows good agreement for the waters, but the electron density does not support the lower spiroadamantane position observed in the GCMC/MD simulations. The source of these discrepancies is unclear. One possibility is that these structural differences are caused by differences in temperature — the X-ray diffraction was carried out at 100 K,[219] and the simulations were performed at 300 K. Another is that the crystallographic binding mode of spiroadamantane is the correct one, and that the upper layer waters do bind, but are poorly resolved in the crystal

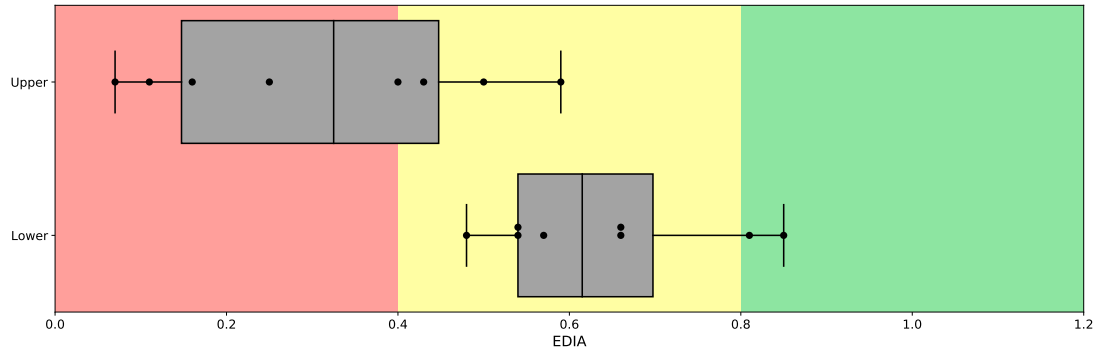
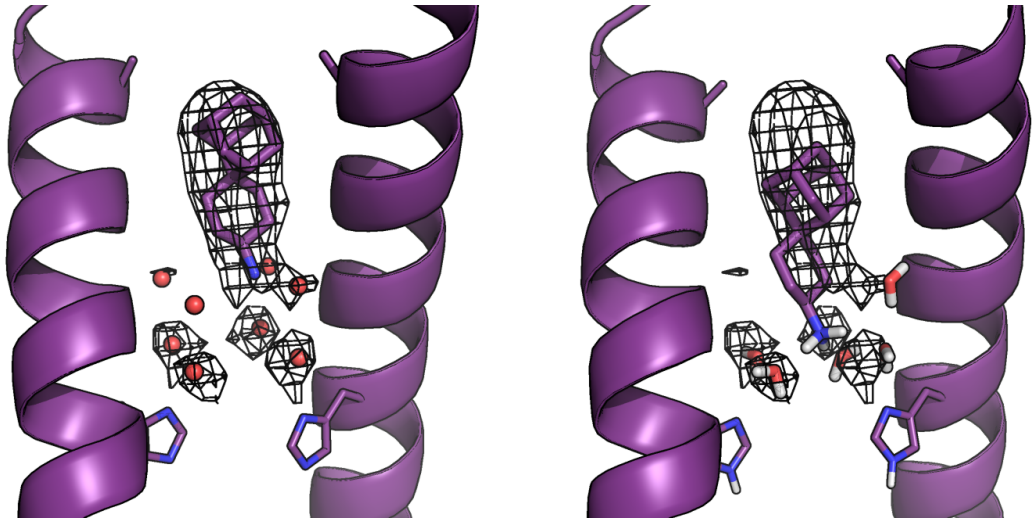


FIGURE 4.16: Box plots for the EDIA scores of the upper and lower layer waters in the spiroadamantane-V27A complex — note that 8 values were obtained for each layer, as the crystal structure (PDB ID: 6NV1[219]) contains two protein tetramers. The raw values for the individual waters are also included as points. The background of the plot is coloured according to the implications of different EDIA scores as described by Meyder *et al.*[232]



(A) (B)

FIGURE 4.17: Comparison with the electron density for the crystallographic binding modes of spiroadamantane in complex with the V27A mutant. The electron density data shown is the $2F_o - F_c$ map contoured at 1σ, showing only the data within 2 Å of the crystallographic spiroadamantane or water positions. (A) Crystal structure (PDB ID: 6NV1,[219] 2.5 Å). (B) Representative frame from the GCMC/MD simulations at B_{equil}.

structure, perhaps owing to disorder in their positions.

4.4 Summary

The work presented in this chapter demonstrates how GCMC titrations can be used in practice for a specific protein of interest, yielding thermodynamic, as well as structural information regarding the binding site waters. Here, GCMC/MD titrations have been applied to the M2 drug target in complex with three key ligands of interest, including two drugs against which it has evolved resistance,[217,218] and a promising inhibitor of the V27A mutant.[216] This has been used to investigate both the role of hydration in stereoselectivity of the chiral drug, rimantadine, and also to investigate the differences in hydration between amantadine and spiroadamantane in complex with both the WT and V27A forms of M2.

The titration calculations performed in this work indicate that the water network binding free energies are very similar between the complexes of WT M2 with the two enantiomers of rimantadine, with that of the *(R)*-enantiomer being slightly more stable. However, the difference ($\sim$1.3 kcal mol^{-1}) is very small,[3] and could be an artefact of the simulation conditions. Therefore, these data are considered to be in support of there being little to no difference in the affinities of the two enantiomers of rimantadine[221–223]

— at least in terms of the role of water in this complex. However, an interesting observation made is that the difference in the stabilities of the water networks is much more substantial at lower levels of hydration (Fig. 4.6b), indicating that there may be a chiral preference when only one of the water layers is bound — this is further supported by the enantiomers adopting different positions within the channel when there are fewer waters bound (Fig. 4.10). It is hypothesised that when two layers of waters are bound, the water network is sufficiently flexible to adapt to the protein-ligand interface, such that the ligand 'sees' an achiral environment, whereas when only one layer is present, the network is not flexible enough to achieve this effect. This observation may be relevant if the next generation of M2 inhibitors displace the upper layer of waters — as does spiroadamantane[211] — as these compounds may exhibit water-mediated stereoselectivity.

When considering the impact of the V27A mutation on the M2-ligand complex, this work has found no clear structural or thermodynamic evidence of a water-mediated mechanism by which this mutant achieves resistance to amantadine, but not to spiroadamantane. The possibility of a water wire being able to bypass the bound amantadine in the V27A structure, allowing the channel to conduct protons even in the presence of the ligand, was considered, but no evidence was found to support this. However, some interesting structural observations were made. For amantadine, the V27A mutation causes the ligand to be shifted up the channel, likely owing to the removal of the Val27 'valve'.[219] Interestingly, this appears to result in the reorganisation of the water network from two layers into three (Fig. 4.13h), although this does not appear to be associated with a significant change in the binding free energy of the network (Fig. 4.12c). It should be noted that this observation cannot be compared to experiment, as there is no known crystal structure of amantadine bound to the V27A mutant. It was suggested by Thomaston *et al.* that spiroadamantane is able to bind to both the WT and V27A forms of M2, because it adjusts its position within the channel according to the identity of residue 27, and, correspondingly, the number of water layers changes from one to two upon the V27A mutation (Fig. 4.2).[219] However, the GCMC/MD simulations carried out at B_{equil} did not support this, with only one water layer observed in the V27A structure, and the spiroadamantane ligand seen in approximately the same position as observed in the WT structure (Figs. 4.13k and 4.13o). Further inspection of the crystallographic data suggested that the electron density evidence for the upper water layer in the V27A structure was quantitatively very weak (Figs. 4.16), but visual inspection of the electron density shows little doubt in the crystallographically observed spiroadamantane position, whereas the position observed in the B_{equil} simulations does not agree well with the ligand electron density (Fig. 4.17). It is therefore expected that the crystallographic binding mode of spiroadamantane is correct, and that the upper water layer is simply poorly resolved (which is very possible for a structure with a resolution of 2.5 Å). It is interesting to note that the crystallographic ligand position is

observed in the GCMC/MD simulations at $B > B_{equil}$ (Figs. 4.13p, 4.14c and 4.15c), indicating a small free energy difference between the one and two layer configurations. In fact, spiroadamantane is observed in the crystallographic position at $B = -4.604$ (Fig. 4.15c), where the corresponding chemical potential is only k_BT is larger than that at B_{equil}, indicating not only that a very subtle bias is needed to cause the second water layer to bind, but that this difference is also thermally accessible. Therefore, the free energy difference between the one and two water layer configurations is likely sufficiently small that observing one or two water layers could be sensitive to the force field — this might explain why the crystallographic pose was observed from MD simulations reported by Thomaston *et al.* [219]

As described above, it does not appear likely (from the results obtained in this work) that the resistance mechanism of the V27A mutation of the M2 protein is water-mediated. Another possibility is that resistance arises from a difference in binding affinity, although such a large drop in amantadine activity[216,228] does not appear likely to arise from a reduction in stability associated with the V27A mutation. It should be noted that whilst it has been demonstrated (via ITC experiments) that the S31N mutation causes a very significant reduction in amantadine binding affinity,[222] such data do not appear to have been published for the V27A mutation, where functional assays seem to be favoured.[216,228] This could be tested computationally by carrying out free energy calculations of M2-ligand complexes, in which the Val27 tetrad is perturbed into an Ala27 tetrad. If the resistance is driven by a difference in binding affinity, the free energy change associated with the mutation would be expected to be large and positive for amantadine, and either small or negative for spiroadamantane. An aspect which has not been considered in this work is the pH-dependence of the ligand binding. Given that the M2 protein is a proton channel, it may be worthwhile to investigate the role of different protein protonation states on the protein-ligand binding. For example, the metadynamics study by Llabrés *et al.* reported that amantadine is significantly less stable in the V27A structure than the WT, citing repulsive interactions between the positively charged amantadine and the His37 tetrad (simulated in the +2 charge state) as a factor.[231] Therefore, constant pH simulations may be of interest in future work on this protein, in order to explicitly capture changes in protonation states.[170,171] It might also be interesting to use this method to carry out the mutation free energy calculations discussed above over a range of pH values. Owing to the possible force field sensitivity noted observed in this work, future work might be best served by carrying out simulations across multiple force fields, to ascertain whether the insights obtained are force field-specific.

Chapter 5

Nonequilibrium Candidate Monte Carlo in the Grand Canonical Ensemble

5.1 Introduction

A weakness of GCMC simulations is that the acceptance rates observed in condensed phases can be very low, and some work in the past has limited grand canonical simulations to low density systems for this reason.[101,105] For example, the equilibrium simulations of bulk water presented in chapter 3 observed acceptance rates of around 0.03 % — whilst the acceptance rate is somewhat system-dependent, they are often approximately on this order of magnitude. This therefore means that a large percentage of the computational effort devoted to GCMC sampling is effectively wasted. This is not ideal, and the very low acceptance rates mean that a very large number of GCMC moves must be attempted in order to yield a sufficient number of accepted moves to provide adequate sampling. Some of the various methods which have been developed to improve the efficiency are described here.

The efficiency of particle insertion/deletion moves can be improved by increasing the acceptance rate, and several approaches have been developed with this aim. The cavity bias method proposed by Mezei seeks to steer the insertion attempts to avoid locations which will result in steric clashes (thereby biasing insertions into cavities), based on pre-sampling a number of points at random within the GCMC volume.[101,102] These pre-sampled points are used to calculate the probability of a site which is not sterically blocked being present, and this probability is used to correct the acceptance criteria of both insertion and deletion moves, in order to maintain detailed balance.[101] For non-spherical molecules such as water, insertion moves can also be rejected even when the

insertion location is good, if the random orientation generated introduces a clash. Orientation biasing methods have been developed which attempt to bias the inserted waters towards more favourable orientations[236,237] — again, this bias must be accounted for in the acceptance criteria. A study by Woo *et al.* combined orientation biasing with a grid-based cavity bias implementation, yielding an acceptance rate of 0.81 % (compared to 0.06 % using unbiased GCMC for the same system).[103] For approaches such as these, it is important to consider whether the additional computational cost introduced by the bias is justified by the increase in acceptance rate: if not, then it would be more efficient to simply perform a larger number of unbiased GCMC moves.

An alternative approach used to improve the efficiency of GCMC moves is to increase the speed of the calculations — particularly those which are rejected — thereby reducing the overall amount of computational time wasted. One such example is the excluded volume mapping method, in which a 3D grid details the sites which are sterically occluded, and if an insertion is attempted at or near one of these positions, the move is quickly rejected, avoiding the need to calculate the full acceptance probability.[64,238,239] Shelley and Patey also point out that, (using such an approach) insertions can typically be rejected much more quickly than deletion moves, so attempting insertions more frequently than deletions (rather than with equal probability, as is more common) can allow a larger number of moves to be executed per unit time, and therefore offering more acceptances[237] — it should be noted that this requires a change in acceptance criteria to account for this bias. In addition to excluded volume mapping, Ben-Shalom *et al.* make use of a rapid, approximate potential energy calculation, and if this indicates a very repulsive interaction, the move is rejected — if the energy difference obtained is less than $+15$ kcal mol^{-1}, then a full energy calculation is used.[64] A unique approach was described by Ross *et al.*, which carries out batches of GCMC moves in a parallelised manner.[110] Many GCMC moves are carried out simultaneously (with each move in the batch assigned an ID number), and if one of the moves in the batch is accepted, the rest of the attempts (those with IDs higher than that of the first accepted move) are discarded, and the accepted microstate is used as the starting point for the next batch. The process is repeated until the prescribed number of moves have been attempted. This approach offers a significant improvement in efficiency by exploiting the low acceptance rates observed during GCMC.[110]

In this chapter, nonequilibrium candidate Monte Carlo (NCMC)[115] is used to increase the acceptance rates of GCMC moves, by allowing the system to relax and adjust in response to the proposed insertion or deletion of a water molecule. In this work, the use of NCMC to enhance GCMC moves is referred to as grand canonical nonequilibrium candidate Monte Carlo (GCNCMC), and simulations which combine this with conventional MD are described as GCNCMC/MD simulations. Whilst this would be

expected to provide an increase in acceptance rate over unbiased GCMC (also referred to as instantaneous GCMC), this approach has a key theoretical advantage over the other efficiency-enhancing measures described above. This is because the relaxation component of GCNCMC offers the possibility of inserting and deleting waters via an 'induced fit' mechanism, where the insertion or deletion of a water site requires cooperativity from the environment. Such moves would almost never be observed using the various biasing approaches, as these would require the spontaneous formation of a cavity, followed by the subsequent and rapid insertion of a water site — or conversely, for a deletion move, this might require the collapse of a cavity containing a water. However, GCNCMC would allow synergistic effects between water insertion/deletion and the rest of the system to occur. Further, owing to the increased computational cost of GCNCMC, any boosts in acceptance rate observed are weighed against the time taken to execute each move, in order to assess the overall benefit.

5.2 Acceptance Ratio Derivation

The derivation for a GCNCMC insertion move is demonstrated here, by combining the derivation of instantaneous GCMC (section 2.6.3.1) with the generalised NCMC acceptance ratio:[115]

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = \frac{P(\tilde{\Lambda}_p|\tilde{x}_T)}{P(\Lambda_p|x_0)} \frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)} \frac{\pi(\tilde{x}_T)}{\pi(x_0)} e^{-\Delta\mathcal{S}(X|\Lambda_p)} \tag{2.112 revisited}$$

where $P(\Lambda_p|x_0)$ is the probability of applying protocol Λ_p to x_0, $\alpha(X|\Lambda_p)$ is the total probability of applying each of the perturbations of the forward move, $\pi(x_0)$ is the equilibrium probability of x_0, and $\Delta\mathcal{S}(X|\Lambda_p)$ is the conditional path action difference. Here, we again make use of the treatment of a grand canonical simulation as a large, canonical simulation which includes the ideal gas reservoir. The equilibrium probability for a microstate with N particles in the system and $M - N$ particles in the ideal gas is given by:

$$\pi_{MVT}(\mathbf{r}^N, \mathbf{r}^{M-N}, \mathbf{p}^M) = Q_{MVT}^{-1} h^{-3M} e^{-\beta E(\mathbf{r}^N, \mathbf{r}^{M-N}, \mathbf{p}^M)} \mathrm{d}\mathbf{r}^M \mathrm{d}\mathbf{p}^M \tag{5.1}$$

where:

$$E(\mathbf{r}^N, \mathbf{r}^{M-N}, \mathbf{p}^M) = U(\mathbf{r}^N) + \sum_{i=1}^{M} \frac{|\mathbf{p_i}|^2}{2m} \tag{5.2}$$

Note that the above is written in terms of the total energy, which has no dependence on the positions of the particles in the ideal gas — the momenta are not separated, as the total energy is not affected by whether a particular momentum value comes from the ideal gas or the system.

The forward protocol for an insertion move first involves translating a particle from the ideal gas to a random location within the system of interest, and then deterministically increasing the interactions of this particle with the surroundings by scaling λ from 0 to 1, interspersed with relaxation steps. The reverse protocol would involve gradually decreasing the interactions of the particle by scaling λ from 1 to 0, and then translating it to a random location in the ideal gas. As the translations between the ideal gas and the system are not actually simulated in practice, these probabilities are absorbed into the probabilities of selecting the forward and reverse protocols. All other perturbation kernels are deterministic, as they involve changing the value of λ between two predetermined values (depending on the direction of the perturbation and the point within the protocol), therefore $\alpha(X|\Lambda_p) = \alpha(\tilde{X}|\tilde{\Lambda}_p)$. Given that insertions and deletions are attempted with equal probability, the probabilities of selecting the forward and reverse protocols are therefore given as:

$$P(\Lambda_p|x_0) = \frac{1}{2}\frac{1}{M-N}\frac{\mathrm{d}\mathbf{r}}{V_{sys}} \tag{5.3}$$

$$P(\tilde{\Lambda}_p|\tilde{x}_T) = \frac{1}{2}\frac{1}{N+1}\frac{\mathrm{d}\mathbf{r}}{V_{ideal}} \tag{5.4}$$

Substituting all of the above into Eq. 2.112 (and using the rearrangements employed in section 2.6.3.1), we arrive at the following, still somewhat generalised, acceptance ratio:

$$\begin{aligned}\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} &= \frac{P(\tilde{\Lambda}_p|\tilde{x}_T)}{P(\Lambda_p|x_0)}\frac{\alpha(\tilde{X}|\tilde{\Lambda}_p)}{\alpha(X|\Lambda_p)}\frac{\pi(\tilde{x}_T)}{\pi(x_0)}e^{-\Delta\mathcal{S}(X|\Lambda_p)} \\ &= \frac{M-N}{V_{ideal}}\frac{V_{sys}}{N+1}e^{-\Delta\mathcal{S}(X|\Lambda_p)}e^{-\beta\Delta E(X|\Lambda_p)} \\ &= \frac{1}{N+1}\frac{V_{sys}}{\Lambda^3}e^{\beta\mu}e^{-\Delta\mathcal{S}(X|\Lambda_p)}e^{-\beta\Delta E(X|\Lambda_p)} \\ &= \frac{1}{N+1}e^{B}e^{-\Delta\mathcal{S}(X|\Lambda_p)}e^{-\beta\Delta E(X|\Lambda_p)}\end{aligned} \tag{5.5}$$

where some of the steps demonstrated in section 2.6.3.1 have been omitted to avoid repetition.

When the propagation kernels preserve the equilibrium distribution, the conditional path action difference — where the conditional path action for the forward move is taken as the negative logarithm of the cumulative probability of all propagation steps,[115,174] and vice versa for the reverse move — can be written in terms of the heat change associated with the forward protocol:[115]

$$\Delta\mathcal{S}(X|\Lambda_p) = -\beta Q(X|\Lambda_p) \tag{5.6}$$

As implemented in *grand*, all relaxation/propagation steps are carried out using the

BAOAB Langevin integrator, which has been empirically observed to sample the equilibrium distribution very well,[158,159] so we can use this expression for the conditional path action difference. The change in total energy of the system can be decomposed into the heat and work of the forward path:

$$\Delta E(X|\Lambda_p) = W(X|\Lambda_p) + Q(X|\Lambda_p) \tag{5.7}$$

Which therefore allows the acceptance ratio to be somewhat simplified:

$$\begin{aligned}\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} &= \frac{1}{N+1} e^{B} e^{-\Delta\mathcal{S}(X|\Lambda_p)} e^{-\beta\Delta E(X|\Lambda_p)} \\ &= \frac{1}{N+1} e^{B} e^{\beta Q(X|\Lambda_p)} e^{-\beta\left(W(X|\Lambda_p)+Q(X|\Lambda_p)\right)} \\ &= \frac{1}{N+1} e^{B} e^{-\beta W(X|\Lambda_p)}\end{aligned} \tag{5.8}$$

There are two contributions to the work done during a nonequilibrium protocol: the work done on the system by each of the perturbation steps, known as the protocol work, W_p; and the work done on the system through integration error, known as the shadow work, W_s.[240] As numerical integration methods employ finite timesteps, they do not strictly obey the equations of motion for the system, and as such the total energy of a simulation is not exactly conserved during deterministic integration steps. These steps conserve the energy of a shadow Hamiltonian, which closely resembles the 'true' Hamiltonian of the simulation when the timestep is small. The change in system energy introduced by this difference is known as the shadow work, which can therefore be considered as a measure of the error introduced to the sampled distribution by the choice of integrator (and timestep).[240] The BAOAB integrator used here has been observed to preserve the equilibrium distribution very well,[158,159] so the shadow work can be assumed to be negligible — this assumption is also made in other NCMC-based methods.[164,168,169] The total work can therefore be well approximated by the protocol work:

$$\begin{aligned}W(X|\Lambda_p) &= W_p(X|\Lambda_p) + W_s(X|\Lambda_p) \\ &\approx \sum_{t=1}^{T} \left[U(x_t^*) - U(x_{t-1})\right]\end{aligned} \tag{5.9}$$

Leading to the following acceptance ratio for a GCNCMC insertion move:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = \frac{1}{N+1} e^{B} e^{-\beta W_p(X|\Lambda_p)} \tag{5.10}$$

An equivalent derivation can be performed to show that the acceptance ratio for a GCNCMC deletion move is:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = Ne^{-B}e^{-\beta W_p(X|\Lambda_p)} \tag{5.11}$$

A comparison between Eqs. 5.10 & 5.11 with Eqs. 2.124 & 2.125 shows that the only difference made by the inclusion of NCMC is the replacement of the potential energy change with the protocol work:

$$\frac{A(\mathbf{r}^{N+1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} = \frac{1}{N+1}e^{B}e^{-\beta\Delta U} \tag{2.124 revisited}$$

$$\frac{A(\mathbf{r}^{N-1}|\mathbf{r}^N)}{A(\mathbf{r}^N|\mathbf{r}^{N+1})} = Ne^{-B}e^{-\beta\Delta U} \tag{2.125 revisited}$$

When these moves are applied in conjunction with a GCMC sphere, it is possible that waters diffuse in or out of the sphere over the course of the move. Therefore, when using a GCMC sphere, Eqs. 5.10 & 5.11 are replaced with Eqs. 5.12 & 5.13, respectively:

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = \frac{1}{N_T}e^{B}e^{-\beta W_p(X|\Lambda_p)} \tag{5.12}$$

$$\frac{A(X|\Lambda_p)}{A(\tilde{X}|\tilde{\Lambda}_p)} = N_0\, e^{-B}e^{-\beta W_p(X|\Lambda_p)} \tag{5.13}$$

where N_0 is the number of waters in the sphere at the beginning of the move, and N_T is the corresponding number at the end of the move. If, at the end of the move, the water which is subject to the nonequilibrium procedure lies outside the sphere, this move must be automatically rejected, as the reverse protocol cannot be proposed by sampling waters within the sphere.

It should be noted that two additional parameters are introduced for GCNCMC moves: the number of perturbation kernels, n_{pert}, and the number of timesteps per propagation kernel, n_{prop}. Given that each protocol begins and ends with propagation (to ensure symmetry of the forward and reverse protocols), the length of each nonequilibrium protocol (known as the switching time, τ) can be related to these parameters:

$$\tau = \left(n_{pert} + 1\right) n_{prop}\delta t \tag{5.14}$$

The length of the protocol should have no impact on the accuracy, but is expected to affect the acceptance rate, and therefore the rate of convergence of results. The effect of different switching times is investigated in this chapter.

5.3 Simulation Details

The TIP3P water model[182] was used in all simulations, with a real-space interaction cutoff of 12 Å (interactions were switched between 10 and 12 Å), with PME used to calculate the effect of long-range electrostatic interactions.[142] Simulations were run at a temperature of 298 K, using the BAOAB Langevin integrator[158] ($\gamma = 1$ ps^{-1}, $\delta t = 2$ fs). All water molecules were constrained using the SETTLE algorithm.[187] Simulations were carried out using versions 7.3.1 and 1.1.0 of OpenMM[114,179] and *grand*,[113] respectively.

5.3.1 Effects of Nonequilibrium Sampling on Performance

As any difference in performance observed between GCNCMC and instantaneous GCMC was expected to be dependent on the protocol used for the GCNCMC moves, a series of protocols were tested, as described here. GCNCMC simulations were carried out using a range of switching times, from $\tau = 1$ ps to 15 ps, using n_{prop} values of 1, 5, 10, 50 and 100, in order to investigate the impact of this parameter — the values of n_{pert} were determined from the switching time and value of n_{prop}, according to Eq. 5.14. All switching times tested are given in Table 5.1, along with the corresponding values of n_{pert} for each value of n_{prop}. In each case, only GCNCMC moves were carried out, with no conventional MD sampling. For comparison, simulations were also carried out using instantaneous GCMC. It should be noted that these simulations are not strictly a fair representation of how GCMC would normally be carried out, but the absence of conventional MD sampling allows a more direct comparison between the two move types, in order to ascertain any benefit offered by the nonequilibrium aspect. All simulations were run for a 12 hour wall time limit, for three independent repeats, starting from a pre-equilibrated box of 2094 water molecules.

5.3.2 Bulk Water Density

To test that the GCNCMC implementation samples the correct distribution — i.e. that the inclusion of nonequilibrium sampling has been correctly accounted for in the acceptance criteria — another test was carried out on a bulk water system. These simulations were carried out on a pre-equilibrated box of 500 water molecules. For reference, constant pressure MD simulations were run for 100 ns each, with simulation frames written out every 50 ps — the pressure was maintained at 1 bar, using a Monte Carlo barostat, with volume changes attempted every 25 timesteps. The average volume observed over the three independent NPT simulations was used for the GCNCMC/MD simulations, which were carried out using a switching time of 7 ps ($n_{prop} = 50$, $n_{pert} = 69$) — this protocol was chosen as it appears to be most efficient for bulk water (Fig. 5.3).

	n_{pert}				
τ / ps	$n_{prop} = 1$	$n_{prop} = 5$	$n_{prop} = 10$	$n_{prop} = 50$	$n_{prop} = 100$
1	499	99	49	9	4
2	999	199	99	19	9
3	1499	299	149	29	14
4	1999	399	199	39	19
5	2499	499	249	49	24
6	2999	599	299	59	29
7	3499	699	349	69	34
8	3999	799	399	79	39
9	4499	899	449	89	44
10	4999	999	499	99	49
11	5499	1099	549	109	54
12	5999	1199	599	119	59
13	6499	1299	649	129	64
14	6999	1399	699	139	69
15	7499	1499	749	149	74

TABLE 5.1: Parameters defining each of the GCNCMC protocols tested. For each value of the switching time, τ, and n_{prop} tested, the corresponding value of n_{pert} is given. These values are all related via Eq. 5.14.

Three independent repeats were run in iterations of one GCNCMC move followed by 3 ps of MD, for 150,000 iterations, with simulation frames saved every 50 iterations.

5.4 Results

5.4.1 Effects of Nonequilibrium Sampling on Performance

The acceptance rates observed for each of the different GCNCMC protocols tested are plotted against switching time in Fig. 5.1. All of these protocols offer acceptance rates better than that of 0.0253 $\pm$ 0.0005 % observed using instantaneous GCMC. First, it seems that the acceptance rate is largely independent of the value of n_{prop} when n_{prop} is fairly small (10 or below, here). Larger values of n_{prop} begin to decrease the acceptance rate, as the correspondingly lower values of n_{pert} (see Table 5.1) result in larger jumps in λ for each perturbation step — this is especially apparent for short switching times, where the acceptance rates for $n_{prop} = 100$ differ significantly from the other values. Interestingly, the acceptance rate appears to increase approximately linearly with the switching time over the tested range of 1 ps $\leq \tau \leq$ 15 ps. For switching times of 15 ps, acceptance rates of around 30 % were observed — an improvement of three orders of magnitude over instantaneous GCMC. Further, the data collected suggest that the acceptance rate could be increased even further with longer switching times, but this would likely not be an efficient use of wall time, as discussed below. For comparison, the combination of cavity and orientation biasing by Woo *et al.* yielded acceptance rates

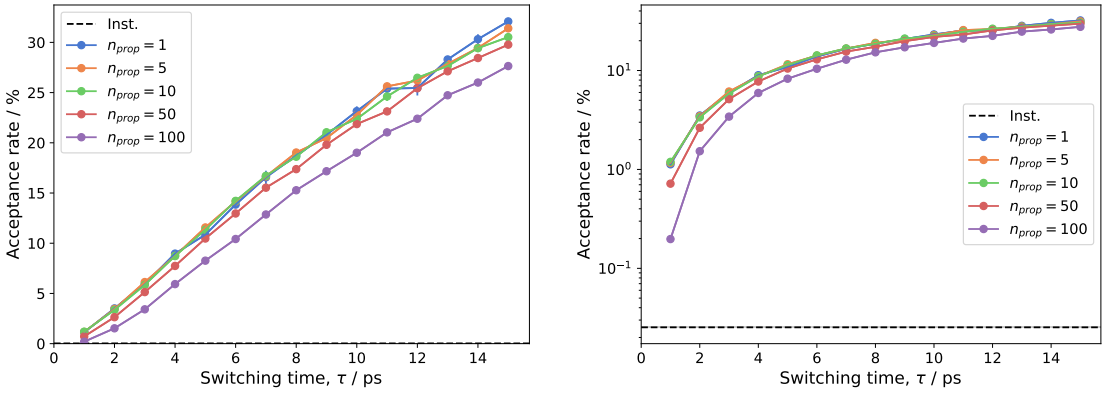


FIGURE 5.1: Acceptance rates observed for the different GCNCMC protocols tested in this chapter, with the acceptance rates observed for instantaneous GCMC included for reference. The data on the right is identical to that on the left, but plotted on a logarithmic scale, in order to better visualise the comparison with the instantaneous acceptance rate. Error bars represent the standard error over the three repeats (as does the shaded region about the dashed line) — note that the errors are very small.

of 0.81 - 0.85 %.[103]

The increase in acceptance rate for longer switching times can be explained in terms of the work distributions obtained from the GCNCMC protocols. Several work distributions for different switching times are plotted in Fig. 5.2, where the work values from rejected moves are also included. As can be seen, increasing the switching time causes the resulting work distribution to shift towards more negative values, as well as becoming narrower and more symmetric. In the limit of very long switching times, the work distributions would be expected to be normally distributed[150] about the excess chemical potential (or the negative value), with a very low variance. The narrowing of the work distributions is significantly more pronounced for the insertion moves than the deletions, where the former are still rather skewed even with a switching time of 15 ps.

Having established that GCNCMC moves can offer significantly higher acceptance rates than instantaneous GCMC, it is important to assess whether these simulations are more efficient, as the improvement may be offset by the increased computational cost of these moves. The relative efficiencies of the different protocols were calculated as the total number of moves accepted during the 12 hour wall time limit of the simulations, divided by the mean number of moves accepted in the same amount of time using instantaneous GCMC. These results are plotted for the different protocols in Fig. 5.3, where it can be seen that both the value of n_{prop} and the switching time have a significant effect on the efficiency. Larger values of n_{prop} increase the relative efficiency of the protocols, as the acceptance rates tend to differ little between different values of n_{prop}, but the wall time required to execute each GCNCMC move is reduced for larger values

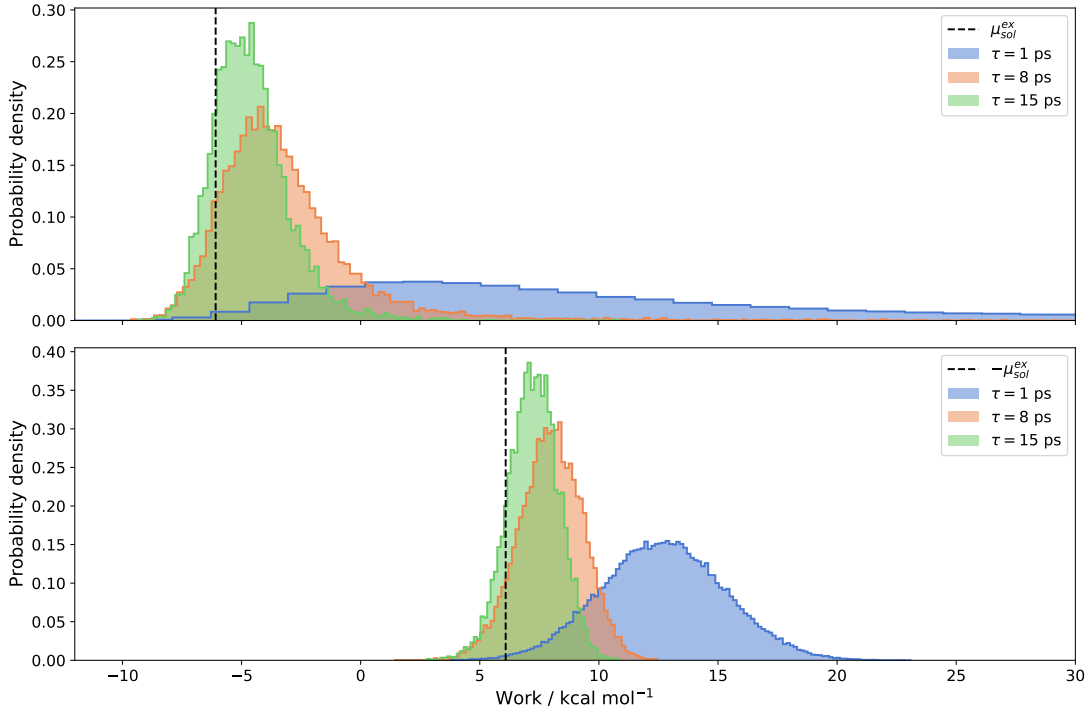


FIGURE 5.2: Work distributions of GCNCMC moves using different switching times. The upper plot corresponds to insertion moves, and the lower plot to deletions, with the positive and negative values of the excess chemical potential included as dashed lines, for reference. The work values were taken from simulations using $n_{prop} = 10$ as a representative example.

of this parameter. It is thought that, for larger relaxation times between perturbations, the simulation becomes slightly more MD-like, which means that more of the simulation time is spent on the native, GPU-optimised MD functionality in OpenMM, relative to the slower perturbation steps carried out in *grand* (executed on the CPU). This fact makes it likely that the exact performance difference between different NCMC protocols is therefore likely to be hardware-dependent. However, the relationship described between n_{prop} and efficiency breaks down when increasing n_{prop} from 50 to 100, as the drop in acceptance rate (Fig. 5.1) is not compensated for by the increase in speed. Of the n_{prop} values tested in this work, $n_{prop} = 50$ provides optimal results (provided that the switching time is not too small), and can be over five times more efficient than instantaneous GCMC. A positive result here is the fact that, for a given value of $n_{prop} < 100$, the relative efficiency does not appear to vary significantly for switching times between approximately 5 ps and 13 ps. This means that the optimal GCNCMC protocol need not be determined too precisely, and should therefore be suitable to simulate different systems with near-optimal efficiency.

An additional advantage of GCNCMC moves over the previously used instantaneous implementation is that the work values calculated during the nonequilibrium protocols can be used to calculate the free energy of inserting a water molecule into the system

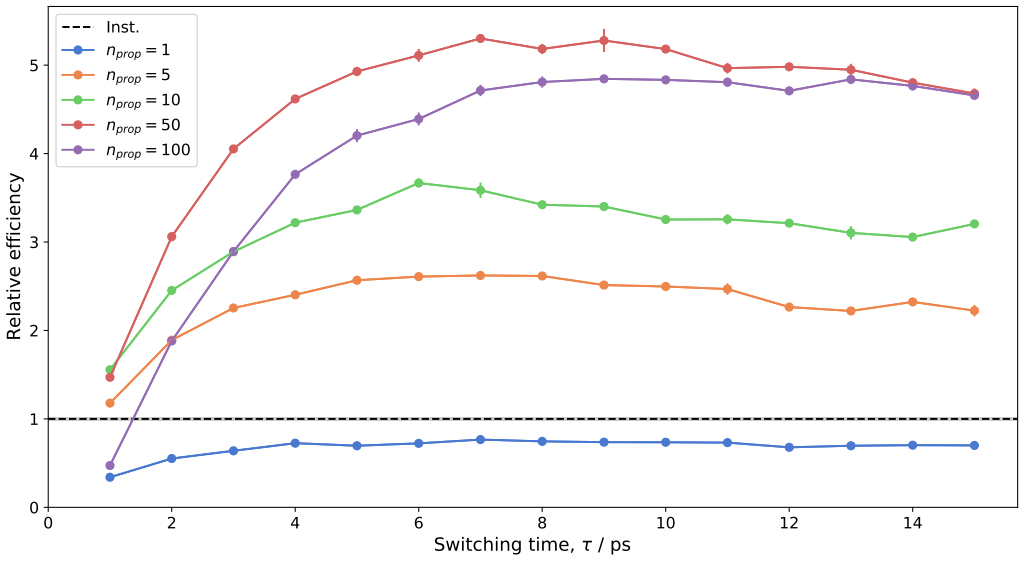


FIGURE 5.3: Efficiencies of the different GCNCMC protocols tested in this chapter, relative to instantaneous GCMC. The efficiencies are calculated as the total number of moves accepted during the same amount of wall time, relative to the mean number of moves accepted in the same amount of time using instantaneous GCMC. As in Fig. 5.1, error bars (and the shaded region about the dashed line) represent the standard error over the three repeats.

(section 2.4.3). This can also make use of the work values from rejected moves, meaning that the time invested in generating configurations which are not accepted into the ensemble is therefore not totally wasted. In order to assess this capability, the work values from each of the simulations were processed using BAR (as implemented in *pymbar*[191]) to calculate the hydration free energy of water (where the insertion works are taken as the forward values, and the deletion works are taken as the reverse values), with the results shown in Fig. 5.4. As can be seen, for longer switching times, the estimated free energies begin to approach the free energy value calculated in chapter 3, although even for the longest switching times employed, the free energies are still systematically overestimated. This is because, as shown in Fig. 5.2, even with a switching time of 15 ps, the work distributions are still distinctly offset from the value calculated (at significantly greater computational expense) using an equilibrium free energy method. Better nonequilibrium free estimates would require either longer switching times, or significantly more samples. However, it should be recalled that the primary aim of these simulations is to sample the grand canonical ensemble, so these free energy data should be considered a bonus, which can be calculated at little to no additional computational cost. Note that here, all work values were combined for each simulation type, as the bulk water system simulated is homogeneous. If a heterogeneous system (such as a protein system) were to be analysed this way, then the work values would have to be separated appropriately to capture the spatial dependence of the insertion free energy — more samples would therefore be needed to give converged free energies

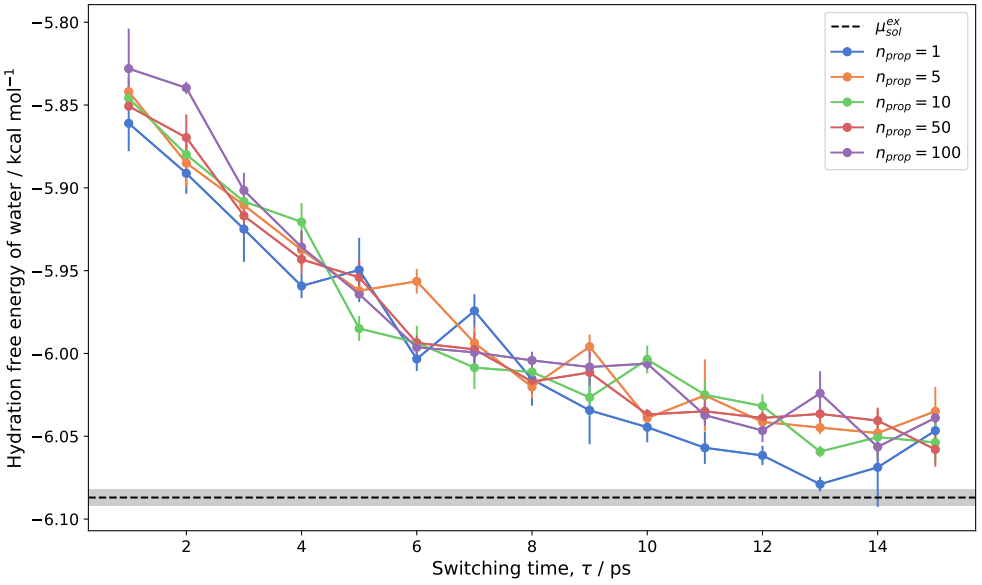


FIGURE 5.4: Nonequilibrium estimates of the free energy to insert a water molecule in water, using the work values obtained from the GCNCMC moves carried out. Each data point is the mean free energy calculated over the three repeats, and the error bars represent the standard error in the mean.

across the space sampled.

5.4.2 Bulk Water Density

As previously mentioned, in order to verify that the inclusion of nonequilibrium switching was correctly accounted for in the acceptance criteria, an analysis of bulk water was carried out, similar to that presented in chapter 3. Histograms of the mass density of the simulated water box are plotted in Fig. 5.5 for both the NPT and GCNCMC/MD simulations. As can be seen, the agreement between the two distributions is qualitatively excellent, which would appear to verify that the acceptance criteria derived for GCNCMC moves in this work are correct. As discussed in chapter 3, bulk water is a much larger system than would typically be simulated with grand canonical methods, but serves as a useful test, because the scale of the system would likely exacerbate any errors in the underlying theory or implementation — this is therefore a reassuring result. It is interesting to note that the distributions shown in Fig. 5.5 are somewhat broader than those in Fig. 3.1a, owing to the smaller volume of the simulations described in this chapter — a similar effect was observed by Ross *et al.*, where fluctuations in salt concentrations were noted to be larger for smaller simulation volumes.[168]

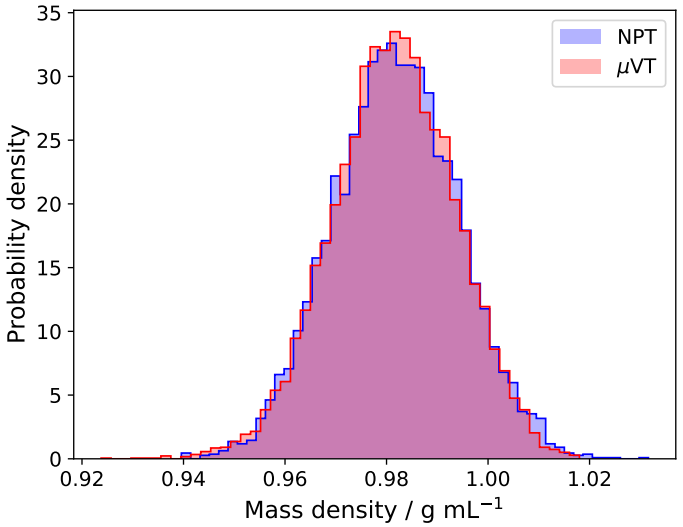


FIGURE 5.5: Distributions of the bulk water density as sampled using constant pressure MD and GCNCMC/MD ($\tau = 7$ ps).

5.5 Summary

This chapter has shown that nonequilibrium candidate Monte Carlo (NCMC) can be applied to drastically improve the acceptance rates and efficiencies of GCMC simulations (yielding GCNCMC), by allowing the simulated system to relax in response to a proposed insertion or deletion move. The derivation is presented for these acceptance criteria, which transpire to be almost identical to those used for instantaneous GCMC simulations, except that the instantaneous potential energy change is replaced with a nonequilibrium work. The results collected in this work show that the acceptance rates for GCMC sampling of water molecules can be increased by three orders of magnitude, when using GCNCMC. Interestingly, the acceptance rate did not plateau over the range of switching times tested, and can therefore likely be increased even further (Fig. 5.1) — however, this will likely be less efficient, as the results obtained indicate that the optimal balance between acceptance rate and computational cost lies somewhere between 5 ps and 13 ps (Fig. 5.3). When accounting for the increased computational cost of GCNCMC moves over instantaneous GCMC, it was found that the efficiency can be increased by over a factor of 5 (Fig. 5.3), although it should be noted that the efficiency differences between different NCMC protocols are likely hardware-dependent. The correctness of the GCNCMC implementation in *grand* was demonstrated by the excellent agreement between the bulk water density distributions observed from GCNCMC/MD simulations and those from constant pressure simulations under the same conditions (Fig. 5.5).

An additional advantage of GCNCMC is that the work values calculated from the nonequilibrium protocols can be used to estimate free energy data from the simulation. Whilst the relatively short nature of the nonequilibrium protocols used for water sampling means that the free energy estimates are not accurate (Fig. 5.4), it should be noted that these free energies are provided as a secondary benefit of using NCMC to enhance the sampling of water molecules, and can be calculated at an additional computational cost which is negligible, relative to that of the simulation itself. Whilst the free energies may not be fully converged, they can still be used to provide some degree of additional insight into the system.

A surprising observation made in related work (*not presented here — the observations discussed in this paragraph were made by Oliver Melling (OJM), using results collected by OJM*) is that the use of GCNCMC sampling of waters in protein binding sites not only improves the sampling of the waters, but can also improve the sampling of other degrees of freedom. During GCNCMC/MD simulations of major urinary protein I (MUP-I), a broader distribution of waters are observed in the protein binding site than observed using instantaneous GCMC/MD. It transpires that these additional microstates are a result of greater motion of the ligand within the binding site, owing to the fact that some insertion and deletion moves appear to require ligand motion in order to stabilise the insertion/deletion. Such stabilisations cannot occur during instantaneous GCMC/MD, and analogous move proposals are therefore overwhelmingly rejected. This surprising observation indicates that GCNCMC/MD simulations may be even more beneficial to the sampling of protein-ligand binding sites than initially anticipated.

In summary, the results presented in this chapter show that NCMC appears to offer a significant and efficient improvement in the very low acceptance rates typically observed using unbiased GCMC. The fact that the acceptance rates can be increased so significantly may mean that NCMC can be applied to GCMC sampling of molecules larger than water, for which instantaneous insertion and deletion moves will become vanishingly unlikely — the smoother nature of GCNCMC insertions and deletions would be expected to be especially beneficial in such cases, where suitable cavities are unlikely to arise spontaneously. This prospect is investigated in the following chapter.

Chapter 6

Grand Canonical Sampling of Small Organic Molecules

6.1 Introduction

6.1.1 Fragment-Based Drug Design

Fragment-based drug design (FBDD) is now a widely used strategy for initial hit identification.[241] Unlike conventional screening, FBDD focuses on the screening of small, weakly binding molecules (fragments), which can then be elaborated upon to design drug-like molecules.[242] Screened fragments typically obey the 'rule of three':[243,244] molecular weight less than, or equal to, 300 Da; fewer than, or equal to, 3 hydrogen bond donors; fewer than, or equal to, 3 hydrogen bond acceptors; and a logP less than, or equal to, 3. The primary benefit of this approach is that a broader range of chemical space can be covered with fewer compounds, compared to screening of larger compounds.[242] However, the reduced size and complexity of fragment-like molecules means that they typically bind weakly to their target, and the lower limit of useful fragment size is therefore determined by the sensitivity with which hits can be identified. Another benefit of FBDD is that the small size of the fragments allows them to bind to protein regions which may not be identified with larger compounds.[245]

Alongside docking — which is widely used in the pharmaceutical industry for the prediction of protein-ligand binding modes[2] — a number of MD-based methods have been developed for the study of fragment binding,[246] some of which are briefly discussed here. The SILCS method involves simulating a protein surrounded by a concentrated fragment solution (with a repulsive potential employed between fragment molecules to prevent aggregation), from which fragment binding locations can be identified.[247]

MixMD is a similar method, which simulates a protein in a mixture of water and another solvent (such as ethanol or acetamide), from which the solvent interactions with the protein are used to identify interaction hotspots.[248] SWISH is a fragment sampling method, primarily intended for the identification of cryptic binding pockets (those which are only visible after ligand binding), which scales the interaction strength between waters and apolar protein atoms, causing more water to bind to the protein than usual.[249] This increased water binding opens apolar pockets on the protein, to which the ligands can then bind by displacing the waters. Multiple replicas of the system are simulated, each using different values of the scaling parameter, with replica exchange used to extract unbiased simulation trajectories.[249] However, these MD-based methodologies would be expected to suffer from the same kinetic limitations as previously discussed for MD simulation of buried water sites, particularly when the fragment binding location is deeply buried. A possible exception is perhaps the BLUES software package, which aims to enhance the sampling of transitions between fragment binding modes, with the first implementation using NCMC to attempt large rotations of the fragment about the centre of mass,[164] and a later version extending this methodology to attempt large dihedral rotations.[167] The use of BLUES for enhanced rotational sampling has been used to refine docked fragment structures, offering an improvement in the identification of crystallographically observed binding modes.[165] More recently, however, this package has been extended to sample transitions between distinct fragment binding sites using a Monte Carlo technique known as molecular darting, which can allow fragments to 'jump' between different regions of the protein.[250]

6.1.2 Application of GCMC to Fragment Binding

Throughout this thesis, grand canonical Monte Carlo (GCMC) has only been discussed in the context of determining the binding locations of water molecules within protein systems, but there is no conceptual reason preventing this method from being used to determine the binding locations of other small molecules, such as molecular fragments. However, it is expected that the acceptance rates would be vanishingly low, making this practically infeasible, given the very small acceptance rates which are typically observed for a molecule even as small as water (chapter 5). Clark *et al.* have employed GCMC sampling to determine the binding modes and free energies of rigid fragments to protein structures,[251,252] although they did not report the acceptance rates observed, a number of biasing techniques were necessary to favour more probable insertions.[251] A GCMC-like method has also been developed by Lakkaraju *et al.* for the sampling of fragment binding to a restrained protein structure.[253] However, it should be noted that this method is not strictly GCMC, as the chemical potential is not constant, instead, the excess chemical potential is allowed to vary such that the fragment concentration within the simulation volume fluctuates about some user-defined value.[253]

In chapter 5, it was shown that nonequilibrium candidate Monte Carlo (NCMC) can be used to increase the acceptance rates of GCMC moves by three orders of magnitude for water molecules. This observation raises the question as to whether NCMC could therefore be used to make grand canonical sampling of small organic molecules more accessible in condensed phases. If this were the case, then it could allow GCNCMC/MD simulations to be used to identify fragment binding locations within protein structures, which would be of significant interest in computer-aided drug design. Additionally, it is possible that GCNCMC moves may allow the detection of cryptic binding pockets by allowing protein rearrangements following the insertion of a fragment into a closed pocket.

It should be noted that the implementation of grand canonical sampling for non-water molecules is slightly more complex. Previously, only sampling of water molecules was considered, where the reference state is bulk water, which has a well defined standard state (pure water under standard conditions). When considering grand canonical sampling of non-water molecules, we consider the reference state to be a solution of the ligand (benzene is used in this chapter) in water. It is preferable to carry out GCNCMC sampling of water, as well as the ligand, and as such, this necessitates the use of two separate Adams values for the ligand and water, denoted B^L and B^W, respectively — that is, the chemical potentials of the both the ligand and the water are separate and constant.[254–257] Similarly, there is an Adams value for each species which will give a simulation in equilibrium with the reference solution:

$$B^L_{equil}(c_L) = \beta\mu^{ex}_L(c_L) + \ln\left(\frac{V_{GCMC}}{V_L(c_L)}\right) \tag{6.1}$$

where c_L is the ligand concentration, μ^{ex}_L is the excess chemical potential of the ligand and V_L is the average volume per ligand in the reference solution[178] — B^W_{equil} is similarly defined. Note that if the exact ligand concentration is known, then the average volume per ligand can be trivially calculated:

$$V_L(c_L) = \frac{1}{N_A c_L} \tag{6.2}$$

where N_A is Avogadro's constant. The standard state for a mixture is more difficult to define than that of a pure substance, as the mixture can be defined in terms of the mole fraction, molarity or molality of the solute,[178] whereas for a single component system such as bulk water, the standard state is simply the pure substance under standard conditions. Even under standard conditions, the chemical potential of the reference solution is dependent on the composition of the mixture[178] (unlike the case of bulk water, where composition is not an issue). As one might wish to define a reference concentration which does not correspond to any definition of standard state solution, the average volume per ligand/water is not referred to as the standard state volume

of the species in this chapter, and the '$\circ$' superscript is therefore not used. Therefore, the grand canonical parameters are not only dependent on the ligand of interest, but also on the concentration of the reference solution, as is made explicit in the equations above.

In this chapter, the use of GCNCMC/MD simulations for fragment-like molecules is demonstrated using benzene as an example. First, the thermodynamic parameters required for grand canonical simulation are calculated for several concentrations of benzene solution. These parameters are then tested by running GCNCMC/MD simulations on a pure water box, in order to identify whether or not the desired concentration is reproduced by the choice of parameters. In order to test the ability of GCNCMC sampling of small molecules to identify fragment binding locations within a protein structure, TEM1 β-lactamase was chosen as a test system. This protein was chosen because it contains a cryptic binding pocket[259] for which benzene binding was observed using the SWISH method for cryptic pocket detection.[249,260] This system was used to test the ability of GCNCMC sampling to bind benzene to both pre-formed pockets and cryptic pockets. The former was tested by running simulations using restraints to hold the cryptic pocket open, and in the latter case, the *apo*-structure was used as a starting point, in which the cryptic pocket is not seen (Fig. 6.1).

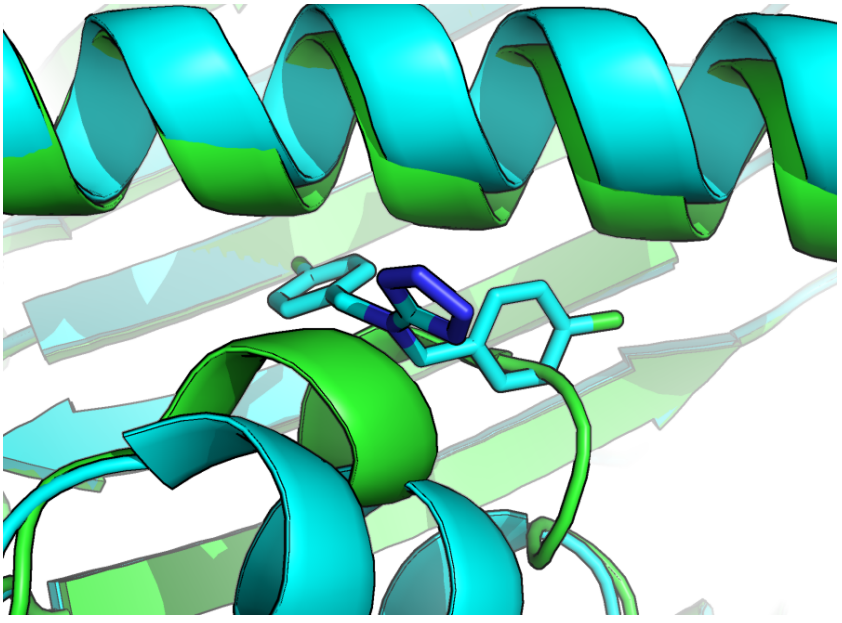


FIGURE 6.1: Location of the TEM1 β-lactamase cryptic pocket in the *apo*- and *holo*-structures. The *apo*-structure (PDB ID: 1JWP,[258] 1.75 Å) is shown in green and the *holo*-structure (PDB ID: 1PZO,[259] 1.90 Å) is shown in cyan, with the bound ligand. As can be seen, the opening of the pocket involves a significant backbone rearrangement.

6.2 Simulation Details

The following conditions were used for all simulations reported in this chapter. The AMBER ff14SB and TIP3P force fields were used to model the protein[181] and water,[182] respectively, and Joung-Cheatham parameters[183,184] were used for any ions present in the simulation. The general AMBER force field[120] (GAFF) was used with AM1-BCC charges,[122,123] (calculated using *antechamber*[230]) to represent benzene molecules. An interaction cutoff of 12 Å was used, with a switching function applied between 10 and 12 Å, and long-range electrostatic interactions were calculated using PME.[142] Simulations were carried out at 298 K, using the BAOAB Langevin integrator[158] to integrate the configurational degrees of freedom ($\gamma = 1$ ps^{-1}, $\delta t = 2$ fs). All bonds involving hydrogen atoms were constrained to their equilibrium values, using the SETTLE algorithm for water molecules,[187] and the SHAKE algorithm for non-water molecules.[185,186] Any NPT simulations were carried out at a pressure of 1 bar, using a Monte Carlo barostat, with volume changes attempted every 25 timesteps. Where GCNCMC moves were carried out, a switching time of 50 ps ($n_{prop} = 50$, $n_{pert} = 499$) was used for benzene, and $\tau = 10$ ps was used for water ($n_{prop} = 50$, $n_{pert} = 99$) — it should be noted that the GCNCMC protocol for benzene has not been optimised. All simulations here used version 7.3.1 of OpenMM,[114,179] and a development version of the *grand* module.[113]

6.2.1 Thermodynamic Parameters

As the grand canonical parameters required for both water and benzene are expected to be dependent on the composition of the reference solution, the excess chemical potential and average volume were calculated for both species at several concentrations of benzene. Concentrations of ∼0.1 M, ∼0.3 M and ∼0.5 M were selected, as this range was observed to provide well mixed solutions (at concentrations closer to 1.0 M, the benzene molecules were observed to aggregate). It should be noted that these concentrations are approximate, as it is very difficult to set an exact concentration for a finite-size simulation — these were prepared by creating systems containing 4, 12 and 20 benzene molecules, each with 2094 water molecules. The exact concentrations observed in each case are provided in Table 6.2 — for notational simplicity, these values are referred to approximately throughout this chapter.

The excess chemical potentials were calculated separately for both benzene and water. In each case, a non-interacting molecule was gradually coupled to the system over 30 equally-spaced λ values from 0 to 1, with ten independent repeats carried out in each case. For each repeat, 1000 samples were collected, with 10 ps of NPT simulation between samples. These samples were post-processed to remove correlated data,[190] and

the free energy was then extracted using the MBAR method.[149,191] The average molecular volume was calculated from a 50 ns simulation at the appropriate concentration. Every 5 ps, samples were collected for both benzene and water, in which the simulation volume was divided by the appropriate number of molecules in the simulation. In this case, five independent repeats were carried out.

6.2.2 Bulk Concentration

In order to test the quality of the thermodynamic parameters determined, the ability of grand canonical simulations to reproduce the bulk benzene concentration was tested, just as the bulk density of water was analysed in chapter 3. Here, GCNCMC/MD simulations were run, starting from a pre-equilibrated box of pure water, containing 2109 water molecules within a volume of (40.101 Å)^3, in order to test if the desired concentration would be reproduced. The system was simulated for 50,000 iterations of the following: one GCNCMC benzene move, three GCNCMC water moves, and 20 ps of conventional MD. Concentration samples were taken after every iteration, with the first 5000 samples discarded as equilibration. The GCNCMC moves sampled the entire system volume, according to the B_{equil} values given in Table 6.1 for each concentration, calculated using Eq. 6.1 with the parameters determined in this work (Table 6.2). Five independent repeats were carried out for each concentration.

For comparison, an additional set of simulations were carried out, under exactly the same conditions, except that every GCNCMC move on benzene was replaced with 5000 instantaneous GCMC moves. In this case, the starting structure for the simulation was an equilibrated benzene solution at approximately the desired concentration. This was run for 2000 iterations (giving a total of 10^7 GCMC moves on benzene). The purpose of this alternate set of simulations was not to test the GCMC methodology, but to obtain a quantitative estimate of how poor the acceptance rate would be for GCMC sampling of benzene, to which the GCNCMC acceptance rate could be compared.

Concentration / M	B^L_{equil}	B^W_{equil}
∼0.1	+0.497	−2.615
∼0.3	+1.459	−2.650
∼0.5	+1.850	−2.634

TABLE 6.1: B_{equil} values of benzene and water for the GCNCMC/MD simulations carried out to reproduce the bulk concentrations. These values were all calculated using Eq. 6.1 (for both the ligand and water) with the parameters given in Table 6.2.

6.2.3 Identification of Fragment Binding Sites

As previously mentioned, two separate sets of simulations were run for the TEM1 system — one with an open binding pocket, and one in which the cryptic pocket is closed. The initial structure of the closed simulation was taken as the *apo*-structure of TEM1 (PDB ID: 1JWP[258]). For the open simulation, the *holo*-structure, in which a ligand is bound to the cryptic pocket (PDB ID: 1PZO,[259] Fig. 6.1), was downloaded, and the ligand molecules were removed, giving a *pseudo-apo*-structure. In order to prevent the pocket from collapsing after the removal of the ligand, harmonic restraints were applied to all protein heavy atoms, restraining them to their initial positions, using a harmonic constant of 10 kcal mol^{-1} $Å^{-2}$.

For both systems, the following setup procedure was applied. The relevant crystal structure was downloaded from the Protein Data Bank (PDB),[31,32] and then hydrogen atoms were added using the *Modeller* tool found in OpenMM[114] — the N-terminus was capped with an acetyl group, and the C-terminus was negatively charged. The protein was solvated in a water box, which extended at least 8 Å from all protein atoms, with sodium ions added to neutralise the system charge. The system was minimised and then subjected to 1 ns of constant pressure equilibration. In each case, the GCMC sphere was centred on the C_α atom of Thr71, with a radius of 30 Å to cover the entire protein, and the benzene concentration was set to ∼0.5 M, using the parameters determined in this work ($B^L_{equil} = +2.412$, $B^W_{equil} = -2.072$). As before, the simulations were carried out in iterations of one GCNCMC benzene move, followed by three GCNCMC moves on water, and then 20 ps of conventional MD. Each simulation was run for 5000 iterations, with simulation frames saved after each iteration — three independent simulations were run for each initial structure.

6.3 Results

6.3.1 Thermodynamic Parameters

The values of the excess chemical potential and average molecular volume calculated for both benzene and water at each of the tested concentrations are given in Table 6.2. As can be seen, these properties are very concentration-dependent, except for the excess chemical potential of water, which appears relatively insensitive to the benzene concentration — likely because the concentrations of benzene are small, relative to those of water. Where these parameters are used in GCNCMC/MD simulations reported in this chapter, the value of each parameter is taken as the mean value, rounded to the same precision as the first significant figure in the standard error, e.g. using the results in Table 6.2, the value of the excess chemical potential of benzene at a concentration of

Concentration / M	μ_L^{ex} / kcal mol^{-1}	V_L / Å^3	μ_W^{ex} / kcal mol^{-1}	V_W / Å^3
0.103564	−0.528	16,033.8	−6.080	30.6281
(0.000001)	(0.015)	(0.2)	(0.010)	(0.0004)
0.305082	−0.600	5442.9	−6.088	31.1915
(0.000007)	(0.012)	(0.1)	(0.013)	(0.0007)
0.499386	−0.656	3325.2	−6.070	31.7589
(0.000022)	(0.026)	(0.1)	(0.010)	(0.0014)

TABLE 6.2: Values of the thermodynamic parameters calculated for both benzene and water for different concentrations of benzene, along with the exact benzene concentrations observed for each simulation configuration. For each parameter, the value quoted is the mean value from the independent repeats, and the value in parentheses is the standard error in the mean. Note that the uncertainty in the concentration arises from fluctuations in the simulation volume at constant pressure.

$\sim$0.1 M is taken as -0.53 kcal mol^{-1}. Having collected these data, the parameters for benzene concentrations between 0.1 and 0.5 M could be obtained by interpolation from the values in Table 6.2, making it easier to simulate alternative concentrations, without requiring these costly parameterisation simulations.

6.3.2 Bulk Concentration

The acceptance rates observed for the GCNCMC moves on benzene, at concentrations of $\sim$0.1 M, $\sim$0.3 M and $\sim$0.5 M were 24.8 $\pm$ 0.1 % , 25.1 $\pm$ 0.1 % and 24.5 $\pm$ 0.0 %, respectively — interestingly, the acceptance rates appear to be relatively independent of the benzene concentration. In order to quantify the improvement in the acceptance rate offered by NCMC, an additional set of simulations were run using instantaneous GCMC moves to insert and delete benzene molecules, as described previously. For these simulations, the acceptance rates were (1.0 $\pm$ 0.5) $\times$ 10^{-5} %, (1.2 $\pm$ 0.4) $\times$ 10^{-5} % and (2.2 $\pm$ 0.9) $\times$ 10^{-5} % , respectively, where such a low acceptance rate would make instantaneous GCMC/MD simulation of benzene infeasible. This therefore indicates that the use of NCMC for benzene insertions and deletions improves the acceptance rates by six orders of magnitude — in chapter 5, the improvement for water was three orders of magnitude, indicating that the benefit of NCMC becomes even more significant for lower probability instantaneous move proposals. This therefore highlights the power of NCMC move proposals, via the relaxation stages in the nonequilibrium protocol.

The concentration distributions observed for the GCNCMC/MD simulations carried out on bulk solvent, using the different sets of parameters are plotted in Fig. 6.2. In general, it appears that the concentration distributions broadly represent the intended macroscopic concentrations at which the parameters were determined. However, some

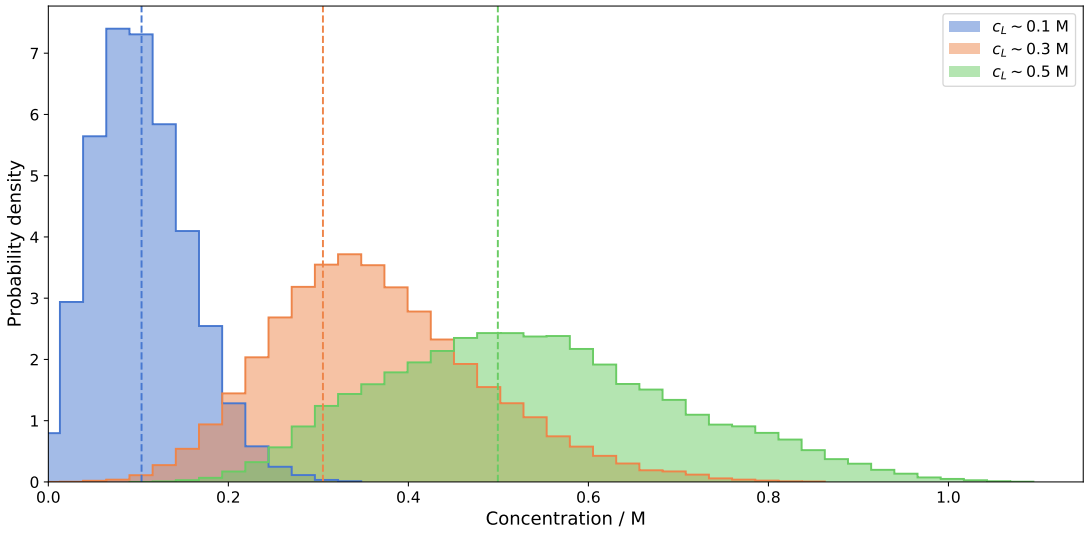


FIGURE 6.2: Distributions of the benzene concentrations observed for the different GCNCMC/MD simulations. The bin widths of the histograms correspond to the difference in concentration caused by changing the number of benzene molecules by one, at the simulated volume. The dashed vertical lines indicate the intended macroscopic concentration for each set of simulations.

discrepancies are noted — particularly for the ∼0.3 M data (discussed below) — but these data generally indicate that the concentration of the ligand within the solution can be controlled via GCNCMC/MD. It is also interesting to note that the distributions here become notably broader for higher concentrations, likely because at lower concentrations, a difference of one benzene molecule corresponds to a larger difference in concentration, relative to the equilibrium concentration, and is therefore less likely to be accepted.

It should be noted that the distribution of the ∼0.3 M data is somewhat shifted from the desired concentration. It is possible that this data is skewed because the parameters are not as well calibrated as one might like — notably, the excess chemical potential of water is slightly lower than those of the other two concentrations (Table 6.2). As shown in chapter 3, even subtle differences in this parameter can cause noticeable differences in the density of bulk water. If the value of μ_W^{ex} is indeed too low the for ∼0.3 M simulations, the slight drop in the density of water could create space for additional benzene molecules to be inserted into the solution, causing the concentration distribution to be shifted. To test this hypothesis, a set of GCNCMC/MD simulations identical to those at ∼0.3 M were carried out, except that the excess chemical potential of water was set to -6.075 kcal mol^{-1} (halfway between the analogous values for ∼0.1 M and ∼0.5 M), giving $B^W_{equil} = -2.625$. The resulting concentration distribution for the two values of μ_W^{ex} are shown in Fig. 6.3. Interestingly, even a small difference of 0.015 kcal mol^{-1} in the excess chemical potential of water causes a notable shift in the concentration distribution of benzene. These data therefore underscore the previously discussed importance

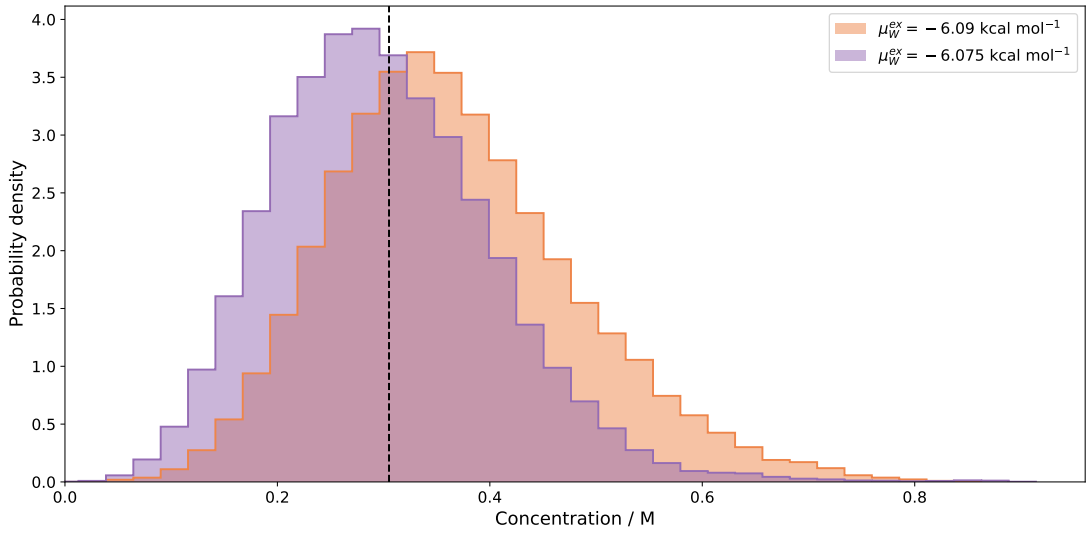


FIGURE 6.3: Distributions of the benzene concentrations observed for a macroscopic concentration of ∼0.3 M, using two different values of the excess chemical potential of water, μ_W^{ex}. The value of −6.09 kcal mol^{-1} was taken by rounding the parameterised value (Table 6.2), and the value of −6.075 kcal mol^{-1} was taken as the midpoint of the values determined at concentrations of ∼0.1 M and ∼0.5 M. The dashed vertical line indicates the intended macroscopic concentration.

of using well calibrated thermodynamic parameters (chapter 3). However, the aim of these simulations was to verify that the ligand concentration is adequately sampled by GCNCMC/MD simulation, which appears to be the case, and in many cases, reproducing the *exact* concentration of a ligand is likely to be of significantly less interest than where the ligands bind. Nonetheless, further work is needed to validate this method, preferably involving multiple solutes at range of concentrations.

6.3.3 Identification of Fragment Binding Sites

First, the simulations in which the cryptic pocket of TEM1 was held open by restraining the protein heavy atoms were analysed. It was observed that binding of benzene to the pocket is rapidly observed in all three repeats, with the first insertion of a benzene molecule into the pocket observed within 150 cycles in all three cases. As the simulation progresses, the pocket is eventually filled with 3-5 benzene molecules, with a representative example of 4 bound benzene molecules shown in Fig. 6.4. This therefore indicates that, despite the GCMC sphere covering the whole protein (not just the binding site), the GCNCMC moves allow the molecule insertions to rapidly find preformed cavities. However, whilst this is a reassuring result, in many cases, the pocket of interest will not be neatly pre-formed, and some degree of protein reorganisation will be necessary upon fragment binding.

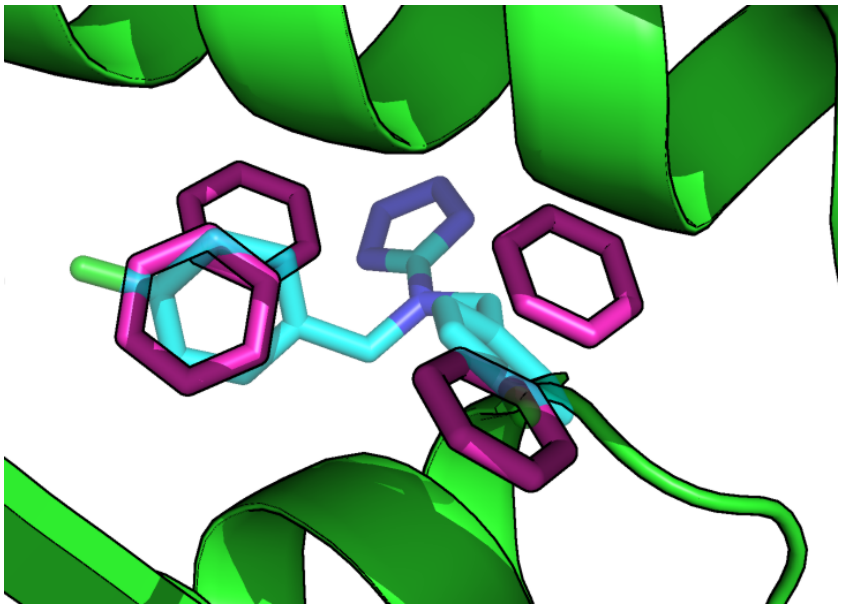


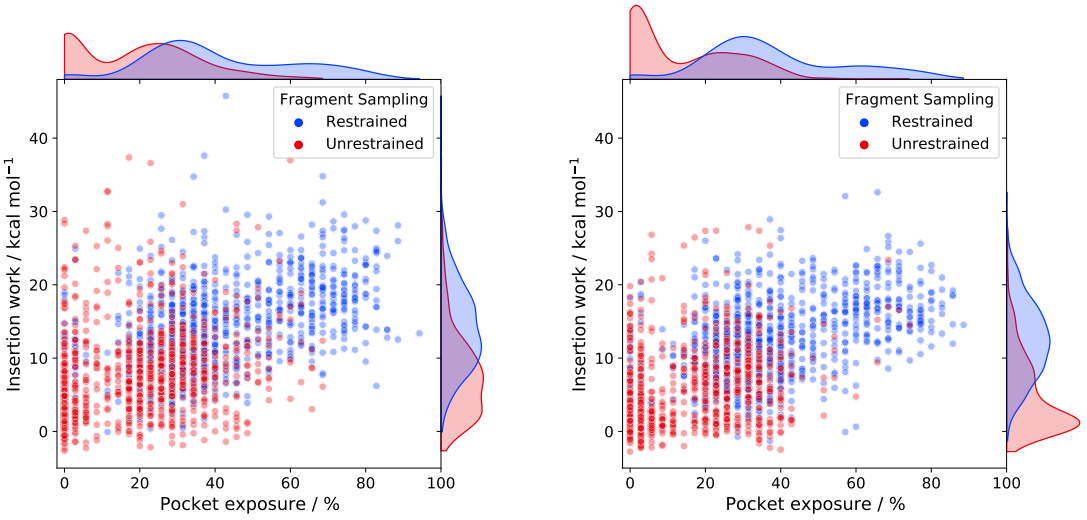
FIGURE 6.4: Representative frame from the GCNCMC/MD simulations, showing the binding of benzene to the TEM1 pocket, where the protein is restrained to maintain the pocket open. The benzene molecules are shown in pink and the crystallographic binding mode of the ligand is shown in transparent cyan.

The unrestrained simulations of *apo*-TEM1 serve as a difficult test case for the use of GCNCMC moves to insert fragments into closed protein binding sites, owing to the significant backbone rearrangement which is required to open the cryptic pocket (Fig. 6.1). Indeed, no binding of benzene to the cryptic pocket was observed during any of the three independent simulations — it should be noted that Oleinikovas *et al.* reported binding of multiple benzene molecules to this pocket when simulating the *apo*-protein using their SWISH method.[249] The fact that binding is rapidly observed in all three cases when the pocket is held open indicates that this is not a limitation of the search of the space by random insertions, but rather that the insertions are hindered by the pocket being closed — previous work by Oleinikovas *et al.* has found the opening of this pocket to be associated with a notable free energy penalty, as determined from a metadynamics analysis of the pocket exposure.[249] During the relaxation stages of GCNCMC moves, the motion of the system is likely to follow the path of least resistance, as governed by the forces acting on each of the particles. Given the fact that the opening of this pocket is thought to be very unfavourable, the binding of a benzene molecule to the closed pocket is likely to face a significant barrier. It is therefore possible that, during the relaxation stages, it is easier for the partially interacting benzene molecule to unbind from the protein (or move to some less obstructed site), than for the protein to adapt to accommodate the fragment.

The hypothesis described above — that fragment unbinding may be preferential to

rearrangement of the protein environment during the relaxation steps of NCMC insertion protocols — was tested using a series of restrained insertions (note that these are not NCMC moves). These simulations attempted to insert benzene molecules into the *apo*-structure, focusing on the location of the cryptic pocket. Here, a random point was selected within a sphere based on the cryptic pocket (centred on the C_α atoms of Arg222 and Ile282, with a radius of 4 Å), and a non-interacting benzene molecule was placed at that location, with a random orientation. First, a harmonic force was gradually introduced over 5 ps (with the harmonic constant linearly increased to 1 kcal mol^{-1} $Å^{-2}$) to restrain the centre of geometry of the benzene to the randomly generated point. The benzene was then alchemically coupled to the system over 50 ps, and then the restraint was removed over a further 5 ps. During this nonequilibrium protocol, 50 timesteps of relaxation were carried out between perturbation steps. Prior to each insertion protocol, the system was sampled for 20 ps, and after the insertion protocol, the coordinates were saved and then reset — this was repeated for 1000 insertions. For comparison, an analogous set of unrestrained insertions were carried out in an identical fashion, except that no restraints were employed and the entire 60 ps of switching was devoted to alchemical coupling of the benzene molecule.

Visual inspection of the generated structures confirmed that the vast majority of the unrestrained insertions result in the inserted benzene leaving the insertion region — either ending up on the protein surface or in bulk solvent. This effect is also observed to some extent from the restrained insertions — as the restraint is removed, the ligand is free to leave — but a large number of these insertions result in benzene molecules bound either to the cryptic pocket or very close. This difference was analysed quantitatively by calculating the extent to which the cryptic pocket is opened by each benzene insertion, using the method of pocket exposure calculation described by Oleinikovas *et al.*[249] This involves first determining the pocket-lining atoms of the crystal structure — those protein heavy atoms within 4.5 Å of the ligand bound to the cryptic pocket — then the pocket exposure of a simulation frame is calculated as the percentage of these atoms which are determined to be pocket-lining using the fpocket tool.[261] Fig. 6.5a shows a plot of the work done during the benzene insertion against the pocket exposure of the protein following the insertion. It can be seen that insertions which induce a greater degree of pocket exposure also require more work to insert the molecule — likely related to the free energy penalty of opening the cryptic pocket.[249] It can also be seen that the restrained insertions consistently cause greater pocket exposure than the unrestrained insertions, where a large majority of the latter have no impact on the pocket exposure. It should be noted that the unrestrained insertions do sometimes bind benzene to the cryptic pocket, but these cases seem to be rare, and extensive sampling would likely be required to observe such an insertion when the sampling is focused on the entire protein. In any case, it is of some concern that the work done by the insertions which induce pocket exposure is typically very positive, and as such, it may be



(A) $\tau = 60$ ps. (B) $\tau = 120$ ps.

FIGURE 6.5: Work done by restrained and unrestrained benzene insertions into the TEM1 cryptic pocket, along with the extent to which the pocket is opened by the insertion — data is shown for switching times of 60 ps and 120 ps. For reference, the pocket exposures of the *apo*- (PDB ID: 1JWP[258]) and *holo*-structures (PDB ID: 1PZO[259]) were calculated as 2.9 % and 80 %, respectively.

very difficult to accept such insertions as GCNCMC moves. However, the distribution of work values for these insertions can be shifted towards more negative values by using longer switching times, as shown in Fig. 6.5b, where the insertions were repeated with doubled switching times. It is interesting to note that the probability of proposing a state with high pocket exposure appears to decrease with longer switching time for the unrestrained insertions — this may be because slower switching times give the partially interacting fragment more time to escape from the binding site. These results therefore suggest that the use of restraints may be of benefit during GCNCMC moves, and this should likely be an avenue of future development in this area.

6.4 Summary

This chapter presents the extension of the *grand* Python module to allow the grand canonical insertion and deletion of non-water molecules, with benzene used as an example here. In this work, GCNCMC moves were used to achieve acceptance rates of around 25 % for simulations on bulk solution, making grand canonical sampling of fragment-like molecules in condensed phases very accessible. For benzene, GCNCMC moves (with a switching time of 50 ps) offer an acceptance rate six orders of magnitude larger than that obtained using conventional GCMC ($\sim 1 \times 10^{-5}$ %). This therefore implies that GCNCMC could be used to sample molecules larger than benzene, whilst still obtaining a reasonable acceptance rate — though in some cases, this

may necessitate a longer switching time, in order to provide sufficient system relaxation. Testing the GCNCMC/MD implementation on a simulation of bulk water yields approximately the correct concentration (Fig. 6.2), for which the thermodynamic parameters (excess chemical potential and average volume for water and benzene) were calibrated. However, it should be noted that the concentration distributions observed are rather sensitive to subtle changes in these parameters — even a small change in the calculated excess chemical of water can have a notable impact on the concentration of benzene in the solution (Fig. 6.3) — further underlining the point made in chapter 3 that the accuracy of grand canonical simulations is dependent on rigorous calibration of these parameters. It should also be noted that benzene may be a sub-optimal choice of compound for this validation (owing to its very hydrophobic nature), and testing of a number of compounds should be carried out, preferably covering a range of solubility.

Application of this methodology to a protein system (TEM1 β-lactamase) yielded mixed results. It was demonstrated that this method is able to rapidly insert fragments into preformed binding sites, by simulating a restrained TEM1 protein structure, in which the cryptic pocket was held open. However, when the pocket is closed, as in the *apo*-structure, no benzene binding to the pocket is observed. Visualisation of these simulations indicated that when benzene molecules were inserted into the closed cryptic pocket, the benzene molecule tends to unbind from the protein during the course of the move, rather than the desired effect of the pocket opening to accommodate the ligand. This observation is likely caused by the fact that there is a free energy barrier associated with the opening of a cryptic pocket — particularly one requiring as significant a protein rearrangement as the TEM1 pocket[249] — which results in the relaxation stages of the GCNCMC insertion favouring the relocation of the inserted fragment. It therefore appears that, whilst this method is able to insert fragments into open pockets (as demonstrated by the simulations in which the cryptic pocket was held open), it will likely be very limited for the insertion of fragment molecules into cryptic binding pockets. However, it is not clear how well this approach will perform for fragment insertions which require only modest rearrangements of the binding pocket — further testing should therefore be performed in future work.

A series of restrained and unrestrained nonequilibrium benzene insertions into the cryptic site revealed that incorporating positional restraints into GCNCMC moves can help to prevent the fragment from unbinding during the relaxation stages (as described above), and were observed to induce a greater degree of pocket exposure than unrestrained insertions. However, it should be noted that this additional pocket opening comes at the cost of the work distributions being positively shifted, which would reduce the acceptance rate if these were Monte Carlo moves. This relates to the 'quantity

versus quality' discussion by Gill *et al.*, that not only the number of NCMC moves accepted is important, but also the types of moves accepted,[164] i.e. in this context, moves which open pockets are more meaningful than those in which the fragment unbinds from the protein, and would justify a slightly lower acceptance rate.

Two points which have not been discussed thus far in this chapter are that the current implementation in *grand* is limited to molecules which are charge neutral and conformationally rigid. Charge neutrality is required in order to maintain that of the system as a whole — in order to circumvent this limitation, the implementation would have to be modified to couple molecular insertion/deletion with the insertion/deletion of an appropriately charged ion, such that the move does not disturb the total system charge. Conformational rigidity is required at this stage because the geometry of an inserted molecule must mimic that of an ideal gas molecule, which typically involves randomising the orientation of the molecule upon insertion. However, for flexible molecules, this would also involve selecting a molecular conformation according to the conformational distribution observed in the ideal gas — this could be generated by running a gas phase simulation for each fragment, from which conformations would be selected at random upon insertion. Future work might seek to extend the *grand* module to remove these limitations.

A practical limitation of this work, as presented, is that the grand canonical simulation of a given ligand requires the calibration of several parameters, which are additionally concentration-dependent. A concern is that this could become prohibitively computationally expensive when simulating a number of fragment molecules, if each fragment must be rigorously parameterised at multiple concentrations (especially as the choice of concentration is somewhat arbitrary). Self-adjusted mixture sampling (SAMS) is an expanded ensemble method which can be used to simulate across a range of states, using adaptive biases which regulate the sampling across these states.[262] These biases can be used to estimate the free energy differences between these states from a single simulation, which can be more efficient than running multiple sets of free energy calculations. SAMS could be used to calculate the free energy differences between states containing different numbers of ligands in solution with water, in order to map the excess chemical potential to the ligand concentration, similarly to the method employed by Ross *et al.*[168] However, a further complication is that the concentrations sampled by GCNCMC/MD simulations appear to be very sensitive to these parameters (Fig. 6.3). An alternative approach is to titrate the system over some range of B^L, as done by Clark *et al.*, in order to identify fragment binding locations and a relative ranking between them (fragments which bind at lower B^L values would be interpreted to bind more strongly).[251,252] It should be noted that this approach would not generate a single equilibrium distribution, as would be obtained from a simulation at B_{equil}. However, a

key difference is that — unlike GCMC titrations of water — the Adams values where $B \neq B_{equil}$ do not correspond to non-physical states, but rather to reference solutions of different ligand concentrations, therefore making GCMC titrations of ligands more akin to experimental titrations. A GCNCMC/MD titration over a range of B^L values would therefore correspond to running a series of simulations over an unknown range of ligand concentrations — though the B^L values could be mapped to a known concentration range by simply parameterising the ligand of interest (principally the excess chemical potential) at several concentrations. Although, if the concentration range of interest is suitably low, then it may not be necessary to parameterise the ligand over a concentration range, as the concentration-dependence of the excess chemical potential is likely to decrease for more dilute solutions — therefore, it may suffice to calculate the hydration free energy for an infinitely dilute ligand solution. This approach could be very useful in a fragment-based drug design context, where the locations and relative stabilities of fragment binding are of great interest.

In summary, further work is required on this project. As previously discussed, there are some limitations to the current implementation (specifically that it is currently restricted to neutral, rigid molecules) which should be resolved, with some suggestions as to how this could be done discussed above. However, significantly more testing of this methodology is needed. First, the validation of GCNCMC/MD for small molecules, in terms of the ability to reproduce concentrations of bulk solutions should be further tested by expanding the tests performed in this work to include more (preferably diverse) fragments, across a range of concentrations. Also, given the observed sensitivity of these results to the GCMC parameters, care should be taken to obtain precise and accurate estimates of these values — particularly the excess chemical potentials, for which the uncertainty is greater (Table 6.2). However, given that the ultimate aim in developing this method is to produce a tool for use in fragment-based drug design, it is especially important that tests be carried out in this regard. Ideally, this would employ a dataset containing a range of proteins with multiple, experimentally verified fragment binding locations. These proteins should then be subjected to GCNCMC/MD titrations (as described above) with a series of fragments, in order to identify a range of fragment binding locations (including a relative ranking between them), which can then be compared to the experimental data, in order to assess the performance of this method for realistic applications.

Chapter 7

Conclusions

7.1 Summary

Protein-bound water molecules are now a key feature in structure-based drug design, owing to the gain in entropy which is typically associated with their displacement by a molecule which binds to the protein.[10] As molecular simulations are now commonplace in computer-aided drug design,[39] the long timescales of water exchange between protein binding sites and bulk solution[106] presents a limitation of these simulations. Grand canonical Monte Carlo (GCMC) is a simulation method which can aid in the sampling of buried water molecules by inserting and deleting waters to/from a region of interest.[104,105,109] During this work, GCMC sampling of water molecules in MD simulations has been made more accessible (via an open source Python module, named *grand*), and several theoretical and methodological developments were made which extend the applicability of GCMC. The work presented in this thesis is summarised below.

Chapter 3 discusses the implementation of GCMC/MD in the *grand* Python module, which was developed during this work. This module makes GCMC sampling of water molecules easily accessible within the OpenMM simulation framework, requiring very little additional knowledge for those already familiar with this simulation engine. The accuracy of this implementation was verified by obtaining the same density distributions of bulk water, using both GCMC/MD and constant pressure simulations. Further validation was carried out using the statistical method proposed by Shirts,[192] demonstrating that the distribution of the particle number, N, generated using *grand* responds correctly to changes in the imposed chemical potential. Finally, for a simple protein system (BPTI), it was shown that GCMC/MD sampling allows rapid equilibration of the binding site waters, where conventional MD is much slower.

Chapter 4 demonstrates how GCMC/MD titrations can be used to investigate the thermodynamics of water binding to a structure of interest. In this work, GCMC/MD titrations were carried out on the binding site of the transmembrane domain of the M2 protein in complex with various inhibitors. The data collected suggest a water-mediated mechanism of inhibitor stereoselectivity, where the presence of a large water network 'cancels' the chirality of the protein, and that when there are fewer waters present, a chiral difference becomes apparent. This is of particular interest for the next generation of M2 inhibitors, such as spiroadamantane, for which only one layer of waters are observed (in complex with wild type M2), rather than two.[216,219] A series of titration calculations were also carried out for both amantadine and spiroadamantane, in complex with the wild type (WT) and V27A structures of M2, where the V27A mutant is resistant to amantadine. No structural or thermodynamic evidence was found to suggest a water-mediated resistance mechanism for the V27A mutation — it was considered that a water wire might be able to bypass amantadine (allowing proton conductance) when bound to the V27A structure, but the simulation data did not support this hypothesis. However, these simulations were able to provide further insight into the structural features of these complexes. Notably, the multiple binding modes adopted by spiroadamantane,[219] depending on the identity of residue 27 appear to be separated by an extremely subtle difference in free energy, which could be problematic for computational analyses of this protein if the force field does not capture this subtlety correctly. Future computational investigations of the V27A resistance mechanism should likely make use of mutation free energy calculations, and also consider the impact of pH on ligand binding.

In chapter 5, nonequilibrium candidate Monte Carlo (NCMC)[115] is used to improve the acceptance rates of GCMC moves, by sampling the configurational degrees of freedom whilst a water is gradually inserted or deleted, referred to as grand canonical nonequilibrium candidate Monte Carlo (GCNCMC). This allows the environment to relax in response to the proposed change, automatically resolving any steric clashes which typically hamper the acceptance of GCMC moves. The acceptance criteria for GCNCMC moves were derived, which are remarkably similar to those of instantaneous GCMC moves. Additionally, it is demonstrated that, by using GCNCMC, the acceptance rates can be improved by three orders of magnitude over conventional GCMC. Despite the significantly increased computational cost of these moves, GCNCMC can be more efficient (in terms of the number of accepted moves per unit wall time) than conventional GCMC by up to a factor of five.

Chapter 6 shows how the *grand* module has been extended to allow GCMC sampling of non-water molecules — reasonable acceptance rates can be achieved using the implementation of GCNCMC presented in chapter 5. It was found that GCNCMC offers

an improvement in the acceptance rate by six orders of magnitude over instantaneous GCMC for sampling of benzene in solution. Grand canonical parameters were calculated for benzene at several concentrations, which were then used to verify that GCNCMC/MD samples suitable concentration distributions for bulk solution — although it should be noted that the results are rather sensitive to the calibration of these parameters. GCNCMC/MD sampling of benzene on the TEM1 β-lactamase protein showed that this method is able to rapidly identify fragment binding locations in pre-formed pockets, but that pockets where fragment binding requires significant protein reorganisation are problematic. As a proof of concept, it was demonstrated that restraining the fragment during nonequilibrium insertion can reduce this issue, and increase the probability of proposing move which opens a closed pocket. The possibility of including positional restraints in GCNCMC moves is therefore an option for future work.

7.2 Future Work

Whilst significant progress has been made during this work, there remain a number of directions in which future work could build upon that presented in this thesis. In this section, some options for future research are discussed.

As briefly mentioned in chapter 5, if GCNCMC moves are used to sample a spherical region (centred on a protein binding site, perhaps), then the move must be automatically rejected if the switched water lies outside the sphere at the end of the move. This is a necessary requirement in order to maintain detailed balance, as the reverse move has zero probability of being proposed. However, in some cases (not presented here), this can cause a large number of otherwise favourable moves to be rejected, and can remove a large portion of the increased efficiency offered by GCNCMC. Solving this issue would be very beneficial, where one possible approach could be to carry out GCNCMC sampling of a larger region around the GCMC sphere, as well as the sphere itself. This would mean that the probability of proposing the reverse move is non-zero for moves where the water leaves the sphere, and therefore, such moves need not be rejected automatically. An additional benefit of sampling the larger region is that this GCMC sampling could aid in allowing the density of the entire system to fluctuate.[110] Similarly, if a convenient solution to preventing water diffusion in/out of the GCMC sphere can be developed for GCMC/MD simulations, this would also make GCMC/MD titrations much more accessible for binding sites which are somewhat solvent-exposed. It may be possible to implement a suitable restraint/constraint protocol — which prevents water diffusion across the boundaries of the GCMC sphere during the MD steps — without adversely affecting the dynamics of the rest of the system.

Previous work using GCMC titrations in ProtoMS has made use of replica exchange, where attempts are periodically made to exchange configurations between adjacent B values. It was found by Ross *et al.* that this significantly reduces the noise in titration curves, and therefore also reduces the uncertainties in the water binding free energies determined.[105] However, this is not implemented in *grand*, as replica exchange simulations are not easily carried out in OpenMM. If a convenient solution to this issue can be found in future work, it would significantly improve the quality of GCMC/MD titration calculations performed using *grand*.

The extension of the *grand* module to allow GCMC sampling of small molecules, as presented in chapter 6, appears promising and could be very useful in computational fragment-based drug design. One limitation noted in this work, is that when fragment binding requires some degree of protein rearrangement, it may be that the relaxation component of GCNCMC causes the fragment to unbind, rather than the protein to rearrange. As a proof of concept, it was shown in chapter 6 that applying restraints during nonequilibrium insertion can alleviate this to some extent — future work could build upon this to make use of restraints during GCNCMC moves in *grand*. It should be noted that there are also some other limitations of the implementation of GCMC moves for small molecules in *grand*. First, the latest development version of *grand* is only correct for molecules with no conformational degrees of freedom. In order to replicate the translation of a flexible molecule from an ideal gas, a conformation would have to be generated upon insertion, and this conformation would have to be drawn at random from the conformational distribution observed in the ideal gas. In practice, this would involve running a gas phase simulation of a single molecule to generate an ensemble from which conformations would be selected at random — though this additional step has not yet been included in *grand*. Secondly, as previously mentioned, *grand* can only be used to carry out GCMC sampling of charge neutral molecules, in order to maintain the charge neutrality of the simulation. This could be resolved by coupling the insertion/deletion of a charged molecule, with that of an appropriately charged ion, similar to the insertion of NaCl pairs in the *saltswap* method presented by Ross *et al.*[168]

A powerful application of GCMC is the execution of relative binding free energy calculations in the grand canonical ensemble, where the hydration of the binding site is automatically adjusted to the ligand perturbation.[107,109,110] Unfortunately, this is not currently possible in *grand* — relative free energy calculations are generally rather difficult in OpenMM, although the *perses* module under development will likely make this more feasible.[263,264] Future work to make *grand* compatible with *perses* would be very beneficial in this regard — although the latest development version of *grand* could be used to carry out absolute binding free energy calculations (in which a ligand is decoupled from the binding site, rather than perturbed into another ligand), these are

typically less useful than relative binding free energy calculations. Additionally, the combination of GCMC/MD sampling with other enhanced sampling methods, such as constant pH simulations, or the sampling of ligand binding modes[164] — thereby allowing multiple orthogonal, slow degrees of freedom to be sampled during the same simulation — would be very interesting. This would be especially useful if these degrees of freedom are likely to be correlated, i.e. if two ligand binding modes or protonation states are hydrated differently, for example.

A factor which has not been discussed in this work is that water models are typically parameterised to reproduce the properties of bulk water.[182] However, as many protein binding sites are likely to be very different to the environment of bulk water, it is likely that these bulk water parameters are not transferable to protein-bound water sites. Specifically, the polarisation of the water sites is likely to be very different. An interesting line of future investigation would be to study the differences introduced by a better treatment of water polarisation — for waters with binding free energies close to zero, this could make the difference between their binding being favourable or unfavourable. However, polarisable force fields, such as AMOEBA,[134–138] tend to significantly increase the computational cost of a simulation. A more attractive approach could be to resample microstates from a fixed charge simulation according to a polarisable force field, in order to reweight the relative probabilities of the configurations, as proposed by Cave-Ayland *et al.*[265]

In summary, whilst significant developments have been made to grand canonical simulations in this work, there are still a number of opportunities for further development in this field, building upon the work presented in this thesis.

Appendix A

Cancellation of the Rotational Partition Function in GCMC

A.1 GCMC Acceptance Criteria

Here, the derivation of GCMC particle insertion moves is presented, as in section 2.6.3.1, but with the ideal rotational partition function explicitly included. First, we make use of the fact that the probability of a given configuration in the canonical ensemble (independent of particle labelling), can be written as:

$$\pi_{NVT}(\mathbf{r}^N) = \frac{Q^{id}_{NVT}}{Q_{NVT}} N!\, e^{-\beta U(\mathbf{s}^N)} \mathrm{d}\mathbf{s}^N \tag{A.1}$$

where the factorial term indicates a sum over all possible particle label arrangements. Therefore, for the large canonical ensemble used in the GCMC derivation (containing $M - N$ particles in the ideal gas, and N particles in the system), the probability is:

$$\pi_{MVT}(\mathbf{x}^N, \mathbf{x}^{M-N}) = \frac{Q^{id}_{(M-N)V_iT} Q^{id}_{NV_sT}}{Q_{MVT}} (M-N)!N!\, e^{-\beta U(\mathbf{x}^N)} \mathrm{d}\mathbf{x}^N \mathrm{d}\mathbf{x}^{M-N} \tag{A.2}$$

where $\mathbf{x}$ is a vector containing the positions and orientations of each particle (scaled such that if unity were integrated over all possible coordinates, the integral would yield unity), and the ideal partition functions for the system and ideal gas are:

$$Q^{id}_{NV_sT} = \frac{(q^{trans}_{V_sT} q^{rot}_T)^N}{N!} \tag{A.3}$$

$$Q^{id}_{(M-N)V_iT} = \frac{(q^{trans}_{V_iT} q^{rot}_T)^{M-N}}{(M-N)!} \tag{A.4}$$

where the fact that q^{trans} depends on both the volume and temperature, and that q^{rot} depends only on temperature[177,266] (for a given molecule) has been made explicit, for

clarity. Given that we are operating in scaled coordinates, the probabilities of proposing the forward and reverse moves are:

$$P(\mathbf{x}^{N+1}|\mathbf{x}^{N}) = \frac{1}{2}\frac{\mathrm{d}\mathbf{s}}{M-N} \tag{A.5}$$

$$P(\mathbf{x}^{N}|\mathbf{x}^{N+1}) = \frac{1}{2}\frac{\mathrm{d}\mathbf{s}}{N+1} \tag{A.6}$$

where $\mathbf{s}$ is a scaled position vector, and insertions and deletions are proposed with equal probability.

Again, we can combine these terms into the acceptance ratio (where terms which immediately cancel have been omitted, due to space constraints):

$$\begin{aligned}\frac{A(\mathbf{x}^{N+1}|\mathbf{x}^{N})}{A(\mathbf{x}^{N}|\mathbf{x}^{N+1})} &= \frac{(N+1)^{-1}}{(M-N)^{-1}}\frac{Q^{id}_{(M-N-1)V_iT}Q^{id}_{(N+1)V_sT}(M-N-1)!(N+1)!\,e^{-\beta U(\mathbf{x}^{N+1})}}{Q^{id}_{(M-N)V_iT}Q^{id}_{NV_sT}(M-N)!N!\,e^{-\beta U(\mathbf{x}^{N})}} \\ &= \frac{M-N}{N+1}\frac{(q^{trans}_{V_iT}q^{rot}_{T})^{M-N-1}(q^{trans}_{V_sT}q^{rot}_{T})^{N+1}}{(q^{trans}_{V_iT}q^{rot}_{T})^{M-N}(q^{trans}_{V_sT}q^{rot}_{T})^{N}}e^{-\beta\Delta U} \\ &= \frac{q^{trans}_{V_sT}}{N+1}\frac{M-N}{q^{trans}_{V_iT}}e^{-\beta\Delta U}\end{aligned} \tag{A.7}$$

We can now make the following substitution (Eq. 2.148):

$$\lim_{M\to\infty}\frac{M-N}{q^{trans}_{V_iT}} = q^{rot}_{T}e^{\beta\mu} \tag{A.8}$$

which allows the derivation of the acceptance ratio to be completed:

$$\begin{aligned}\frac{A(\mathbf{x}^{N+1}|\mathbf{x}^{N})}{A(\mathbf{x}^{N}|\mathbf{x}^{N+1})} &= \frac{q^{trans}_{V_sT}}{N+1}q^{rot}_{T}e^{\beta\mu}e^{-\beta\Delta U} \\ &= \frac{1}{N+1}\frac{q^{rot}_{T}V_{sys}}{\Lambda^{3}}e^{\beta\mu}e^{-\beta\Delta U} \\ &= \frac{1}{N+1}e^{B'}e^{-\beta\Delta U}\end{aligned} \tag{A.9}$$

where the following, modified Adams parameter has been introduced:

$$B' = \beta\mu + \ln\left(\frac{q^{rot}_{T}V_{sys}}{\Lambda^{3}}\right) \tag{A.10}$$

Whilst it appears above that the rotational partition function does change the acceptance ratio, this is only true if GCMC is carried out by setting the chemical potential

of the simulation directly. However, in practice, the chemical potential tends to be set via the Adams value.[99,100] The equilibrium Adams value, when accounting for ideal rotations, is the following:

$$\begin{aligned} B'_{equil} &= \beta\mu^{\circ}_{sol} + \ln\left(\frac{q_T^{rot}V_{sys}}{\Lambda^3}\right) \\ &= \beta\left(\mu^{ex}_{sol} + k_BT\ln\left(\frac{\Lambda^3}{q_T^{rot}V^{\circ}}\right)\right) + \ln\left(\frac{q_T^{rot}V_{sys}}{\Lambda^3}\right) \\ &= \beta\mu^{ex}_{sol} + \ln\left(\frac{V_{sys}}{V^{\circ}}\right) \end{aligned} \tag{A.11}$$

where the rotational partition function cancels. Therefore, when carrying out GCMC simulations where the chemical potential is regulated via the Adams value, the rotational partition function need not be accounted for. An equivalent derivation can demonstrate this for GCMC deletion moves.

A.2 Titration Calculations

Whilst the above demonstrates that the rotational partition function does not impact GCMC sampling of water, it may be the case that this term needs to be included in the analysis of titration calculations. Whilst this appears unlikely, given the excellent agreement reported between the free energies calculated from GCMC titrations and those using rigorous double decoupling calculations,[105] the cancellation of q^{rot} is demonstrated from first principles in this section.

First, we recall Eq. 2.133:

$$\Delta G_{bind} = \Delta F_{trans} + \Delta F_{ideal} - \Delta G_{sol} \tag{2.133 revisited}$$

Ross *et al.* stated that the calculation of ΔF_{trans} is unaffected by ideal rotations of water molecules.[104] However, no such statement has been made in the literature for ΔG_{bind}, so this is demonstrated here. The Helmholtz free energy of N non-spherical ideal gas particles can be calculated analytically:

$$F_{ideal}(N) = -k_BT\ln\left(\frac{(q_{V_iT}^{trans}q_T^{rot})^N}{N!}\right) \tag{A.12}$$

Therefore, the difference in free energy between states containing N_i and N_f particles is:

$$\beta\Delta F_{ideal}(N_i \to N_f) = \ln\left(\frac{N_f!}{N_i!}\right) - (N_f - N_i)\ln\left(\frac{q_T^{rot}V_{ideal}}{\Lambda^3}\right) \tag{A.13}$$

The difference in solvent free energy (accounting for the rotational variation) is:

$$\Delta G_{sol}(N_i \rightarrow N_f) = (N_f - N_i)\left(\mu_{sol}^{ex} + k_B T \ln\left(\frac{\rho_{sol}\Lambda^3}{q_T^{rot}}\right)\right) \tag{A.14}$$

It therefore transpires that the rotational contributions to ΔF_{ideal} and ΔG_{sol} cancel exactly. This confirms that the rotational partition function of water need not be included in the free energy analysis of GCMC titration calculations.

References

[1] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea and J. P. Overington, *Nat. Rev. Drug Discov.*, 2017, **16**, 19–34.

[2] R. D. Taylor, P. J. Jewsbury and J. W. Essex, *J. Comput. Aided. Mol. Des.*, 2002, **16**, 151–166.

[3] A. S. Mey, B. K. Allen, H. E. Bruce Macdonald, J. D. Chodera, D. F. Hahn, M. Kuhn, J. Michel, D. L. Mobley, L. N. Naden, S. Prasad, A. Rizzi, J. Scheen, M. R. Shirts, G. Tresadern and H. Xu, *Living J. Comp. Mol. Sci.*, 2020, **2**, 18378.

[4] C. Bissantz, B. Kuhn and M. Stahl, *J. Med. Chem.*, 2010, **53**, 5061–5084.

[5] P. Ball, *Chem. Rev.*, 2008, **108**, 74–108.

[6] J. D. Dunitz, *Science*, 1994, **264**, 670.

[7] C. S. Poornima and P. M. Dean, *J. Comput. Aided. Mol. Des.*, 1995, **9**, 500–512.

[8] C. S. Poornima and P. M. Dean, *J. Comput. Aided. Mol. Des.*, 1995, **9**, 513–520.

[9] C. S. Poornima and P. M. Dean, *J. Comput. Aided. Mol. Des.*, 1995, **9**, 521–531.

[10] J. E. Ladbury, *Chem. Biol.*, 1996, **3**, 973–980.

[11] Y. Lu, R. Wang, C.-Y. Yang and S. Wang, *J. Chem. Inf. Model.*, 2007, **47**, 668–675.

[12] J. M. Chen, S. L. Xu, Z. Wawrzak, G. S. Basarab and D. B. Jordan, *Biochemistry*, 1998, **37**, 17735–17744.

[13] C. Liu, S. T. Wrobleski, J. Lin, G. Ahmed, A. Metzger, J. Wityak, K. M. Gillooly, D. J. Shuster, K. W. McIntyre, S. Pitt, D. R. Shen, R. F. Zhang, H. Zhang, A. M. Doweyko, D. Diller, I. Henderson, J. C. Barrish, J. H. Dodd, G. L. Schieven and K. Leftheris, *J. Med. Chem.*, 2005, **48**, 6261–6270.

[14] A. Joncour, N. Desroy, C. Housseman, X. Bock, N. Bienvenu, L. Cherel, V. Labeguere, C. Peixoto, D. Annoot, L. Lepissier, J. Heiermann, W. J. Hengeveld, G. Pilzak, A. Monjardet, E. Wakselman, V. Roncoroni, S. Le Tallec, R. Galien,

C. David, N. Vandervoort, T. Christophe, K. Conrath, M. Jans, A. Wohlkonig, S. Soror, J. Steyaert, R. Touitou, D. Fleury, L. Vercheval, P. Mollat, N. Triballeau, E. van der Aar, R. Brys and B. Heckmann, *J. Med. Chem.*, 2017, **60**, 7371–7392.

[15] A. Wissner, D. M. Berger, D. H. Boschelli, M. B. Floyd, L. M. Greenberger, B. C. Gruber, B. D. Johnson, N. Mamuya, R. Nilakantan, M. F. Reich, R. Shen, H.-R. Tsou, E. Upeslacis, Y. F. Wang, B. Wu, F. Ye and N. Zhang, *J. Med. Chem.*, 2000, **43**, 3244–3256.

[16] N. N. Nasief, H. Tan, J. Kong and D. Hangauer, *J. Med. Chem.*, 2012, **55**, 8283–8302.

[17] P. Y. S. Lam, P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bacheler, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, C.-H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe and S. Erickson-Viitanen, *Science*, 1994, **263**, 380–384.

[18] M. Kim and A. E. Cho, *Sci. Rep.*, 2016, **6**, 36807.

[19] M. Aldeghi, G. A. Ross, M. J. Bodkin, J. W. Essex, S. Knapp and P. C. Biggin, *Commun. Chem.*, 2018, **1**, 19.

[20] J. G. Kettle, R. Anjum, E. Barry, D. Bhavsar, C. Brown, S. Boyd, A. Campbell, K. Goldberg, M. Grondine, S. Guichard, C. J. Hardy, T. Hunt, R. D. O. Jones, X. Li, O. Moleva, D. Ogg, R. C. Overman, M. J. Packer, S. Pearson, M. Schimpl, W. Shao, A. Smith, J. M. Smith, D. Stead, S. Stokes, M. Tucker and Y. Ye, *J. Med. Chem.*, 2018, **61**, 8797–8810.

[21] M. S. Bodnarchuk, M. J. Packer and A. Haywood, *ACS Med. Chem. Lett.*, 2020, **11**, 77–82.

[22] J. R. Tame, S. H. Sleigh, A. J. Wilkinson and J. E. Ladbury, *Nat. Struct. Mol. Biol.*, 1996, **3**, 998–1001.

[23] A. M. Davis, S. J. Teague and G. J. Kleywegt, *Angew. Chem. Int. Ed.*, 2003, **42**, 2718–2736.

[24] A. McPherson, *Methods*, 2004, **34**, 254–265.

[25] G. J. Kleywegt, *Acta Crystallogr. D Biol. Crystallogr.*, 2000, **56**, 249–265.

[26] D. H. Ohlendorf, *Acta Crystallogr. D Biol. Crystallogr.*, 1994, **50**, 808–812.

[27] B. A. Fields, H. H. Bartsch, H. D. Bartunik, F. Cordes, J. M. Guss and H. C. Freeman, *Acta Crystallogr. D Biol. Crystallogr.*, 1994, **50**, 709–730.

[28] D. Myles, *Curr. Opin. Struc. Biol.*, 2006, **16**, 630–637.

[29] W. B. O'Dell, A. M. Bodenheimer and F. Meilleur, *Arch. Biochem. Biophys.*, 2016, **602**, 48–60.

[30] A. S. Gardberg, A. R. Del Castillo, K. L. Weiss, F. Meilleur, M. P. Blakeley and D. A. A. Myles, *Acta Crystallogr. D Biol. Crystallogr.*, 2010, **66**, 558–567.

[31] *Protein Data Bank*, `rcsb.org`, (Date accessed: March 31, 2020).

[32] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.

[33] G. Otting, E. Liepinsh and K. Wuthrich, *Science*, 1991, **254**, 974–980.

[34] G. Otting, *Prog. Nucl. Magn. Reson. Spectrosc.*, 1997, **31**, 259–285.

[35] B. Halle, *Phil. Trans. R. Soc. Lond. B*, 2004, **359**, 1207–1224.

[36] K. Modig, E. Liepinsh, G. Otting and B. Halle, *J. Am. Chem. Soc.*, 2004, **126**, 102–114.

[37] J. M. Gruschus and J. A. Ferretti, *J. Biomol. NMR*, 2001, **20**, 111–126.

[38] M. S. Bodnarchuk, *Drug. Discov. Today*, 2016, **21**, 1139–1146.

[39] A. P. Graves, I. D. Wall, C. M. Edge, J. M. Woolven, G. Cui, A. Le Gall, X. Hong, K. Raha and E. S. Manas, *Curr. Top. Med. Chem.*, 2017, **17**, 1–18.

[40] M. L. Samways, R. D. Taylor, H. E. Bruce Macdonald and J. W. Essex, *Chem. Soc. Rev.*, 2021, 10.1039.D0CS00151A.

[41] W. R. Pitt and J. M. Goodfellow, *Protein Eng. Des. Sel.*, 1991, **4**, 531–537.

[42] W. R. Pitt, J. Murray-Rust and J. M. Goodfellow, *J. Comput. Chem.*, 1993, **14**, 1007–1018.

[43] I. J. Bruno, J. C. Cole, J. P. Lommerse, R. S. Rowland, R. Taylor and M. L. Verdonk, *J. Comput. Aided. Mol. Des.*, 1997, **11**, 525–537.

[44] M. L. Verdonk, J. C. Cole and R. Taylor, *J. Mol. Biol.*, 1999, **289**, 1093–1108.

[45] M. L. Verdonk, J. C. Cole, P. Watson, V. Gillet and P. Willett, *J. Mol. Biol.*, 2001, **307**, 841–859.

[46] G. Rossato, B. Ernst, A. Vedani and M. Smieško, *J. Chem. Inf. Model.*, 2011, **51**, 1867–1881.

[47] E. Nittinger, F. Flachsenberg, S. Bietz, G. Lange, R. Klein and M. Rarey, *J. Chem. Inf. Model.*, 2018, **58**, 1625–1637.

[48] W. Xiao, Z. He, M. Sun, S. Li and H. Li, *J. Chem. Inf. Model.*, 2017, **57**, 1517–1528.

[49] M. Zheng, Y. Li, B. Xiong, H. Jiang and J. Shen, *J. Comput. Chem.*, 2013, **34**, 583–592.

[50] M. Jukič, J. Konc, S. Gobec and D. Janežič, *J. Chem. Inf. Model.*, 2017, **57**, 3094–3103.

[51] H. Patel, B. A. Gruning, S. Gunther and I. Merfort, *Bioinformatics*, 2014, **30**, 2978–2980.

[52] P. C. Sanschagrin and L. A. Kuhn, *Protein Sci.*, 1998, **7**, 2054–2064.

[53] C. A. Bottoms, T. A. White and J. J. Tanner, *Proteins*, 2006, **64**, 404–421.

[54] M. L. Raymer, P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman and L. A. Kuhn, *J. Mol. Biol.*, 1997, **265**, 445–464.

[55] A. T. García-Sosa, R. L. Mancera and P. M. Dean, *J. Mol. Model.*, 2003, **9**, 172–182.

[56] G. A. Ross, G. M. Morris and P. C. Biggin, *PLoS ONE*, 2012, **7**, e32036.

[57] P. J. Goodford, *J. Med. Chem.*, 1985, **28**, 849–857.

[58] D. Beglov and B. Roux, *J. Phys. Chem. B*, 1997, **101**, 7821–7826.

[59] T. Imai, R. Hiraoka, A. Kovalenko and F. Hirata, *J. Am. Chem. Soc.*, 2005, **127**, 15334–15335.

[60] T. Imai, R. Hiraoka, A. Kovalenko and F. Hirata, *Proteins*, 2006, **66**, 804–813.

[61] D. J. Sindhikara, N. Yoshida and F. Hirata, *J. Comput. Chem.*, 2012, **33**, 1536–1543.

[62] L. Fusani, I. Wall, D. Palmer and A. Cortes, *Bioinformatics*, 2018, **34**, 1947–1948.

[63] P. Setny and M. Zacharias, *J. Phys. Chem. B*, 2010, **114**, 8667–8675.

[64] I. Y. Ben-Shalom, C. Lin, T. Kurtzman, R. C. Walker and M. K. Gilson, *J. Chem. Theory Comput.*, 2019, **15**, 2684–2691.

[65] A. Miranker and M. Karplus, *Proteins*, 1991, **11**, 29–34.

[66] H.-H. Bui, A. J. Schiewe and I. S. Haworth, *J. Comput. Chem.*, 2007, **28**, 2241–2251.

[67] L. Zhang and J. Hermans, *Proteins*, 1996, **24**, 433–438.

[68] A. Morozenko, I. V. Leontyev and A. A. Stuchebrukhov, *J. Chem. Theory Comput.*, 2014, **10**, 4618–4623.

[69] A. Morozenko and A. A. Stuchebrukhov, *Proteins*, 2016, **84**, 1347–1357.

[70] A. Sridhar, G. A. Ross and P. C. Biggin, *PLoS ONE*, 2017, **12**, e0172743.

[71] M. Rarey, B. Kramer and T. Lengauer, *Proteins*, 1999, **34**, 17–28.

[72] V. Schnecke and L. A. Kuhn, *Perspect. Drug Discov. Des.*, 2000, **20**, 171–190.

[73] R. A. Friesner, R. B. Murphy, M. P. Repasky, L. L. Frye, J. R. Greenwood, T. A. Halgren, P. C. Sanschagrin and D. T. Mainz, *J. Med. Chem.*, 2006, **49**, 6177–6196.

[74] T. Young, R. Abel, B. Kim, B. J. Berne and R. A. Friesner, *Proc. Natl. Acad. Sci. USA*, 2007, **104**, 808–813.

[75] R. Abel, T. Young, R. Farid, B. J. Berne and R. A. Friesner, *J. Am. Chem. Soc.*, 2008, **130**, 2817–2831.

[76] R. B. Murphy, M. P. Repasky, J. R. Greenwood, I. Tubert-Brohman, S. Jerome, R. Annabhimoju, N. A. Boyles, C. D. Schmitz, R. Abel, R. Farid and R. A. Friesner, *J. Med. Chem.*, 2016, **59**, 4364–4384.

[77] M. L. Verdonk, G. Chessari, J. C. Cole, M. J. Hartshorn, C. W. Murray, J. W. M. Nissink, R. D. Taylor and R. Taylor, *J. Med. Chem.*, 2005, **48**, 6504–6515.

[78] C. R. Corbeil, P. Englebienne and N. Moitessier, *J. Chem. Inf. Model.*, 2007, **47**, 435–449.

[79] C. N. Nguyen, T. Kurtzman Young and M. K. Gilson, *J. Chem. Phys.*, 2012, **137**, 044101.

[80] S. Uehara and S. Tanaka, *Molecules*, 2016, **21**, 1604.

[81] H. Sun, L. Zhao, S. Peng and N. Huang, *Proteins*, 2014, **82**, 1765–1776.

[82] T. E. Balius, M. Fischer, R. M. Stein, T. B. Adler, C. N. Nguyen, A. Cruz, M. K. Gilson, T. Kurtzman and B. K. Shoichet, *Proc. Natl. Acad. Sci. USA*, 2017, **114**, E6839–E6846.

[83] M. A. Lie, R. Thomsen, C. N. S. Pedersen, B. Schiøtt and M. H. Christensen, *J. Chem. Inf. Model.*, 2011, **51**, 909–917.

[84] F. Österberg, G. M. Morris, M. F. Sanner, A. J. Olson and D. S. Goodsell, *Proteins*, 2002, **46**, 34–40.

[85] G. Lemmon and J. Meiler, *PLoS ONE*, 2013, **8**, e67536.

[86] M. S. Bodnarchuk, R. Viner, J. Michel and J. W. Essex, *J. Chem. Inf. Model.*, 2014, **54**, 1623–1633.

[87] M. K. Gilson, J. A. Given, B. L. Bush and J. A. McCammon, *Biophys. J.*, 1997, **72**, 1047–1069.

[88] C. Barillari, J. Taylor, R. Viner and J. W. Essex, *J. Am. Chem. Soc.*, 2007, **129**, 2577–2587.

[89] T. Lazaridis, *J. Phys. Chem. B*, 1998, **102**, 3531–3541.

[90] T. Lazaridis, *J. Phys. Chem. B*, 1998, **102**, 3542–3550.

[91] C. N. Nguyen, T. Kurtzman and M. K. Gilson, *J. Chem. Theory Comput.*, 2016, **12**, 414–429.

[92] B. Hu and M. A. Lill, *J. Comput. Chem.*, 2014, **35**, 1255–1260.

[93] A. H. Mahmoud, M. R. Masters, Y. Yang and M. A. Lill, *Commun. Chem.*, 2020, **3**, 19.

[94] G. Cui, J. M. Swails and E. S. Manas, *J. Chem. Theory Comput.*, 2013, **9**, 5539–5549.

[95] R. H. Henchman, *J. Chem. Phys.*, 2007, **126**, 064504.

[96] G. Gerogiokas, M. W. Y. Southey, M. P. Mazanetz, A. Hefeitz, M. Bodkin, R. J. Law and J. Michel, *Phys. Chem. Chem. Phys.*, 2015, **17**, 8416–8426.

[97] G. Gerogiokas, M. W. Y. Southey, M. P. Mazanetz, A. Heifetz, M. Bodkin, R. J. Law, R. H. Henchman and J. Michel, *J. Phys. Chem. B*, 2016, **120**, 10442–10452.

[98] J. Michel, J. Tirado-Rives and W. L. Jorgensen, *J. Phys. Chem. B*, 2009, **113**, 13337–13346.

[99] D. Adams, *Mol. Phys.*, 1974, **28**, 1241–1252.

[100] D. Adams, *Mol. Phys.*, 1975, **29**, 307–311.

[101] M. Mezei, *Mol. Phys.*, 1980, **40**, 901–906.

[102] M. Mezei, *Mol. Phys.*, 1987, **61**, 565–582.

[103] H.-J. Woo, A. R. Dinner and B. Roux, *J. Chem. Phys.*, 2004, **121**, 6392–6400.

[104] G. A. Ross, M. S. Bodnarchuk and J. W. Essex, *J. Am. Chem. Soc.*, 2015, **137**, 14930–14943.

[105] G. A. Ross, H. E. Bruce Macdonald, C. Cave-Ayland, A. I. Cabedo Martinez and J. W. Essex, *J. Chem. Theory Comput.*, 2017, **13**, 6373–6381.

[106] D. Laage, T. Elsaesser and J. T. Hynes, *Chem. Rev.*, 2017, **117**, 10694–10725.

[107] Y. Deng and B. Roux, *J. Chem. Phys.*, 2008, **128**, 115103.

[108] J. Wahl and M. Smieško, *J. Chem. Inf. Model.*, 2019, **59**, 754–765.

[109] H. E. Bruce Macdonald, C. Cave-Ayland, G. A. Ross and J. W. Essex, *J. Chem. Theory Comput.*, 2018, **14**, 6586–6597.

[110] G. A. Ross, E. Russell, Y. Deng, C. Lu, E. D. Harder, R. Abel and L. Wang, *J. Chem. Theory Comput.*, 2020, **16**, 6061–6076.

[111] J. Michel, J. Tirado-Rives and W. L. Jorgensen, *J. Am. Chem. Soc.*, 2009, **131**, 15403–15411.

[112] F.-X. Coudert and A. H. Fuchs, *Coord. Chem. Rev.*, 2016, **307**, 211–236.

[113] M. L. Samways, H. E. Bruce Macdonald and J. W. Essex, *J. Chem. Inf. Model.*, 2020, **60**, 4436–4441.

[114] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLoS Comput. Biol.*, 2017, **13**, e1005659.

[115] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh and J. D. Chodera, *Proc. Natl. Acad. Sci. USA*, 2011, **108**, E1009–E1018.

[116] A. R. Leach, *Molecular modelling: principles and applications*, Prentice Hall, Harlow, England ; New York, 2nd edn., 2001.

[117] D. Frenkel and B. Smit, *Understanding molecular simulation: from algorithms to applications*, Academic Press, San Diego, 2nd edn., 2002.

[118] D. L. Mobley, C. C. Bannan, A. Rizzi, C. I. Bayly, J. D. Chodera, V. T. Lim, N. M. Lim, K. A. Beauchamp, D. R. Slochower, M. R. Shirts, M. K. Gilson and P. K. Eastman, *J. Chem. Theory Comput.*, 2018, **14**, 6076–6092.

[119] C. Zanette, C. C. Bannan, C. I. Bayly, J. Fass, M. K. Gilson, M. R. Shirts, J. D. Chodera and D. L. Mobley, *J. Chem. Theory Comput.*, 2019, **15**, 402–423.

[120] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157–1174.

[121] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.

[122] A. Jakalian, B. L. Bush, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2000, **21**, 132–146.

[123] A. Jakalian, D. B. Jack and C. I. Bayly, *J. Comput. Chem.*, 2002, **23**, 1623–1641.

[124] C. I. Bayly, P. Cieplak, W. Cornell and P. A. Kollman, *J. Phys. Chem.*, 1993, **97**, 10269–10280.

[125] A. J. Stone and M. Alderton, *Mol. Phys.*, 2002, **100**, 221–233.

[126] R. F. W. Bader, *Chem. Rev.*, 1991, **91**, 893–928.

[127] R. S. Mulliken, *J. Chem. Phys.*, 1955, **23**, 1833–1840.

[128] F. L. Hirshfeld, *Theoret. Chim. Acta*, 1977, **44**, 129–138.

[129] T. C. Lillestolen and R. J. Wheatley, *Chem. Commun.*, 2008, 5909.

[130] T. C. Lillestolen and R. J. Wheatley, *J. Chem. Phys.*, 2009, **131**, 144101.

[131] T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck, M. Waroquier and P. W. Ayers, *J. Chem. Theory Comput.*, 2016, **12**, 3894–3912.

[132] C. A. Hunter and J. K. M. Sanders, *J. Am. Chem. Soc.*, 1990, **112**, 5525–5534.

[133] J. G. Vinter, *J. Comput. Aided. Mol. Des.*, 1994, **8**, 653–668.

[134] P. Ren and J. W. Ponder, *J. Comput. Chem.*, 2002, **23**, 1497–1506.

[135] P. Ren and Jay W. Ponder, *J. Phys. Chem. B*, 2003, **107**, 5933–5947.

[136] A. Grossfield, P. Ren and J. W. Ponder, *J. Am. Chem. Soc.*, 2003, **125**, 15671–15682.

[137] P. Ren and J. W. Ponder, *J. Phys. Chem. B*, 2004, **108**, 13427–13437.

[138] P. Ren, C. Wu and J. W. Ponder, *J. Chem. Theory Comput.*, 2011, **7**, 3143–3161.

[139] J. E. Lennard-Jones, *Proc. R. Soc. Lond. A*, 1924, **106**, 463–477.

[140] H. A. Lorentz, *Ann. Phys.*, 1881, **248**, 127–136.

[141] P. P. Ewald, *Ann. Phys.*, 1921, **369**, 253–287.

[142] T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

[143] M. E. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, Oxford University Press, New York, NY, 2010.

[144] J. W. Gibbs, *Elementary Principles in Statistical Mechanics*, Charles Scribner's Sons, New York, NY, 1902.

[145] B. Widom, *J. Chem. Phys.*, 1963, **39**, 2808–2812.

[146] R. W. Zwanzig, *J. Chem. Phys.*, 1954, **22**, 1420–1426.

[147] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber and W. F. van Gunsteren, *Chem. Phys. Lett.*, 1994, **222**, 529–539.

[148] C. H. Bennett, *J. Comput. Phys.*, 1976, **22**, 245–268.

[149] M. R. Shirts and J. D. Chodera, *J. Chem. Phys.*, 2008, **129**, 124105.

[150] C. Jarzynski, *Phys. Rev. E*, 1997, **56**, 5018–5035.

[151] M. R. Shirts, E. Bair, G. Hooker and V. S. Pande, *Phys. Rev. Lett.*, 2003, **91**, 140601.

[152] B. P. Cossins, S. Foucher, C. M. Edge and J. W. Essex, *J. Phys. Chem. B*, 2009, **113**, 5508–5519.

[153] L. Verlet, *Phys. Rev.*, 1967, **159**, 98–103.

[154] W. C. Swope, H. C. Andersen, P. H. Berens and K. R. Wilson, *J. Chem. Phys.*, 1982, **76**, 637–649.

[155] G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.

[156] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *J. Chem. Phys.*, 1984, **81**, 3684–3690.

[157] H. C. Andersen, *J. Chem. Phys.*, 1980, **72**, 2384–2393.

[158] B. Leimkuhler and C. Matthews, *Appl. Math. Res. eXpress*, 2012, **2013**, 34–56.

[159] J. Fass, D. A. Sivak, G. E. Crooks, K. A. Beauchamp, B. Leimkuhler and J. D. Chodera, *Entropy*, 2018, **20**, 318.

[160] S. Nosé, *Mol. Phys.*, 1984, **52**, 255–268.

[161] W. G. Hoover, *Phys. Rev. A*, 1985, **31**, 1695–1697.

[162] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *J. Chem. Phys.*, 1953, **21**, 1087–1092.

[163] W. K. Hastings, *Biometrika*, 1970, **57**, 97–109.

[164] S. C. Gill, N. M. Lim, P. B. Grinaway, A. S. Rustenburg, J. Fass, G. A. Ross, J. D. Chodera and D. L. Mobley, *J. Phys. Chem. B*, 2018, **122**, 5579–5598.

[165] N. M. Lim, M. Osato, G. L. Warren and D. L. Mobley, *J. Chem. Theory Comput.*, 2020, **16**, 2778–2974.

[166] S. Sasmal, S. C. Gill, N. M. Lim and D. L. Mobley, *J. Chem. Theory Comput.*, 2020, **16**, 1854–1865.

[167] K. H. Burley, S. C. Gill, N. M. Lim and D. L. Mobley, *J. Chem. Theory Comput.*, 2019, **15**, 1848–1862.

[168] G. A. Ross, A. S. Rustenburg, P. B. Grinaway, J. Fass and J. D. Chodera, *J. Phys. Chem. B*, 2018, **122**, 5466–5486.

[169] T. D. Bergazin, I. Y. Ben-Shalom, N. M. Lim, S. C. Gill, M. K. Gilson and D. L. Mobley, *J. Comput. Aided. Mol. Des.*, 2021, **35**, 167–177.

[170] Y. Chen and B. Roux, *J. Chem. Theory Comput.*, 2015, **11**, 3919–3931.

[171] B. K. Radak, C. Chipot, D. Suh, S. Jo, W. Jiang, J. C. Phillips, K. Schulten and B. Roux, *J. Chem. Theory Comput.*, 2017, **13**, 5933–5944.

[172] T. Lelièvre, G. Stoltz and M. Rousset, *Free energy computations: a mathematical perspective*, Imperial College Press, London ; Hackensack, N.J, 2010.

[173] J. A. Wagoner and V. S. Pande, *J. Chem. Phys.*, 2012, **137**, 214105.

[174] D. A. Sivak, J. D. Chodera and G. E. Crooks, *J. Phys. Chem. B*, 2014, **118**, 6466–6474.

[175] G. E. Crooks, *J. Stat. Phys.*, 1998, **90**, 1481–1487.

[176] H. E. Bruce Macdonald, *Ph.D. thesis*, University of Southampton, Southampton, UK, 2018.

[177] P. Atkins and J. de Paula, *Atkins' Physical Chemistry*, Oxford University Press, Oxford, UK, 10th edn., 2014.

[178] H. Kokubo, J. Rösgen, D. W. Bolen and B. M. Pettitt, *Biophys. J.*, 2007, **93**, 3392–3407.

[179] P. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts and V. S. Pande, *J. Chem. Theory Comput.*, 2013, **9**, 461–469.

[180] C. J. Woods, J. Michel, M. S. Bodnarchuk, R. T. Bradshaw, G. A. Ross, C. Cave-Ayland, H. E. Bruce Macdonald, A. I. Cabedo Martinez, M. L. Samways and J. A. Graham, *ProtoMS 3.4*, 2018, http://protoms.org, (Date accessed: January 17, 2020).

[181] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.

[182] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926–935.

[183] I. S. Joung and T. E. Cheatham, *J. Phys. Chem. B*, 2008, **112**, 9020–9041.

[184] I. S. Joung and T. E. Cheatham, *J. Phys. Chem. B*, 2009, **113**, 13279–13290.

[185] J.-P. Ryckaert, G. Ciccotti and H. J. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327–341.

[186] M. Yoneya, H. J. C. Berendsen and K. Hirasawa, *Mol. Simulat.*, 1994, **13**, 395–405.

[187] S. Miyamoto and P. A. Kollman, *J. Comput. Chem.*, 1992, **13**, 952–962.

[188] A. Ben-Naim and Y. Marcus, *J. Chem. Phys.*, 1984, **81**, 2016–2027.

[189] G. S. Kell, *J. Chem. Eng. Data*, 1975, **20**, 97–105.

[190] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok and K. A. Dill, *J. Chem. Theory Comput.*, 2007, **3**, 26–41.

[191] K. A. Beauchamp, J. D. Chodera, L. Naden, M. R. Shirts, S. Martiniani, C. D. Stern, R. T. McGibbon, R. Gowers and J. Barnett, *pymbar*, 2017, https://github.com/choderalab/pymbar, (Date accessed: March 3, 2020).

[192] M. R. Shirts, *J. Chem. Theory Comput.*, 2013, **9**, 909–926.

[193] A. Wlodawer, J. Walter, R. Huber and L. Sjölin, *J. Mol. Biol.*, 1984, **180**, 301–329.

[194] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York and P. A. Kollman, *AMBER 2018*, 2018.

[195] B. W. Zhang, D. Cui, N. Matubayasi and R. M. Levy, *J. Phys. Chem. B*, 2018, **122**, 4700–4707.

[196] A. W. Milne and M. Jorge, *J. Chem. Theory Comput.*, 2019, **15**, 1065–1078.

[197] J. Hermans, A. Pathiaseril and A. Anderson, *J. Am. Chem. Soc.*, 1988, **110**, 5982–5986.

[198] D. Beglov and B. Roux, *J. Chem. Phys.*, 1994, **100**, 9050–9063.

[199] R. Pomès, E. Eisenmesser, C. B. Post and B. Roux, *J. Chem. Phys.*, 1999, **111**, 3387–3395.

[200] W. L. Jorgensen and C. Jenson, *J. Comput. Chem.*, 1998, **19**, 1179–1186.

[201] C. Vega and J. L. F. Abascal, *Phys. Chem. Chem. Phys.*, 2011, **13**, 19663.

[202] SciPy 1.0 Contributors, P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa and P. van Mulbregt, *Nat. Methods*, 2020, **17**, 261–272.

[203] L. H. Pinto, L. J. Holsinger and R. A. Lamb, *Cell*, 1992, **69**, 517–528.

[204] K. Shimbo, D. Brassard, R. Lamb and L. Pinto, *Biophys. J.*, 1996, **70**, 1335–1346.

[205] J. A. Mould, J. E. Drury, S. M. Frings, U. B. Kaupp, A. Pekosz, R. A. Lamb and L. H. Pinto, *J. Biol. Chem.*, 2000, **275**, 31038–31050.

[206] T.-I. Lin and C. Schroeder, *J. Virol.*, 2001, **75**, 3647–3656.

[207] I. V. Chizhmakov, D. C. Ogden, F. M. Geraghty, A. Hayhurst, A. Skinner, T. Betakova and A. J. Hay, *J. Physiol.*, 2003, **546**, 427–438.

[208] K. Martin and A. Heleniust, *Cell*, 1991, **67**, 117–130.

[209] K. Duff and R. Ashley, *Virology*, 1992, **190**, 485–489.

[210] C. Ma, A. L. Polishchuk, Y. Ohigashi, A. L. Stouffer, A. Schon, E. Magavern, X. Jing, J. D. Lear, E. Freire, R. A. Lamb, W. F. DeGrado and L. H. Pinto, *Proc. Natl. Acad. Sci. USA*, 2009, **106**, 12283–12288.

[211] J. L. Thomaston, N. F. Polizzi, A. Konstantinidi, J. Wang, A. Kolocouris and W. F. DeGrado, *J. Am. Chem. Soc.*, 2018, **140**, 15219–15226.

[212] A. L. Stouffer, R. Acharya, D. Salom, A. S. Levine, L. Di Costanzo, C. S. Soto, V. Tereshko, V. Nanda, S. Stayrook and W. F. DeGrado, *Nature*, 2008, **451**, 596–599.

[213] R. Liang, J. M. J. Swanson, J. J. Madsen, M. Hong, W. F. DeGrado and G. A. Voth, *Proc. Natl. Acad. Sci. USA*, 2016, **113**, E6955–E6964.

[214] J. J. Skehel, A. J. Hay and J. A. Armstrong, *J. Gen. Virol.*, 1978, **38**, 97–110.

[215] L. C. Watkins, W. F. DeGrado and G. A. Voth, *J. Am. Chem. Soc.*, 2020, **142**, 17425–17433.

[216] J. Wang, C. Ma, G. Fiorin, V. Carnevale, T. Wang, F. Hu, R. A. Lamb, L. H. Pinto, M. Hong, M. L. Klein and W. F. DeGrado, *J. Am. Chem. Soc.*, 2011, **133**, 12834–12841.

[217] R. A. Bright, M.-J. Medina, X. Xu, G. Perez-Oronoz, T. R. Wallis, X. M. Davis, L. Povinelli, N. J. Cox and A. I. Klimov, *Lancet*, 2005, **366**, 1175–1181.

[218] T. Lampejo, *Eur. J. Clin. Microbiol. Infect. Dis.*, 2020, **39**, 1201–1208.

[219] J. L. Thomaston, A. Konstantinidi, L. Liu, G. Lambrinidis, J. Tan, M. Caffrey, J. Wang, W. F. Degrado and A. Kolocouris, *Biochemistry*, 2020, **59**, 627–634.

[220] A. K. Wright, P. Batsomboon, J. Dai, I. Hung, H.-X. Zhou, G. B. Dudley and T. A. Cross, *J. Am. Chem. Soc.*, 2016, **138**, 1506–1509.

[221] A. Drakopoulos, C. Tzitzoglaki, C. Ma, K. Freudenberger, A. Hoffmann, Y. Hu, G. Gauglitz, M. Schmidtke, J. Wang and A. Kolocouris, *ACS Med. Chem. Lett.*, 2017, **8**, 145–150.

[222] A. Drakopoulos, C. Tzitzoglaki, K. McGuire, A. Hoffmann, A. Konstantinidi, D. Kolokouris, C. Ma, K. Freudenberger, J. Hutterer, G. Gauglitz, J. Wang, M. Schmidtke, D. D. Busath and A. Kolocouris, *ACS Med. Chem. Lett.*, 2018, **9**, 198–203.

[223] P. E. Aldrich, E. C. Hermann, W. E. Meier, M. Paulshock, W. W. Prichard, J. A. Synder and J. C. Watts, *J. Med. Chem.*, 1971, **14**, 535–543.

[224] Y. Abed, N. Goyette and G. Boivin, *Antimicrob. Agents Chemother.*, 2005, **49**, 556–559.

[225] A. N. Brown, J. J. McSharry, Q. Weng, E. M. Driebe, D. M. Engelthaler, K. Sheff, P. S. Keim, J. Nguyen and G. L. Drusano, *Antimicrob. Agents Chemother.*, 2010, **54**, 3442–3450.

[226] Y. Furuse, A. Suzuki and H. Oshitani, *Antimicrob. Agents Chemother.*, 2009, **53**, 4457–4463.

[227] Y. Furuse, A. Suzuki, T. Kamigaki and H. Oshitani, *Virol. J.*, 2009, **6**, 67.

[228] V. Balannik, J. Wang, Y. Ohigashi, X. Jing, E. Magavern, R. A. Lamb, W. F. DeGrado and L. H. Pinto, *Biochemistry*, 2009, **48**, 11872–11882.

[229] C. J. Dickson, B. D. Madej, Å. A. Skjevik, R. M. Betz, K. Teigen, I. R. Gould and R. C. Walker, *J. Chem. Theory Comput.*, 2014, **10**, 865–879.

[230] J. Wang, W. Wang, P. A. Kollman and D. A. Case, *J. Mol. Graph. Model.*, 2006, **25**, 247–260.

[231] S. Llabrés, J. Juárez-Jiménez, M. Masetti, R. Leiva, S. Vázquez, S. Gazzarrini, A. Moroni, A. Cavalli and F. J. Luque, *J. Am. Chem. Soc.*, 2016, **138**, 15345–15358.

[232] A. Meyder, E. Nittinger, G. Lange, R. Klein and M. Rarey, *J. Chem. Inf. Model.*, 2017, **57**, 2437–2447.

[233] E. Nittinger, N. Schneider, G. Lange and M. Rarey, *J. Chem. Inf. Model.*, 2015, **55**, 771–783.

[234] R. Fährrolfes, S. Bietz, F. Flachsenberg, A. Meyder, E. Nittinger, T. Otto, A. Volkamer and M. Rarey, *Nucleic Acids Res.*, 2017, **45**, W337–W343.

[235] K. Schöning-Stierand, K. Diedrich, R. Fährrolfes, F. Flachsenberg, A. Meyder, E. Nittinger, R. Steinegger and M. Rarey, *Nucleic Acids Res.*, 2020, **48**, W48–W53.

[236] R. F. Cracknell, D. Nicholson, N. G. Parsonage and H. Evans, *Mol. Phys.*, 1990, **71**, 931–943.

[237] J. C. Shelley and G. N. Patey, *J. Chem. Phys.*, 1995, **102**, 7656–7663.

[238] G. L. Deitrick, L. E. Scriven and H. T. Davis, *J. Chem. Phys.*, 1989, **90**, 2370–2385.

[239] M. R. Stapleton and A. Z. Panagiotopoulos, *J. Chem. Phys.*, 1990, **92**, 1285–1293.

[240] D. A. Sivak, J. D. Chodera and G. E. Crooks, *Phys. Rev. X*, 2013, **3**, 011007.

[241] D. A. Erlanson, in *Fragment-Based Drug Discovery and X-Ray Crystallography*, ed. T. G. Davies and M. Hyvönen, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, vol. 317, pp. 1–32.

[242] D. A. Erlanson, S. W. Fesik, R. E. Hubbard, W. Jahnke and H. Jhoti, *Nat. Rev. Drug Discov.*, 2016, **15**, 605–619.

[243] M. Congreve, R. Carr, C. Murray and H. Jhoti, *Drug. Discov. Today*, 2003, **8**, 876–877.

[244] H. Jhoti, G. Williams, D. C. Rees and C. W. Murray, *Nat. Rev. Drug Discov.*, 2013, **12**, 644–644.

[245] P. Kirsch, A. M. Hartman, A. K. H. Hirsch and M. Empting, *Molecules*, 2019, **24**, 4309.

[246] M. Bissaro, M. Sturlese and S. Moro, *Drug. Discov. Today*, 2020, **25**, 1693–1701.

[247] O. Guvench and A. D. MacKerell, *PLoS Comput. Biol.*, 2009, **5**, e1000435.

[248] D. Alvarez-Garcia and X. Barril, *J. Med. Chem.*, 2014, **57**, 8530–8539.

[249] V. Oleinikovas, G. Saladino, B. P. Cossins and F. L. Gervasio, *J. Am. Chem. Soc.*, 2016, **138**, 14257–14263.

[250] S. C. Gill and D. L. Mobley, *J. Chem. Theory Comput.*, 2021, **17**, 302–314.

[251] M. Clark, F. Guarnieri, I. Shkurko and J. Wiseman, *J. Chem. Inf. Model.*, 2006, **46**, 231–242.

[252] M. Clark, S. Meshkat and J. S. Wiseman, *J. Chem. Inf. Model.*, 2009, **49**, 934–943.

[253] S. K. Lakkaraju, E. P. Raman, W. Yu and A. D. MacKerell, *J. Chem. Theory Comput.*, 2014, **10**, 2281–2290.

[254] P. R. Van Tassel, H. T. Davis and A. V. McCormick, *Langmuir*, 1994, **10**, 1257–1267.

[255] R. F. Cracknell, D. Nicholson, S. R. Tennison and J. Bromhead, *Adsorption*, 1996, **2**, 193–203.

[256] S. R. Challa, D. S. Sholl and J. K. Johnson, *J. Chem. Phys.*, 2002, **116**, 814–824.

[257] N. A. Mahynski, J. R. Errington and V. K. Shen, *J. Chem. Phys.*, 2017, **147**, 234111.

[258] X. Wang, G. Minasov and B. K. Shoichet, *J. Mol. Biol.*, 2002, **320**, 85–95.

[259] J. R. Horn and B. K. Shoichet, *J. Mol. Biol.*, 2004, **336**, 1283–1291.

[260] F. Comitani and F. L. Gervasio, *J. Chem. Theory Comput.*, 2018, **14**, 3321–3331.

[261] V. Le Guilloux, P. Schmidtke and P. Tuffery, *BMC Bioinformatics*, 2009, **10**, 168.

[262] Z. Tan, *J. Comput. Graph. Stat.*, 2017, **26**, 54–65.

[263] P. B. Grinaway, J. M. Behr, H. E. Bruce Macdonald, D. A. Rufa and J. D. Chodera, *perses*, 2020.

[264] D. A. Rufa, H. E. Bruce Macdonald, J. Fass, M. Wieder, P. B. Grinaway, A. E. Roitberg, O. Isayev and J. D. Chodera, *Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning / molecular mechanics potentials*, biorxiv preprint, 2020.

[265] C. Cave-Ayland, C.-K. Skylaris and J. W. Essex, *J. Chem. Theory Comput.*, 2017, **13**, 415–424.

[266] K. S. Schmitz, in *Physical Chemistry*, Elsevier, 2017, pp. 559–632.