University of Southampton

Faculty of Engineering and Physical Sciences

Thesis for the degree of
Doctor of Philosophy

The role of water in drug binding: Calculating positions and binding free energies of active site water molecules, and their influence on ligand binding

Author: Hannah Bruce Macdonald

Supervisor: Prof. Jonathan Essex

September 2018

UNIVERSITY OF SOUTHAMPTON ABSTRACT
FACULTY OF ENGINEERING AND PHYSICAL SCIENCES

Chemistry Doctor of Philosophy

The role of water in drug binding: Calculating positions and binding free energies of active site water molecules, and their influence on ligand binding by Hannah Bruce Macdonald

This thesis studies the ability of computer simulation to determine the location and free energy of binding of active site water molecules, and the energetic effect water molecules can have on ligand binding. The primary method used involves sampling within the grand canonical ensemble, using grand canonical Monte Carlo (GCMC).

The first results chapter looks at the introduction of replica exchange (RE) to GCMC simulations, and the improvements this yields in the reliability of calculated water binding free energies. The results show that GCMC can determine water binding free energies that are consistent with double-decoupling methods, while being able to calculate multiple water free energies simultaneously, without a priori knowledge of water locations.

The second chapter explores the accuracy of GCMC at determining the locations of active site water molecules, using a large dataset of molecules and targets of pharmaceutical interest. Understanding the accuracy of GCMC to reproduce crystallographic water locations allows for reliable calculation of protein-ligand complexes without experimentally known water locations being known. Focus will be placed on the variation of quoted water placement success rates with different published protocols.

The final chapter of this thesis involves the integration of two techniques; GCMC and ligand alchemical perturbation simulations. Grand canonical Alchemical Perturbations (GCAP) will be presented, whereby relative binding free energies of pairs of ligands are calculated, while active site water molecules are sampled using the grand canonical ensemble. This GC sampling of water allows the ligands water network to dynamically adapt. GCAP will be demonstrated for two example systems, where active site water molecules are a key factor in the ligand binding affinities.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Jonathan Essex for the continuous support throughout my Ph.D, for the guidance, encouragement and advice that he has imparted during my time in his group. I am extremely lucky to have a supervisor who started me on such an interesting project, and allowed me to pursue the research that interests me most. He has supported me to attend many conferences, and collaborate with multiple groups, which has not only benefitted me for these past three years, but will help set me up for future too.

I would like to thank all of the members of the TMCS CDT, who have been there from the first day in Oxford and it has been a pleasure to share this experience with. Thank you to the Engineering and Physical Sciences Research Council for funding to carry out my research.

I would like to thank the whole of the Essex Group past and present for their help throughout my PhD work. I have learnt so much from all of you and for that I am extremely grateful. Dr. Greg Ross for getting me started on such a great thesis project and Marley Samways for his contribution to the work presented in Chapter 3. Particular thanks to Dr. Chris Cave-Ayland, whose help and opinion has had an impact on every page of this thesis, thank you for always helping me no matter how stupid the question, or busy you are. The whole of the Essex group, and the other computational chemistry groups have made these years fun. Thank you also to Dr. Richard Taylor for showing such enthusiasm in our research, and having me join UCB for a few months — it was a great experience and thank you also to everyone I met at UCB during the time.

Thank you to Georgia, Cathy, Scarlet, Emily, Erin, Roxy, Aleks, Rachel, Steph, Sam, Claire, and to all the other women who I am extremely lucky to call my friends. They have always provided me with both sanctuary and sanity. I owe a huge thank you to my partner, Nick Lees, whom I watched go through all this, while not fully appreciating how hard the final stage of this was — I am sorry. He

has supported me throughout with calmness and advice, no matter how difficult I have found it. Finally, thank you for being such a huge motivation — if you hadn't have moved to a different continent, it probably would have taken me a year longer to finish. I cannot wait to join you.

I would like to thank my grandparents; Frankie and John, and also to my other grandparents, Syb and Bill, who did not get to see me finish. Finally, and most importantly, I would like to dedicate this to my parents who stopped understanding what I do about 10 years ago, but have supported me unwaveringly throughout. To you I owe everything.

Chapter 1

Introduction

Many cases are known where water molecules are known to directly influence ligand binding affinity. For example, OppA is a non-specific tri-peptide binder, capable of binding to a class of ligands of the structure Lys-X-Lys, where X is any of the natural amino acids.[1] The active site of OppA is therefore capable of binding ligands of a range of sizes and properties. This promiscuity is made possible by a varying network of water molecules occupying the active site volume around the central amino acid, confirmed by X-ray crystallography studies.[1] Binding data of the ligand class shows that displacement of water molecules in the active site corresponds to a decreasing binding affinity. In contrast, Scytalone Dehydratase (SD) is a fungicidal protein-target for rice-blast disease in crops.[2] The SD enzyme catalyses two steps in the pathogenic fungus, *M. grisea*, in its biosynthesis of melanin. Melanin is required by the fungus for its structural integrity, without which cell penetration, which is required in its mechanism of infection is not possible. A range of ligands are known to bind to the protein, causing inhibition of the enzyme and disrupting the melanin pathway. Design of high-affinity ligands for this system has focussed on displacing known active site water molecules, and displacement of one particular water molecule can increase ligand affinity 100-30,000 fold.[3] These two cases illustrate the lack of consistency that arises, whereby in OppA, disruption of a water network weakens binding while in SD, the displacement of a water has the opposite effect. Quantifying these changes in affinity are of significant importance in drug design.

Various protein-ligand systems will be studied in this thesis. SD will be used in Chapters 2 and 4. As it is a single water system, where displacement of the said water molecule has a large impact on ligand binding affinity, it is a useful test system. Bovine pancreatic trypsin inhibitor (BPTI) will also be used as a test system in Chapter 2. BPTI is useful as it is a small protein that has a small pocket that ligands do not bind, but contains three water molecules. This is useful for empirically testing the effect of multiple water binding and water network effects. Adenosine A_{2A}receptor is a membrane protein, which has a dataset of binding affinity data for twelve related ligands.[4] This is an interesting system, as the two associated crystallographic structures are low resolution (3UZA: 3.273 Å, 3UZC:

3.341 Å) where no water locations are resolved. This shows how grand canonical (GC) methods can be advantageous for a case where the experimental data is of low quality. A_{2A}will be used in Chapter 4 to demonstrate GCAP where multiple water molecules are displaced by ligand perturbation. Chapter 3 will present analysis of the hydration of a dataset of 105 protein-ligand complexes, and demonstrate a large scale test of GCMC.

This thesis uses GCMC methodologies to try reproduce experimental water locations, and ligand binding affinities. If GCMC is shown to be reliable at reproducing known experimental results, it can in future be applied to novel target systems with confidence. GCMC will be used first, in the calculation of binding free energies of active site water molecules; second, in the determination of the locations of active site water molecules, and finally, for the calculation of ligand binding free energies, in cases where water molecules are displaced resulting in changes in affinity. GCMC involves simulating in the grand canonical ensemble, the μVT ensemble, where μ is chemical potential, V is volume and T is temperature. This involves the fluctuation of N (the number of atoms or molecules) within the simulation through insertion and deletion Monte Carlo moves. The molecules allowed to insert and delete in the protein-ligand systems studied in this thesis are water molecules, with insertion and deletion moves attempted within a certain user defined region of a protein-ligand complex. Insertion and deletion of water molecules allows for the location of active site hydration sites to be predicted, as well as their binding free energy calculated through using the grand canonical integration (GCI) equation. GCMC is beneficial over other water location methods as it is able to calculate multiple waters simultaneously i.e. a network of water molecules, without prior knowledge of where the waters are located, while providing binding affinities consistent with double decoupling (DD) simulations. The theoretical basis and computational methodology of GCMC will be discussed in Section 1.4.

Several methods are capable of locating and classifying water molecules in a system, but this is only the starting point from a pharmaceutical perspective.

Primarily, the binding affinity of a ligand is of interest, which can be calculated by free energy methods, which are discussed in Section 1.2. These methods can be used to determine the binding free energy of a ligand once the correct surrounding water structure is known. Computationally, binding is often handled in relative terms, and when comparing binding of two ligands with differing water structures convergence issues can arise, solved only by lengthy simulations and associated thermodynamic cycles.[5] The method of grand canonical Alchemical Perturbation (GCAP) is able to avoid this, using free energy methods between two ligands, while simultaneously optimising their respective water networks through the water-location method of GCMC. GCAP will be introduced in Section 1.4.

1.1 Computational methods

Force fields

A force field is a set of parameters and a functional form that have been designed to reproduce known properties of a system — either experimental values, or properties determined from a higher level of computational theory. Force fields are often designed for a specific use, i.e. Amber forcefields[6] for proteins, and the general Amber force field (GAFF) for small organic molecules.[7] Herein, fixed-charge all-atom force fields will be considered, however various coarse-grained and united-atom models also exist. The functional form of the force field consists of both the bonded and the non-bonded parameters. The bonded energy is determined between covalently bonded ligands, through bond, angle and dihedral terms, while the non-bonded terms are calculated for non-bonding atoms by considering the electrostatic and van der Waals forces between atoms, Equation 1.1c.

$$E_{total} = E_{bonded} + E_{non-bonded} \tag{1.1a}$$

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral} \tag{1.1b}$$

$$E_{non-bonded} = E_{ele} + E_{vdW} \tag{1.1c}$$

The bonded terms (E_{bonded}) are calculated using the following;

$$E_{bond} = k_{bond}(r - r^o)^2 \tag{1.2a}$$

$$E_{angle} = k_{angle}(\theta - \theta^o)^2 \tag{1.2b}$$

$$E_{dihedral} = \sum_{i=1}^{n} k_i[1 + k_j(cos(k_k\phi + k_l))] \tag{1.2c}$$

both the bond and angle terms take the same form of a harmonic potential, whereby the difference between a bond or angle (r or θ) to a minimum value (r^o or θ^o) with a bond or angle strength (k_{bond} or k_{angle}). The form of the dihedral energy is calculated is using the dihedral angle ϕ, and a set of dihedral parameters, k_{i-l}. The bonded energy terms account for the energetic interactions of covalently bonded atoms that are one, two or three bonds distance. For pairs of atoms that are separated by four or more bonds, or not covalently linked, non-bonded energies are calculated. These consist of the electrostatic terms; which are calculated using the Coulomb equation, and intermolecular electron dispersion forces are calculated using the Lennard-Jones potential;[8]

$$E_{ele} = \frac{q_i q_j}{4\pi\epsilon_o r} \tag{1.3a}$$

$$E_{vdW} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^m - \left(\frac{\sigma_{ij}}{r}\right)^n\right] \tag{1.3b}$$

The Coulomb equation is shown in Equation 1.3a, where the electrostatic interaction between two atoms at distance r can be calculated using their respective charges (q_i and q_j), where ϵ_o is the permittivity of free space. The Lennard-Jones m-n potential, Equation 1.3b, is a pairwise approximation of many-body interactions that would be computationally prohibitive to calculate directly. Many other forms of pair potentials exist,[9,10] of which the Lennard-Jones 12-6 is the most common. Both σ and ϵ, the collision radius and well-depth respectively, are empirically determined parameters for an atom, which for a pair of atoms (σ_{ij} and ϵ_{ij}) are calculated using arithmetic combining for σ and geometric combining for

ϵ;

$$\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j) \tag{1.4a}$$

$$\epsilon_{ij} = (\epsilon_i \epsilon_j)^{\frac{1}{2}} \tag{1.4b}$$

these are the combining rules used in Amber, and other force fields use differing combining rules.

In a system of more than a few atoms, the number of non-bonded interactions will quickly outnumber the number of bonded interactions. To reduce this expense, non-bonded interactions may be truncated by only calculating them for atoms, or groups of atoms within a given cutoff distance, (r_{cut}). To smooth the non-bonded interaction energy, the energy is scaled for some region ($r_{feather}$).[11] Two atoms at distance r will be scaled accordingly;

$$E_{non-bonded} = scale(r)E_{non-bonded} \tag{1.5a}$$

$$r > r_{cut} \rightarrow scale(r) = 0.0 \tag{1.5b}$$

$$r_{cut} - r_{feather} < r < r_{cut} \rightarrow scale(r) = \frac{r_{cut}^2 - r^2}{r_{cut}^2 - (r_{cut} - r_{feather})^2} \tag{1.5c}$$

$$r < r_{cut} - r_{feather} \rightarrow scale(r) = 1.0 \tag{1.5d}$$

as $scale(r)$ is 0.0 where $r > r_{cut}$, these energies do not need to be evaluated.

Statistical mechanics

A force field allows for the energy of a state of a system to be calculated. Statistical mechanics is able to relate details of all states of a given system to macroscopic properties. All possible states, or replicas of a given system is known as an ensemble of states, whereby the type of ensemble is defined by the properties that are constant between all replicas of states. The canonical ensemble (NVT) is where the number of atoms (N), the volume (V) and the temperature (T) are consistent;

the microcanonical ensemble (NVE) where E is energy, and the ensemble that is exploited within this thesis; the grand canonical ensemble (μVT). The canonical ensemble allows for calculation of the Helmholtz free energy (A) of a system;[11]

$$A = -k_B T ln(Q) \tag{1.6}$$

where k_B is the Boltzmann constant, and Q is the partition function. Different ensembles provide different types of free energy. The partition function is the sum of the energies of all microstates in the ensemble;

$$Q = \sum_i e^{-\frac{E_i}{k_B T}} \tag{1.7}$$

where E_i is the energy of the i^{th} microstate. In the classical limit, the canonical partition function of N atoms, can be treated as an integral over all states,

$$Q = \frac{1}{h^{3N} N!} \int_r e^{\frac{E_i(r)}{k_B T}} \quad \mathrm{d}r \tag{1.8}$$

where h is Planck's constant. The $\frac{1}{N!}$ term removes the overcounting of microstates which are fundamentally the same, but differ only in the exchange of identical atoms with differing labels. this can be substituted into Equation 1.6;

$$A = -k_B T ln \left(\frac{1}{h^{3N} N!} \int_r e^{-\frac{E_i(r)}{k_B T}} \mathrm{d}r \right) \tag{1.9}$$

this leads to a result where the free energy of a system can be determined from the ensemble of a canonical system.

Free energy

The following section leads to the resulting equation, Equation 1.9, where the absolute Helmholtz free energy of a canonical ensemble can be calculated from potential energy of each microstate of the system. However, this cannot be solved for large systems for two reasons; firstly the number of microstates of the system that must be integrated over is prohibitively large, and the $e^{E_i(p)}$ term results in poor numerical behaviour.

The issue of the large number of microstates will first be addressed using the Boltzmann-weighted distribution of the phase space, and the issue of the numerical behaviour of $e^{E_i(p)}$ will be reduced by considering relative free energy calculations.[11]

All of the j microstates of a system at temperature, T, will follow the Boltzmann distribution;

$$P_i = \frac{e^{-E_i\beta}}{\int_j e^{-E_j\beta}} \tag{1.10}$$

Where the probability of the system being in microstate i is P_i, where E_x is the energy of microstate x, and $\beta = \frac{1}{k_B T}$ is thermodynamic beta. The Boltzmann distribution means that only a subsection of microstates will contribute significantly to the ensemble observables. This means that integral over all microstates can be simplified to simply summing over the important microstates of a system — which is the states that are proximal to the minima and therefore contributing to the denominator. The integral over all states can be replaced with a sum over all states — or in practise a sum over contributing states.

$$P_i = \frac{e^{-E_i\beta}}{\sum_j e^{-E_j\beta}} \tag{1.11}$$

An average property of the system can be calculated from the Boltzmann distribution of states, using the following;

$$\langle X \rangle = \frac{\int_r X(r)e^{-\beta E_i(r)}\mathrm{d}r}{\int_r e^{-\beta E_i(r)}\mathrm{d}r} \tag{1.12a}$$

$$\langle X \rangle = \int_r X(r)P_i(r) \tag{1.12b}$$

This can be combined with Helmholtz free energy equation, Equation 1.9, where the numerator has been multiplied by $e^{-\beta E_r(r)}e^{\beta E_r(r)} = 1$ to give;

$$
\begin{aligned}
A &= k_B T ln \left(\frac{\int_r e^{-\beta E_r(r)} e^{\beta E_r(r)}}{\int_r e^{\frac{-E_r(r)}{k_B T}}} \mathrm{d}r \right) && \text{(1.13a)} \\
&= k_B T ln \left(\int_r P_i(r) e^{\beta E_r(r)} \right) && \text{(1.13b)} \\
&= k_B T ln \left(< e^{\beta E_r(r)} > \right) && \text{(1.13c)}
\end{aligned}
$$

This results in an equation whereby the Helmholtz free energy can be determined from the average potential energy of observed microstates, rather than the entire ensemble. This allows for thermodynamic results to be calculated from a sampling regime, which will be discussed in Sections 1.1 and 1.1. The result calculates the absolute free energy of a system, and there is still the issue of the e^E term whereby the result will be unstable with addition of additional microstates as additional states will cause large flucctuations to the free energy, and is only stable for small systems with a small configurational phase space that can be sampled adequately. Absolute free energies are not viable for biomolecular systems.

Relative free energies

To circumvent the issue of large energetic terms, relative free energies can be calculated.[12]

$$
\Delta A_{AB} = A_B - A_A \tag{1.14}
$$

where A is the Helmholtz free energy, calculated from the NVT ensemble.

$$\begin{aligned}
\Delta A_{AB} &= -k_B T ln \left(\frac{Q_B}{Q_A} \right) && (1.15a)\\
&= -k_B T ln \left(\frac{\int_r e^{-\beta E_B(r)}}{\int_r e^{-\beta E_A(r)}} \right) && (1.15b)\\
&= -k_B T ln \left(\frac{\int_r e^{-\beta E_B(r)} e^{-\beta E_A(r)} e^{\beta E_A(r)}}{\int_r e^{-\beta E_A(r)}} \right) && (1.15c)\\
&= -k_B T ln \left(P_A(r) \int_r e^{-\beta(E_B(r)-E_A(r))} \right) && (1.15d)\\
&= -k_B T ln \left(\left\langle e^{-\beta \Delta E_{AB}(r)} \right\rangle \right) && (1.15e)
\end{aligned}$$

Where ΔE_{AB} is the difference in energy of a microstate of system A in systems A and B. This quantity will be smaller than the absolute energy, E_A, and means that the result can be evaluated.

With Equation 1.15e we now have a method of calculating the relative free energy of a system, by calculating the energy difference between two systems, for thermally significant states of the systems. To generate thermally significant states, sampling methods will be used. States of a system can be generated using either molecular dynamics or Monte Carlo simulations.

Molecular dynamics

Molecular dynamics (MD) simulation is the method of studying atomic systems following the equations of classical dynamics.[13] Newton's equations of motions are solved repeatedly over short time steps.

$$\begin{aligned}
\mathbf{F} &= m\mathbf{a} && (1.16a)\\
\mathbf{F} &= \frac{dv}{d\mathbf{r}} && (1.16b)
\end{aligned}$$

The positions and velocities of all the particles in a system are all coupled, which results in a many-body problem, meaning that the equations of motion need to be integrated using a finite difference method, rather than being solved analytically. The most common methods of integrating the equations of motions is using either the Verlet[14] or the velocity Verlet algorithm. The Verlet algorithm is derived by first approximating the positions and momenta using a Taylor series expansion;

$$\mathbf{x}(t+\delta t) = \mathbf{r}(t) + \delta t \tag{1.17}$$

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) + \mathcal{O}(\delta t^3)... \tag{1.18a}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \delta t\mathbf{a}(t) + \mathcal{O}(\delta t^3)... \tag{1.18b}$$

for the positions (r), velocity (v) and acceleration (a) at time (t) and an incrementally small time after t, $t + \delta t$. The Verlet algorithm provides the positions at time $t + \delta t$ using;

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2\mathbf{a}(t) \tag{1.19}$$

which is dependent on the coordinates, the coordinates at the previous time step and the acceleration. The velocity can also be determined by dividing the difference in positions at $t + \delta t$ and $t - \delta t$. Other methods that also integrate the equations of motions exist.[15,16] Updating the positions and velocities of the atoms for a sufficient number of timesteps, allows the motion of the system for a time Xδt, where X is the number of iterations that are performed. If the time is sufficient, behaviours and properties of the system can be studied. From a sufficiently long MD simulation, it is possible to calculate system properties that are a function of atom coordinates and momenta. As MD simulations follow a time trajectory, they are useful for understanding diffusion motions of systems, and other time-dependent properties. However, even for reasonable numbers of atoms, many computing hours are required to achieve simulation timescales on the order of nanoseconds. MD, with the use of thermostats, is able to generate a set of microstates, which can be used with Equation 1.15e.

Monte Carlo

Monte Carlo is an alternate method for generating states of a systems, whereby instead of a time-evolving set of states being generated as by MD, states are generated by making random changes to the system.[11] The potential energy of each sampled state can be generated from the positions of atoms in the system, but as the 'motions' are randomly generated, there is no momenta component. The energy of a microstate is the sum of the potential energy, which is a function of the atomistic coordinates, and the kinetic energy, a property of the atomistic momenta. This allows the partition function to be decomposed into two parts — coordinates and momentum;

$$Q = \frac{1}{h^{3N}N!}\int_r\int_q e^{-\frac{E_r(p)}{k_BT}} e^{-\frac{E_k(q)}{k_BT}} \mathrm{d}p\mathrm{d}q \tag{1.20}$$

where E_r and E_k are the potential and kinetic energy respectively, which are in turn, a function of the coordinates, r, and momenta, p, of the system. These are separable;

$$Q = \frac{1}{h^{3N}N!}\int_r e^{-\frac{E_r(r)}{k_BT}} \mathrm{d}r \int_p e^{-\frac{E_k(p)}{k_BT}} \mathrm{d}p \tag{1.21a}$$

$$Q = Q_rQ_p \tag{1.21b}$$

The kinetic partition function (Q_p);[17]

$$Q_p = \frac{V^N}{N!\Lambda^{3N}} \tag{1.22}$$

where $\Lambda = \sqrt{h^2/2\pi k_BTm}$, and m is atomic mass. The kinetic contribution to the partition function is the partition fucntion of an ideal gas and is analytically solvable. The potential energy part of the partition function can now be considered;

$$Q_r = \int_r e^{\frac{-E_r(r)}{k_BT}} \mathrm{d}r \tag{1.23}$$

where when used in Equation 1.9, allows the Helmholtz free energy to be calculated from the potential energy of the generated microstates.

States are generated by making random changes to the system of interest, with much freedom with regards to the type of change that can be made. One example of this are the insertion and deletion moves of GCMC that will be discussed in much more detail. A MC move consists of making a random change, and assessing this change based on the energetic difference between the system and the trial system. Either the previous microstate, or the trial microstate will be accepted or rejected into the ensemble of states, based on the Metropolic criteria. If the energy of the trial state is lower than the previous, the trial move is accepted, and this configuration is then used as the starting point for the following step. However, if the trial configuration is higher in energy than the previous, the move is accepted if a randomly generated number between 0 and 1 is smaller than the Boltzmann factor;

$$rand(0,1) \leq e^{-\beta\Delta E} \tag{1.24}$$

this ensures that the states generated follow the correct Boltzmann distribution of states.

In terms of the moves that can be made, these are generally applied to the translation and rotation of atoms, functional groups or molecules. The number of atoms moved, and the magnitude of the change will increase the likelihood of ΔE being large, and therefore reduce the likelihood of the move being accepted. On the other hand, if the moves are very small, then the majority of moves will be accepted, but the states generated are likely to be very similar. The move size is often optimised such that approximately 50% of moves are accepted.

Metropolis sampling requires that the condition of detailed balance to hold for systems in equilibrium.[18,19] This states that the step from configuration a to configuration b should be equally likely as the step from b to a, and should hold for all configurations of a system. The likelihood of moving from configuration a to configuration b (referred to as the flow, $\kappa(a \rightarrow b)$) is a product of the likelihood of being in state a ($\mathcal{N}(a)$), the likelihood of the move to state B being proposed ($\alpha(a \rightarrow b)$), and finally the likelihood of accepting the proposed move, $acc(a \rightarrow b)$;

$$\kappa(a \rightarrow b) = \kappa(b \rightarrow a) \tag{1.25a}$$

$$\kappa(a \rightarrow b) = \mathcal{N}(a)\alpha(a \rightarrow b)acc(a \rightarrow b) \tag{1.25b}$$

where in the canonical ensemble, the likelihood of being in configuration a is the Boltzmann factor; $e^{-\beta E(r)}$. If all random moves proposed are equally likely, then $\alpha(a \rightarrow b) = \alpha(b \rightarrow a)$, and the appropriate acceptance rates can be derived;

$$\frac{acc(a \rightarrow b)}{acc(b \rightarrow a)} = \frac{e^{-\beta E_b(r)}}{e^{-\beta E_a(r)}} \tag{1.26}$$

which leads to the Metropolis acceptance criteria;

$$acc(a \rightarrow b) = min\left[1, e^{-\beta(E_b(r)-E_a(r))}\right] \tag{1.27}$$

1.2 Free energy calculations

The previous section refers to a and b, which are two microstates of the same state. Here, we would like to compare the free energy of two different states, A and B, using Equation 1.15e. This involves sampling in state A, following the Boltzmann distribution of states, that MC sampling upholds, and evaluating the difference in energy between each microstate of A, for both state A and B, $\Delta E_{AB}(r)$. Integrating between these two states can be performed with various rigorous free energy methods; including thermodynamic integration (TI),[20] Bennett Acceptance Ratio (BAR),[21] and Multistate BAR (MBAR)[22] methods.

Accuracy of free energy methods require the sampling of state A, to reflect the Boltzmann distribution of state B. To improve overlap between the two states, intermediate states can be introduced such to bridge the difference. The collection of intermediate states are referred to as a a λ coordinate, where λ is a coupling parameter. States A and B refer to the λ end points, 0 and 1, while for intermediate states a fictitious potential is defined. One common definition;

$$U(\lambda) = (1-\lambda)U_A + \lambda U_B. \tag{1.28}$$

where $U(\lambda)$ is the fictitious potential at the λ intermediate. The free energy at each λ value can be determined from the partition function of an ensemble at that λ using Equation 1.29.

$$Q(N,V,T,\lambda) = \sum_i e^{-\beta U_i(\lambda)} \tag{1.29}$$

where i are the microstates of the λ ensemble. The average derivative of the potential energy with respect to each λ value can be used to compute the integral over each of these derivatives. forming a path between $\lambda = 0$ and 1, i.e. states A and B, Equation 1.30.

$$\Delta F(A \rightarrow B) = \int_0^1 \frac{\delta F(\lambda)}{\delta\lambda} d\lambda = \int_0^1 \left\langle \frac{\delta U(\lambda)}{\delta\lambda} \right\rangle_\lambda d\lambda \tag{1.30}$$

Computationally, this is often performed using parallelised code, where individual λ values of a free energy path will be simulated on individual processors of a node, commonly 12 or 16 λ windows. As the method relies on overlapping phase space, errors can occur if the λ coupling between the two states is not sufficient. RE can be applied to TI to reduce the effect of limited sampling.[23] The general method of RE is when swaps are attempted between multiple replicas of a system, where each replica differs by a given property. Attempts to swap replicas are made and swaps are accepted or rejected following acceptance criteria.[24,25] The method has been successfully applied to simulations of various properties such as temperature[26] and pH.[27] The application of RE to TI by attempting swaps between neighbouring λ windows allows simulations to share coordinates of trajectories, optimising sampling between the states and removing errors from the integration path. RE methods have been applied to GCMC methods in this thesis, and the benefits of this will be discussed in Chapter 2.

BAR is an alternate method of determining the free energy difference between two states.[21] BAR calculates the difference in free energy between states A and B, with the assumption that the two systems share all their microstates. With this assumption, complete sampling of the phase space of state A, with the Hamiltonian of A should cover the entire phase space of state B. Reversibly, A should be sampled by the simulation of state B with the Hamiltonian of B. The difference in free energy between the two states, will be the ratio of probabilities of sampling one state from the system of the other, Equation 1.31

$$e^{-\beta(\Delta F - C)} = \frac{\langle f(\beta(U_B - U_A - C)) \rangle_A}{\langle f(\beta(U_A - U_B - C)) \rangle_B} \tag{1.31}$$

where ΔF is the free energy difference between the two states, $\langle U_x \rangle_y$ is the potential energy U evaluted using the Hamiltonian of $_x$, while sampling in the ensemble of y. Equation 1.31 will hold for any function, f, that also meets the detailed balance condition. In practise $f(x) = \frac{1}{1+e^x}$ is used as it is the optimal solution. C is an energy offset between the two systems, i.e. the value of interest, ΔF. This requires the equation to be solved iteratively.[13] Iterations will only converge if there is sufficient overlap between the two states. Multistate BAR (MBAR) is a derivative of BAR, where all intermediate λ states are considered in the calculation, rather than only neighbouring states.[22] MBAR has been shown to be the most statistically efficient method to abstract free energy differences from simulations[28] .

Single and dual topology calculations

For calculations of free energies of systems, thermodynamic cycles are often required to calculate the energies of interest. For example, to calculate the relative solvation free energy of two molecules, the free energy difference between the two species is required both in solution and in the gas phase.[29]

$$\begin{tikzcd}
A_{gas} \arrow[r, "\Delta G_2"] \arrow[d, "\Delta G^{A}{}_{solv}"'] & B_{gas} \arrow[d, "\Delta G^{B}{}_{solv}"] \\
A_{sol} \arrow[r, "\Delta G_1"'] & B_{sol}
\end{tikzcd}$$

Figure 1.1: Free energy cycle. Vertical legs are the solvation free energy of A and B respectively, the difference of which can be determined by computational calculation of the horizontal legs, the free-energy calculation of peturbing A to B in the gas and solution phase.

Shown in Figure 1.1, the difference between the two vertical legs is the relative solvation free energies of the species, while the horizontal legs are the perturbation between the two species when solvated, and in the gas phase. The direct calculation of the solvation free energy of each individual molecule is prohibitively difficult, as the overlap of the phase space between a molecule in gas phase and in solution is too poor to use the methods outlined above. Assuming the two molecules are sufficiently similar, their phase space within a given environment should overlap, allowing the computational methods of free energy calculation to be possible. The two alchemical transformations are equal to the two solvation free energies, $\Delta G^{A}_{solv} - \Delta G^{B}_{solv} = \Delta G_2 - \Delta G_1$. Thermodynamic cycles are also used for the calculation of molecular association, where gas and solv instead represent bound and unbound complexes. The calculation of the free energy between two states, requires a pathway to exist between the two states through the λ coupling parameter outlined above. The pathway taken between the two states is known as the molecular mechanical topology. There are two main approaches to this, known as single topology and dual topology, illustrated in Figure 1.2.

Illustrations of both single and dual topology are shown in Figure 1.2. In dual topology, the pathway between the two states is generated by retaining two in-

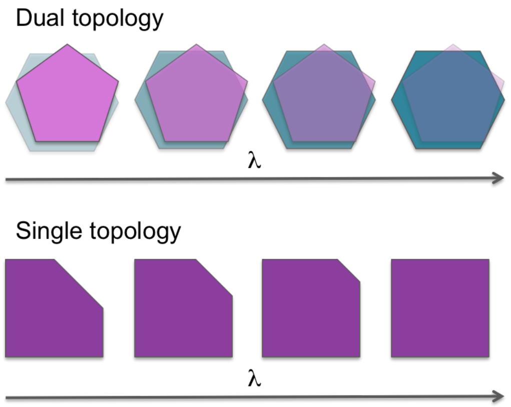


Figure 1.2: Two protocols for alchemical perturbations; dual topology and single topology. Dual topology contains two ligands, while one is decoupled from the system across λ, the other has its interactions turned on. Single topology only contains one ligand, which is geometrically and electrostatically altered between the two ligands considered.

dependent topologies of each state (A and B), both of which are present in the calculation. Each state does not interact with the other, but interacts with its environment with an energy scaled by λ. At each λ value, two topologies exist, with state A interacting with its environment with a value of λ, and B, a value of (1-λ). At $\lambda = 0$ and 1, only one state will be 'on' and the other 'off'.[30] Single topology differs from this as only one independent topology exists at intermediate states generated by λ scaling of the force field and geometric parameters. A molecular geometry at each λ value is required for the simulation, which at intermediate λ values will refer to an alchemical molecular structure. This involves the mapping of the two structures onto each other, and changing differing bond lengths and atom types across the λ path. Where an atom is not present in the map of the other, dummy atoms are required. When an atom is perturbed to a dummy state, the bond to the dummy atom is retracted. The determination of the state of the hybrid topology at intermediate λ values can sometimes be non-obvious for two states, and the accuracy of the result may vary depending on the protocol, if the two end points are not clearly defined.[29] When an atom is perturbed to a dummy state, the bond to the dummy atom is retracted, such that the dummy atom is within the vdW radius of the bound atom. If the protocol differs in end states, the free energy will be dependent on the method of shrinking/disappearing atoms or groups,[31, 32] requiring a bond length pseudo-potential of mean force correction to ensure the free energy is independent to the shrinking of these groups. This issue only arises when the dummy-bond is sampled within each λ window. As bond lengths are not sampled within the ProtoMS software, this is not a consideration for these results. Any contributions due to a choice in single-topology protocol should cancel directly between the two legs of the calculation, i.e. solvation and bound legs.

At the end values of the λ pathway, both single and dual topology may have either created or annihilated of atoms in the system. If a molecule is 'off' it has no interaction with its environment and is able to overlap with surrounding atoms. If an atom is then turned on from this position of overlap, the energy will be infinite, even if the interaction of the group is scaled to be very small at the neighbouring λ value. This is known as a singularity problem, occurring due to the r^{-12} repulsion term in the Lennard-Jones equation. Soft-core potential functions are able to stop infinite energies by removing the points of singularity and ensuring the energies are finite in these high energy conformations.[33] One possible form of soft-core potential function has the form of Equation 1.32,[34] where a value of 0.5 for δ is suggested in the original work. As the interatomic distance, r_{ij} approaches zero, an unsoftened functional would result in an energy of infinity, while Equation 1.32 causes the energy to go to a constant, finite value $(\lambda\delta)^{-6}\sigma_{ij}^{6}$, where ϵ and σ are the combined Lennard-Jones parameters for a pair of atoms, and δ controls the degree of softness.

$$V_{ij}^{LJ} = 4\epsilon_{ij}(1-\lambda)\left(\frac{\sigma_{ij}^{12}}{(\lambda\delta\sigma_{ij})^{6}} - \frac{\sigma_{ij}^{6}}{(\lambda\delta\sigma_{ij})^{3}}\right) \tag{1.32}$$

While the Lennard-Jones softening allows molecules to interact with a finite value at short distances a consequence of this that charged molecules are able to move closer together than if a normal potential is used. This can be bypassed by two methods; either a two-step decoupling can occur, whereby the Lennard-

Jones interactions and electrostatic interactions are decoupled from the system in separate steps (a two-step decoupling), or an additional electrostatic softening term can be used, Equation 1.33.

$$V_{ij}^{ele} = (1-\lambda)\frac{q_i q_j}{4\pi\epsilon_0\sqrt{(\lambda + r_{ij}^2)}} \tag{1.33}$$

Various forms of the soft-core potentials exist and have been applied to a multitude of energy calculations, from binding energies, solvation energies and solubility of additives in amorphous materials.[35–38] The calculated free energy difference will be independent of the soft-core potential system, within a range of sensible parameters.

The above describes the protocol for relative binding free energies, however this can be applied for absolute binding free energies. For absolute binding free energies, state A will correspond to the ligand, and state B will be no ligand. Absolute binding free calculations involve the decoupling of the entire ligand, across the λ pathway. Depending on the size of the ligand, this change is generally much larger change than a relative perturbation between two similar ligands, thus reducing the phase space overlap. Not only is the removal of the ligand a large change in the system, but the protein itself may also adjust, differing in structure between the apo and holo form. Absolute free energy packages counteract this by introducing many λ windows to the perturbation, increasing the computational cost.[39] An interesting application is the Waterswap implementation of absolute free energy calculations in Sire,[40,41] where instead of fully decoupling the ligand, the ligand is perturbed across the λ coordinate into N water molecules that occupy the equivalent volume of the ligand. This prevents the creation of a vacuum on decoupling, and will solvate the apo form of the protein. Waterswap assumes that the apo form of the protein is solvated at a density of bulk water, and that the conformation of the holo and apo protein do not significantly differ.

Restraints and constraints

The dual topology method used for calculating ligand binding involves simulating one molecule in the off state, at $\lambda = 0$ and 1. When a molecule is entirely non-interacting with the environment, it is able to move by a random walk through the volume of the simulation. This causes the sampled phase space at $\lambda = 0$ and 1 to significantly differ from intermediate λ values as the molecule can move into clashing regions owing to being non-interacting, which can result in a lack of overlap in the phase spaces of neighbouring coupling parameters. The prevention of the non-interacting molecule sampling configurational space that is unavailable to it when in an interacting state can be achieved by trapping the molecule in the locality of the relevant configurational space. The relevant part of configurational space can be defined by where the ligand is considered to be bound to the protein. This is defined by Hill et al. as being a region in which all configurations with a significant contribution to the chemical potential of the bound state are included, without including large regions of unbound states i.e. states that contribute to the chemical potential in the unbound state.[42] Restraints and constraints differ in application between MC and MD, as in MC, only the configurational partition function is affected, while in MD, restraints also have an effect on the kinetic motion, and therefore the kinetic partition function. Here, discussion will focus on restraints and constraints in MC simulations. Several methods of trapping the molecule exist, including associating the movement of the two states present in a dual topology simulation,[43–45] or associating the ligand to a relevant region of the protein.[46,47] In MC simulations where the macroscopic environmental coordinates do not largely shift through the simulation, particularly if regions are treated as rigid, the molecule can be associated to a location defined by cartesian coordinates. In MD, where phase space is generally better explored, defining restraints or constraints can often be more difficult, as the system may shift from the initial cartesian frame of reference, and restraints need to be defined based on dynamic atom locations.

Trapping a molecule in a given location has the effect of changing its chemical potential from that of a standard concentration.[48] This can be corrected for, by calculating the energy associated with trapping the ligand, which depends on the restraint or constraint method used. The two major methods that will be considered here, are restraints and constraints. A constraint is where a hard-wall potential is applied to the molecule, such that when the molecule occupies a region outside its allowed volume (typically spherical) its energy will become infinite and the move will be rejected. This has the effect of trapping the molecule into the defined volume. Restraining a molecule typically involves applying a harmonic potential to its energy, $k(x - x_o)^2$, where the minimum of the harmonic potential is at x_o, the expected location of the molecule, and its current location is x, where k is the force constraint. The free energy calculated using a restraint or constraint needs to be corrected, so that the trapped molecule effectively occupies the same volume in the trapped phase as its standard state. The correction required is shown in Equation 1.34.

$$\Delta G^{\ominus}_{volcorr} = k_B T ln \left(\frac{V^{sim}}{V^{\ominus}} \right) \tag{1.34}$$

$V^{\ominus}$ is the volume occupied by a molecule at standard concentration, which for 1 M solution is a volume per molecule of 1660 Å^3. V^{sim} is the volume available to the molecule in the simulation. For constraint calculations it is simply the volume within the hardwall potential. For restraint calculations the volume available due to the harmonic potential is calculated as $\left(\frac{2\pi k_B T}{k}\right)^{\frac{3}{2}}$.[48] This volume correction is required whenever a restraint or a constraint is used. A harmonic restraint was first used for the calculation of the binding free energy of a xenon atom to myoglobin,[49] however the statistical basis for the correction was first presented by Roux et al.[50] when studying the affinity of water molecules in protein cavities. The volume correction is required to relate free energy of the restrained or constrained simulation back to a well-defined standard state. The resulting free energy should be independent of the strength of harmonic restraint or volume of a constraint, within the limit that the volume is consistent with the definition provided by Hill et al.[42]

With the GCMC method, which will be introduced in Section 1.4, a correction

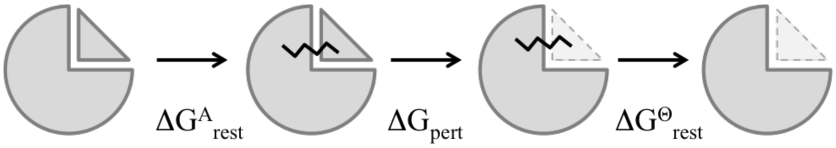


Figure 1.3: Schematic of double decoupling (DD). The ligand protein system is shown, with the restraint illustrated using a black line. When the ligand is 'off' it is shown with a dashed border.

of a similar form to that required for double decoupling (DD) methods will be presented in Section 2.3.5. DD is the process whereby the absolute binding free energy of a species can be determined. First, a restraint is applied to the species, shown as a ligand in a protein in Figure 1.3. The energetic cost of applying the restraint, ΔG^{A}_{rest} is often negligible, but can be calculated using the Zwanzig equation. ΔG_{pert} is the free energy of perturbing the species from the fully interacting, to the fully decoupled state. The final term, $\Delta G^{\ominus}_{rest}$ is standard state term defined in equation 1.34 that can be solved analytically.

Further corrections

Gilson et al. present several other corrections to calculate the standard free energy change of decoupling a ligand from a binding site[48] . The main result is shown in Equation 1.35.

$$
\begin{aligned}
\Delta G_1^{\ominus} = & \underbrace{\left\langle \frac{\delta U(\lambda, r_A, r_B, \zeta_B, r_S)}{\delta \lambda} \right\rangle_{\lambda} d\lambda}_{\text{free energy}} - \underbrace{k_B T ln \left(\frac{\sigma_{AB}}{\sigma_A \sigma_B} \right)}_{\text{symmetry}} \\
& + \underbrace{k_B T ln \left(\frac{V^{sim}}{V^{\ominus}} \right)}_{\text{ligand volume}} + \underbrace{k_B T ln \left(\frac{\xi_1}{8\pi^2} \right)}_{\text{rotation}} + \underbrace{P^{\ominus}(V_A - V_{AB})}_{\text{system volume}} \qquad (1.35)
\end{aligned}
$$

The free energy term in Equation 1.35 is determined from a simulation, with methods outlined in Section 1.2. The symmetry correction arises from the denominator of the molecular partition function of the bound complex and the unbound specices, where σ is the symmetry number of each state. The ligand volume correction is the correction of the chemical potential of the restrained volume to that in the standard state, discussed in Section 1.2 and Equation 1.34. The rotational term is equivalent to the ligand volume, in the case where restraints on the molecule prevents full orientational sampling, e.g. if a symmetrical molecule is prevented from sampling any of its symmetry mates. The system volume correction is the pressure-volume work associated with the overall protein-solvent system when a ligand is decoupled. In most cases, the protein-solvent system is significantly larger than the decoupled ligand, causing the change in system volume on decoupling to be small, and the correction negligible. The difficulty of understanding the various corrections to the free energy, and methods of applying them will be discussed.

The symmetry number is the number of states of a molecule that are interchangeable through the permutation of indistinguishable atoms.[51] The symmetry number of a molecule can be determined by inspection, whereby σ is the number of unique configuration of atoms possible through the symmetry operations of its point group. Note it is the number of unique permutations that contribute to the symmetry number, rather than the range of operations. The symmetry correction is required as computational modelling assigns distinguishable labels to simulations i.e. (H1, H2) to atoms, which through molecular symmetry are equivalent. If sampling allows H1 and H2 to interchange, the phase space of the molecule is twice as large than if they do not interchange. More generally, a molecule with a symmetry of σ will have an available phase space proportional to σ, depending on the sampling of the system. A point of uncertainty is the contribution of internal symmetry number of a molecule. A methyl group of a molecule is considered to contribute a symmetry number of 3, if the group is free to rotate. Should the group be unable to rotate such as at low energy, then the symmetry number of the group is 1. This introduces both internal symmetry, and temperature dependence if the likelihood of rotation has a thermally accessible barrier.[52]

When calculating the binding free energy of ligand A with protein B, the assumption would be that both the protein and the complex would have no symmetry (σ_B, $\sigma_{AB} = 1$). However considering internal symmetry, a protein may have a higher symmetry number, provided by each methyl, carboxylic acid or other rotational R group. The assumption can be made that the protein has the same symmetry in the bound and unbound state, then the two terms cancel leaving only the term for the ligand symmetry. A large symmetrical molecule such as benzene may lose its ability to rotate and fully sample its symmetry states, however small molecules such as water should be free to rotate when in complex. In this case where the ligand is mobile and can sample as many orientations in the bound leg as the free leg, the symmetry term will cancel, as the symmetry number of the complex A-B should be a product of the symmetry number of both A and B.

Mobley et al. have worked on the use of orientational restraints to prevent ligand flipping, and determined the appropriate symmetry correction.[53] If a ligand has been restrained to only one of its possible orientations, then a symmetry correction is required. If a ligand is unrestrained, however does not fully sample all of its possible rotations, then the correction would also be required. Mobley et al. state if " [a ligand's] orientations were sampled a number of times, no correction factor would be necessary". This is difficult to implement as it is unclear how much of the symmetrical phase space needs to be sampled, and how frequent the transitions between the two orientations would be required for a correction factor to be applied. Ross et al.[54] applied the symmetry correction when only one orientation was observed during the fully 'on' state of the ligand during decoupling, however Mobley et al. suggest that the correction factor would need to be applied to each individual λ replica, depending on the orientations sampled at each window.

The application of symmetry corrections and rotational corrections can be difficult to navigate. Corrections should be used if the sampling of the ligand in the bound state differs to that of the free state, whether the difference arises due to applied restraints or constraints, or as an artefact of the ligand being 'trapped' in the active site.[53,54] The understanding that corrections are required when the sampling between two legs is inconsistent supports the volume correction presented in Chapter 2.

replica exchange

Replica exchange (RE) is a computational tool developed to both improve sampling and reduce the correlation times of simulations.[55] The premise is that multiple repeats of the same system are set up, each varying in a given parameter, where the parameter may be, but limited to; temperature,[55] λ[56] and the Hamiltonian.[57] Along the simulation trajectory, swaps are attempted between the different repeats, and accepted or rejected following the Metropolis criterion,

$$P_{swap} = min\left[1, e^{(\beta_i - \beta_j)(U_i - U_j)}\right] \tag{1.36}$$

where a swap between replicas i and j is proportional to the potential energy U_x of each state, and thermodynamic beta $\beta_x = \frac{1}{k_B T_x}$ at the temperature T_x.

Following the example of temperature, this allows configurations that are accessible at higher temperature to exchange with those at lower temperature. This can allow for transitions that would not typically be observed in the lower temperature repeats to be observed and overcome barriers in the simulation. These swaps do not affect the Boltzmann distribution of any of the ensembles, and RE methodologies have been expanded to molecular dynamics (REMD).[26]

The introduction of replicas comes at additional computational expense. For the additional cost to be of value, RE needs to be both efficient (a fair number of accepted swaps) and useful (replicas are enhancing the sampling). These two conditions are somewhat contradictory; the more different and interesting the two states are, the less likely exchange is to occur. Multiple closely spaced replicas are used, and various attempts have been made to most efficiently distribute different replicas.[58] If the replicas are too closely spaced however, while exchange will be frequent, the benefits in terms of phase space accessed will be small. Mixing between states has been shown to be most efficient when exchange attempt rates are high.[59] High exchange attempt rates are possible if the computational cost of trialling an exchange is cheap. For some protocols such as temperature RE the acceptance probability for which is shown in Equation 1.36, the attempt is cheap, as the temperature and the total energy of a configuration are known, but for replica exchange between states of differing λ or Hamiltonian, the acceptance test requires additional energetic evaluations.

Typically exchanges are attempted between neighbouring replicas, so as to increase the likelihood of accepted swaps. While this improves exchange rates, this can result in slow diffusion of a replica across the replica space. Several exchange schemes have been suggested to speed up the random walk of the replica. One such method is to attempt an all-pairs exchange,[60] which was found to result in a four-fold speed up of replica diffusion for an 8 replica system of aniline dipeptide while maintaining detailed balance. Instead of only calculating P_{ij} between states i and j where $j = i+1$, all-pairs exchange calculates P_{ij} for all other replicas. The j state that is then swapped with state i is randomly chosen from the normalised probabilities of all swaps. All-pairs RE allows for quicker sampling as larger steps of states are possible. An alternate protocol can be to use self-adjusted mixture sampling (SAMS) whereby the parameter of a single walker is able to adjust along the simulation, within a parameter's locality.[61]

1.3 Methods of calculating water binding

Two details of active site water molecules are of interest; their location and their binding free energy. The location of an active site water molecule can be observed experimentally from a crystal structure, however for a given protein-ligand complex, the experimental structure may not exist, the protein may be too difficult to crystallise, or it may have been studied in the apo form, or bound to a different ligand. Depending on the similarity between the different ligands, it can be difficult to assume the ligands will bind in the same manner to each other, or if the water network for one ligand is conserved with the other ligand. Understanding where water molecules are in a crystal structure can be difficult, and will be discussed in detail in Section 1.5. The electron densities gained from crystallographic studies are a superposition of all possible positions of the electron density during the course of the experiment, which means that only well ordered water molecules will be observed. If the thermal fluctuation of a water molecule is greater than 1 Å, the limit at which electron densities can be resolved, it will not be seen in the crystallographic results.[62] NMR studies of protein systems are based on the intermolecular nuclear Overhauser effect (NOE), whereby the distance between water molecules and protein atoms are monitored, rather than the electron density.[63] This means that NMR can be used to observe more transient water molecules in protein complexes that would be blurred in the corresponding electron density.[64] NMR also has limitations however, as the NOE intensity decays with proton-proton distance at a rate of r^{-6}, where r is the inter-proton distance, which requires the active site water molecules to be directly interacting. For all of the stated reasons, it can be experimentally difficult to conclude where active site water molecules are located.

The other factor of interest is the binding affinity of active site water molecules. Rationalising if a water molecule should be retained or displaced in drug design is difficult and requires knowing how tightly bound the water molecules are. A weakly bound water molecule will be easy to displace, and doing so will release entropy. A tightly bound water molecule will come at a larger energetic cost to displace, although it may be occupying a region of protein that a ligand could interact with more favourably. These factors also need to be balanced with the cost of disrupting the hydrogen bonding network within the active site, as displacing a water molecule may destabilise other adjacent molecules. While it is possible to locate water molecules using crystallographic or NMR results, it is more difficult to calculate a binding affinity of active site water. As this is impossible to directly evaluate experimentally, it marks a region where computational techniques can be helpful. Various methods exist to both locate active site water molecules and calculate their binding affinity. This section will cover a few of these methods -DD methods, Watermap, and Just Add Waters (JAWS).

The binding free energy of a water molecule to an active site could be calculated from its relative binding and unbinding rates from determining residence times from simulations. It is not currently possible to observe multiple binding events as standard in a typical simulation. The residence time of a water molecule in an active site water network has been suggested to be on the order of microseconds,[65] which is significantly longer than the timescale of a typical simulation. This means that the binding and unbinding of active site water molecules is not typically observed in the timescale of a simulation. These sampling limitations are worsened in cases where the active site is occluded from the bulk, or if the water molecule is 'pinned' by a binding partner that would need to unbind to allow for a pathway for the waters of interest to vacate. In addition, ProtoMS and other MC software packages often reduce the sampling of parts of the system such as the protein backbone, where it could reasonably be expected that large scale motions of the protein are required for water or ligand binding or unbinding to be observed. All of these factors mean that it is not currently possible to determine the binding free energy of a water molecule in a typical simulation through monitoring simulation residence times. This means that enhanced sampling methods are required.

The binding free-energy of a water can be calculated using the methods outlined in Section 1.2, where an individual water molecule can be decoupled from its environment in the way that ligands are treated. While absolute binding free energy calculations are generally avoided due to the large changes in energies involved, absolute decoupling of a water molecule can be well-behaved as the molecule is small, meaning the phase-space overlap is better than for a ligand, where the ligands disappearance would cause a large change in the surrounding system. In efforts to classify waters in protein systems and determine their use for drug design, Barillari et al. calculated the binding free energies of 54 water molecules in systems of interest using the DD method with TI.[66] For the calculations, hard-wall constraints were used to prevent the water leaving its location during the calculation, and to exclude the volume to other water molecules. The interactions between the water and its system were decoupled in two stages; firstly decoupling electrostatic charges, followed by van der Waals interactions. This work demonstrates a method of calculating a binding energy of a water molecule that will be used herein, while also seen in other research.[5] The work by Barillari et al. was able to demonstrate with a 95% level of confidence that water molecules that are tightly bound are more likely to be conserved between structures. However, no statistical correlation was found between the affinity of a ligand, and the binding energy of water molecules it displaced on binding. This highlights the need for a method that can incorporate water binding within the active site, and the end goal — the ligand affinity.

Another example whereby water binding free energies have been determined using the DD method is a paper by Michel et al.[5] Michel el al.'s paper is the basis for some of the research performed herein. For three systems, where a water molecule is known to affect the binding of a ligand, the relative binding free energy of the ligands has been calculated, both with and without the water molecule present. The two thermodynamic cycles between the ligands are mapped onto each other using the free energy of binding of the water in the presence of each of the ligands. This provides a thermodynamic map, whereby a ligand without a water molecule can be compared to a different ligand with a water molecule through a free energy pathway of various steps. Multiple pathways can exist between the two states, which can result in different free energy differences due to errors in cycle closures. This provides a method for comparing ligands for which water occupancies are different, but involves simulation of high-energy states, which can introduce errors into the calculations. This method, as with other methods that rely on the decoupling of individual water molecules, quickly become laborious as the number of waters in a system increases, particularly if appropriate care is taken for the order in which water molecules are decoupled. Double decoupling requires *a priori* knowledge of a hydration site, as restraints or constraints are required. Michel et al. determined the location of water molecules using JAWS.

While it is possible to perform DD on a collection of water molecules simultaneously, typically simulations are done only on one water at a time to increase the accuracy of the results. DD is restrictive as the water location is required *a priori*. Methods such as WaterMap and JAWS have been developed with these issues in mind. At the time of writing, WaterMap is moving to a GCMC-type method (release 2018-2), but the previous method (2018-1 and prior) will be described herein. WaterMap is an MD method, where the protein and ligand are simulated in bulk solvent, and water positions are calculated based on the locations of waters throughout the simulation, typically 2 ns in length.[67,68] The locations of the water molecules throughout the MD simulation are clustered before inhomogeneous solvation theory (IST) is used to determine thermodynamic properties of each water.[69] Watermap has been used to locate water molecules in protein-ligand systems, typically with a water network of a holo-protein structure determined, followed by analysis of which molecules would be displaced if a ligand is overlaid with the structure. This method of looking only at displaced molecules has a tendency to overlook subtle changes and shifts in the water network which may influence ligand binding.[70] One limitation of WaterMap, as is the case with DD methods, is that each hydration site is considered as its own entity rather than as part of a network. It can be misleading to consider a water's binding free energy in isolation from the rest of the system, as the effects of perturbing the network through secondary interactions can be missed. For this reason, grid inhomogeneous solvation theory (GIST) has been developed, which considers the thermodynamic properties of the grid, rather than each hydration site,[71] which is advantageous as it is rapid to calculate. Both WaterMap and GIST have been applied to systems of pharmaceutical interest.[72–74]

Just Add Waters (JAWS) is a λ-dynamics MC based method,[75] whereby water molecules sample, and are mapped onto grid points. Water molecules are able to scale in λ, i.e. the degree to which they interact with their system through, sampled using a MC test. The locations in which the water molecules spend much time in the 'on' interacting state are understood to be favourable binding sites. The water sampling locations are clustered to identify possible binding sites. A second simulation is required for each of the possible binding sites found in the initial calculation, whereby the binding free energy of that water is estimated from the ratio of simulation in which the water molecule is 'on' or 'off'.[75] The 'on' and 'off' states are defined as $\lambda > 0.95$ and $\lambda < 0.05$ respectively. The transfer free energy is calculated using this ratio of on and off probabilities using Equation 1.37.

$$\Delta G_{trans} \simeq -k_B T ln \left(\frac{P(\lambda > 0.95)}{P(\lambda < 0.05)} \right) \tag{1.37}$$

The choice of 0.95 and 0.05 for the definition of on and off is arbitrary, and results in energies that are estimates, rather than rigorous. While still contributing to the system sampling, the time a water molecule is in the λ region between 0.05 and 0.95 it is not contributing to the free energy calculation.

The methods described are examples of a class of simulation-based predictions of water molecules. Many other methods exist that also use IST to calculate the free energy of water molecules.[69,76,77] Other simulation based methods include 3D-grids to probe an area, including by Setny et al. and 3D-RISM. Methods exist that do not rely on simulation, but instead predict the locations of water molecules based on the locations of water molecules in other crystal structures. These methods are referred to as knowledge-based,[78–82] and are advantageous as no force field or lengthy simulation time is required, but the methods will only be as good as the data on which they are trained. One recently published knowledge-based method is WarPP,[82] and will be looked at in detail in Chapter 3, as their success rate of 80% of water molecules correctly predicted within 1.0 Å cutoff for a large dataset of 20,000 waters is — to our knowledge — the highest published success rate.

To conclude, experimental techniques exist that are able to locate water molecules in protein systems while none exist that are able to directly determine their binding free energy. This indicates that computational techniques may be able to provide information that is useful for drug design. Double decoupling methods are limited, as the water binding site needs to be known *a priori*. Additionally the simulations often need restraints or constraints which can be non-trivial to perform. As DD can only determine one water molecule at a time with ease, water network methods have been developed, including WaterMap (MD) and JAWS (MC). Both methods involve a two stage simulation — one to identify hydration sites and one to calculate the binding free energy. WaterMap calculates binding free energies using IST, which may be limited when water network energies are of primary interest.

GIST has been developed to bridge this gap, but the underlying method is still only as good as sampling allows. While JAWS enhances sampling using λ scaling, the free energies calculated are estimates. GCMC is a method that can determine the hydration sites of water networks, as well as rigorously calculating their binding free energy in a single simulation. Where other methods available are a compromise between the number of water molecules that can be simulated, and the quality of resulting binding free energy, the GCMC method is able to calculate both of these rigorously, within a single simulation. The theoretical basis of the grand canonical ensemble, the MC insertion criteria, the computational methodology and the rigorous calculation of Gibbs free energies will be outlined in Section 1.4, and developed further throughout this thesis.

1.4 Grand canonical Monte Carlo

As discussed, understanding where water molecules are and their binding free energies is not always easy to do experimentally. Various computational methods have been developed to perform this task, and a selection of these have been presented in the previous section. While each method has its advantages, none of the presented methods can calculate both the locations of multiple water molecules and their binding free energies rigorously, within a single simulation. While the free energies determined from DD are theoretically exact, the method does not scale well for water networks. Methods that do scale to multiple water molecules, such as JAWs or WaterMap, do not give energies as accurate as DD. GCMC is able to handle both water placement and binding free energy calculation of many waters in a single simulation.

The grand canonical (GC) ensemble is the statistical ensemble of states of a given chemical potential, temperature and volume, μVT, where μ is chemical potential, V is volume and T is temperature. States in the GC ensemble, can vary in both total energy and the number of particles. This can be thought of as being open to exchanging both energy and particles with a reservoir, where the reservoir

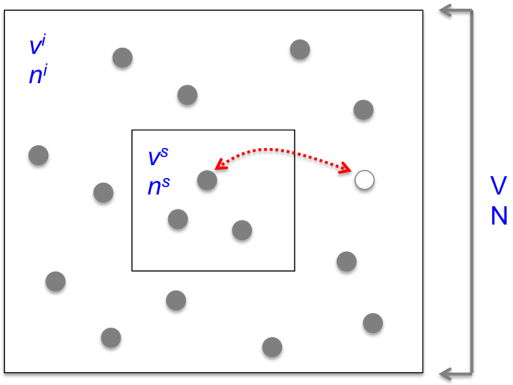


Figure 1.4: Illustration of the grand canonical ensemble, consisting of two, canonical ensembles, with an interface permeable to molecules. The inner ensemble will be the system of interest, while the outer is a non-interacting ideal gas, of infinite size. Adapted from *Understanding Molecular Simulation.*[13]

is an ideal gas. This allows properties dependent on an average number of particles to be calculated as a function of their external conditions. One example of a use of the GC ensemble is determining the extent of gas adsorption on a surface, at constant temperature and pressure.[13] In principle, a system like this, where the the number of adsorbed particles varies, could be simulated in an NPT ensemble (where P is pressure). However, equilibration between the surface and gas phase may be far longer than feasibly computable due to slow diffusion. Large simulations in the NPT ensemble would be required to gain the correct average molecular occupancy, where the GC ensemble is able to bypass the slow diffusion processes that are limiting in other ensembles.

theoretical basis

The grand canonical partition function can be calculated from the canonical partition function. In the canonical ensemble, the system and a bath are in thermal equilibrium, whereby energy is able to pass between the two systems. This is known as the NVT ensemble. The grand canonical ensemble (μVT) can be understood by considering a canonical (NVT) ensemble that has been divided into two, where the divide is permeable to atoms. The two halves of the canonical system, which will now be considered as one ideal part, and one 'system' part, subscripted i and s respectively, have their own chemical potential (μ), volume (v_x), temperature (T), and number of atoms (n_x). The partition functions of each of the subsystems, are shown in Equations 1.38 and 1.39.[13]

$$Q_i(n_i, v_i, T) = \frac{{v_i}^{n_i}}{\Lambda^{3n_i} n_i!} \int ds^{n_i} e^{-\beta U(s^{n_i})} \tag{1.38}$$

$$Q_s(n_s, v_s, T) = \frac{{v_s}^{n_s}}{\Lambda^{3n_s} n_s!} \int ds^{n_s} e^{-\beta U(s^{n_s})} \tag{1.39}$$

where Λ is the thermal de Broglie wavelength, U is the potential energy of the system defined using scaled coordinates, s^{n_x}, where $s^{n_x} = V^{-\frac{1}{3}} r^{n_x}$ with r^{n_x} are the unscaled system coordinates. The partition function of the overall, NVT ensemble is the product of the two subsystems;

$$Q(n_i, n_s, v_i, v_s, T) = \frac{v_i^{n_i} v_s^{n_s}}{\Lambda^{3n_i} n_i! \Lambda^{3n_s} n_s!} \int ds^{n_i} \int ds^{n_s} e^{-\beta U(s^{n_s})} \cancelto{1}{e^{-\beta U(s^{n_i})}} \tag{1.40}$$

where the integral over the non-interacting ideal gas will be one. The division between the two subsystems is permeable, allowing n_i and n_s to interchange, however the total number of particles, N is constant, $n_i + n_s = N$. The partition function of the ensemble needs to consider every possible division of atoms between the two systems.

$$Q(N, v_i, v_s, T) = \sum_{n_s=0}^{N} \frac{v_i^{n_i} v_s^{n_s}}{\Lambda^{3n_i} n_i! \Lambda^{3n_s} n_s!} \int ds^{n_s} e^{-\beta U(s^{n_s})} \tag{1.41}$$

The chemical potential of an ideal gas can be determined from the ideal partition function,

$$F = -k_B T ln(Q_p) \tag{1.42a}$$

$$F = -k_B T ln\left(\frac{V^N}{N!\Lambda^{3N}}\right) \tag{1.42b}$$

$$F = -k_B T \left[N ln\left(\frac{V}{\Lambda^3}\right) - N ln(N) + N\right] \tag{1.42c}$$

Equation 1.42c can be reached using Stirling's approximation. The free energy can be related to chemical potential using $F = \mu N$;

$$\frac{\delta F}{\delta N} = \mu = -k_B T ln\left(\frac{V}{N\Lambda^3}\right) \tag{1.43}$$

which when particle density, $\rho = \frac{N}{V}$ is used;

$$\mu = k_B T ln(\Lambda^3 \rho) \tag{1.44}$$

. Considering the limit where the ideal gas reservoir is infinitely larger than the system, $n_i \rightarrow \infty$, using Stirling's approximation, Equation 1.41, becomes;

$$Q(N, v_i, v_s, T) = \sum_{n_s=0}^{\infty} \frac{v_i^{n_i} v_s^{n_s}}{\Lambda^{3n_i} n_i^{n_i} \Lambda^{3n_s} n_s!} \int ds^{n_s} e^{-\beta U(s^{n_s})} \tag{1.45}$$

which it is possible to rearrange to;

$$Q(N, v_i, v_s, T) = \sum_{n_s=0}^{\infty} \left(\frac{v_i}{\Lambda^3 n_i}\right)^{n_i} \frac{v_s^{n_s}}{\Lambda^{3n_s} n_s!} \int ds^{n_s} e^{-\beta U(s^{n_s})} \tag{1.46}$$

Substituting in Equation 1.44

$$Q(N, v_i, v_s, T) = \sum_{n_s=0}^{\infty} e^{-\beta\mu n_i} \frac{v_s^{n_s}}{\Lambda^{3n_s} n_s!} \int ds^{n_s} e^{-\beta U(s^{n_s})} \tag{1.47}$$

. To remove dependency on n_i, $n_i = N - n_s$ can be used, leading to;

$$Q(N, v_i, v_s, T) = \sum_{n_s=0}^{\infty} \cancel{e^{-\beta\mu N}} e^{\beta\mu n_s} \frac{v_s^{n_s}}{\Lambda^{3n_s} n_s!} \int ds^{n_s} e^{-\beta U(s^{n_s})} \tag{1.48}$$

where e^{-N} will cancel to 1 in the limit where N is large. This results in the grand canonical partition function, Equation 1.49, where there is no longer any dependence on the ideal gas reservoir,

$$Q(\mu, v_s, T) = \sum_{n_s=0}^{\infty} e^{\beta\mu n_s} \frac{v_s^{n_s}}{\Lambda^{3n_s} n_s!} \int ds^{n_s} e^{-\beta U(s^{n_s})} \tag{1.49a}$$

$$= \sum_{n_s=0}^{\infty} e^{\beta\mu n_s} Q(n_s, v_s, T) \tag{1.49b}$$

which is related to the canonical partition function. Here, v_s and n_s are used to refer to the volume and number of particles of the system, but as there is no longer any dependence on the ideal gas reservoir N and V are often used to refer to the system, as opposed to how they are illustrated in Figure 1.4 and used in this derivation. Using N and V to refer to the interacting system will be used onwards.

The detailed balance condition can be used to derive the acceptance criteria for GC insertion and deletion moves. Detailed balance is discussed for configurational sampling moves in Section 1.1, and the same method is used here. The probability density of a grand canonical state is shown in Equation 1.50.

$$\mathcal{N} \propto \frac{e^{\beta\mu N} V^N}{\Lambda^{3N} N!} e^{-\beta U(s^N)} \tag{1.50}$$

While the probability of a particular insertion and deletion move are equal ($\alpha(N \rightarrow N+1) = \alpha(N+1 \rightarrow N)$), the acceptance criteria depend only on the density of the states.

$$\frac{acc(N \rightarrow N+1)}{acc(N+1 \rightarrow N)} = \frac{\mathcal{N}(N+1)\alpha(N+1 \rightarrow N)}{\mathcal{N}(N)\alpha(N \rightarrow N+1)} \tag{1.51}$$

Substituting Equations 1.50 and 1.51;

$$\frac{acc(N \to N+1)}{acc(N+1 \to N)} = \frac{\Lambda^{3N} N!}{V^N e^{\beta\mu N} e^{-\beta U(s^N)}} \frac{V^{(N+1)} e^{\beta\mu(N+1)} e^{-\beta U(s^{N+1})}}{\Lambda^{3(N+1)}(N+1)!} \tag{1.52a}$$

$$= \frac{V}{\Lambda^3(N+1)} e^{\beta\mu} e^{-\beta(U(s^{N+1})-U(s^N))} \tag{1.52b}$$

This leads to the insertion and deletion Metropolis criteria;

$$acc(N \to N+1) = min\left[1, \frac{V}{\Lambda^3(N+1)} e^{\beta\mu} e^{-\beta[U(s^{N+1})-U(s^N)]}\right] \tag{1.53a}$$

$$acc(N \to N-1) = min\left[1, \frac{\Lambda^3 N}{V} e^{-\beta\mu} e^{-\beta[U(s^{N-1})-U(s^N)]}\right] \tag{1.53b}$$

which hold as long as insertion and deletion moves are attempted with equal likelihood.

methods

Above, the GCMC insertion and deletion acceptance criteria are shown, required for simulating in the μVT ensemble. As N is a variable in GCMC simulations, the value of N is controlled by μ, the chemical potential. The excess chemical potential of a system is the difference between the given chemical potential, and the equivalent ideal gas system. This is the Helmholtz free energy required to move a particle between a system and the ideal gas. This allows the excess chemical potential to be related to the Helmholtz free energy difference between a system, and the system with one fewer molecule, Equation 1.54, where $\Delta N = 1$

$$\mu' = \frac{\Delta F_{ex}}{\Delta N} = F_{ex}(N+1) - F_{ex}(N) \tag{1.54}$$

The excess chemical potential of a system can be determined computationally using Widom's particle insertion method.[83] Widom particle insertion involves repeated attempts to insert a test particle into a system, with the excess chemical potential calculated from the exponential of the energy of the insertion, Equation 1.55.

$$\mu' = -\frac{1}{\beta} ln \int ds_{N+1} \left\langle e^{-\beta \Delta U} \right\rangle_N \tag{1.55}$$

In practise, this will be a summation over microstates observed, rather than an integral over all states of the system. The GC acceptance criteria, Equation 1.53, were derived by Adams, who first included the acceptance and deletion moves to simulate with the μVT ensemble, for both a hard-sphere fluid[84] and a Lennard-Jones fluid.[85] The Adams formulation of GCMC is the method used in this work. The method allows the chemical potential to be chosen for the simulation, however choosing a sensible value of μ is less intuitive than other parameters such as V and T, as it is not an experimental observable. Adams combined the chemical potential with other required constants, to use the B parameter (or Adams parameter), Equation 1.56,

$$B = \mu\beta + ln\left(\frac{V_{sys}}{\Lambda^3}\right) \tag{1.56}$$

Where V_{sys} is the volume of the GCMC region. This can be substituted into Equation 1.53, to provide the equivalent Metropolis criteria shown;

$$acc(N \rightarrow N+1) = min\left[1, \frac{1}{(N+1)} e^{B} e^{-\beta[U(N+1)-U(N)]}\right] \tag{1.57a}$$

$$acc(N+1 \rightarrow N) = min\left[1, N e^{-B} e^{-\beta[U(N-1)-U(N)]}\right] \tag{1.57b}$$

As the chemical potential controls the average number of particles in a system, the correct chemical potential can be established deterministically — where simulations are repeated with different μ until an expected value of N is observed.[86] At lower chemical potentials, fewer particles are inserted into the system, and therefore the lower the chemical potential at which a particle is first inserted, the more favourable the inserted particle is interacting with the system. A consequence of this is that the particles can be rank-ordered by affinity according to the chemical potential at which they insert. This however requires knowing the expected value of N for the system, which is a parameter that would ideally be calculated without prior knowledge.

The GCMC method has previously applied to ligand-protein systems,[86, 87] with water molecules treated as the GC species present in the simulation. The chemical potential was determined by matching the known experimental value of number of waters for a system, and the waters are rank ordered by the chemical potential at which they first insert into the system. This implementation results in the method being deterministic. Grand canonical integration (GCI) and other theoretical developments will be described that allow the method to be used predictively.

cavity bias

Whether GCMC is considered for protein-ligand systems, or other applications such as interfaces or porous materials,[88,89] the method can suffer from poor acceptance rates for insertion and deletion. GCMC is advantageous for systems where the property of interest is dependent on a slow rate of diffusion, whereby the GCMC methodology is able to computationally speed up the sampling of the effect of the diffusion. One factor that may cause diffusion to be slow, is the density of the system of interest. If a system is high density, attempting an insertion into the GCMC region can be difficult due to a high probability of overlapping, high energy configurations. Attempts have been made to improve the acceptance rates of GCMC, most notably through cavity bias. Cavity bias was presented by Mezei to study a dense LJ fluid at the triple point.[90,91]

Cavity bias GCMC (CB-GCMC) has an additional stage in the algorithm, where prior to an insertion, a grid search of N_t points is attempted over the GCMC region, to estimate the probability (P_c^N) of finding a cavity of radius larger than R_c. Several methods of calculating P_c^N are suggested by Mezei and are labelled following the original notation; P_c^N can be an average of all $P_c^N(r^N)$ observed previously the simulation (mean, M), or an average of $P_c^N(r^N)$ where the insertion or deletion move is accepted (accepted mean, AM) or $P_c^N(r^N)$ is determined by a frame-wise grid search at each step, (grid search, GX). The insertion is then attempted into one of the grid points identified, and accepted or rejected based on Equation 1.58a. This prevents attempting insertions that result in overlap and high energy structures, by only attempting insertions where there is space. This means that ΔU of insertion is finite and small, and will increase the likelihood of acceptance. As the insertion acceptance criteria has changed, so must the deletion acceptance to maintain detailed balance, Equation 1.58b.

$$acc(N \rightarrow N+1)^{CB} = min\left[1, \frac{V}{\Lambda^3(N+1)} P_c^N e^{\beta\mu} e^{-\beta[U(N+1)-U(N)]}\right] \tag{1.58a}$$

$$acc(N \rightarrow N-1)^{CB} = min\left[1, \frac{\Lambda^3 N}{V P_c^{N-1}} e^{\beta\mu} e^{-\beta[U(N+1)-U(N)]}\right] \tag{1.58b}$$

If the system has very high density ($P_c^N = 0$), then there is no point within the grid system for cavity bias to attempt an insertion, and 1.58a approaches zero. In this case, the simulation will revert to unbiased GCMC insertion criteria, where the insertion will be attempted at any location, with the acceptance criteria following Equation 1.53a. As it is possible for the insertion to revert to the unbiased scheme, the deletion moves must be balanced with respect to this, whereby the unbiased deletion move will be attempted with a probability of $(1 - P_c^{N-1})$.

The work by Mezei considers cavity-bias GCMC simulations of densely packed LJ fluids. Roux et al. have extended this methodology to make orientational-bias cavity-bias GCMC for systems where the GC solute has orientational degrees of freedom.[92] This follows the basic methodology outlined by Mezei. However, once a grid-based search has been performed to find a suitable cavity site, an orientation l of the species is chosen with the probability;

$$P_{CO}^N = \frac{e^{-\beta U_l}}{\sum_{i=1}^m e^{-\beta U_i}} \tag{1.59}$$

where U_i is the potential energy of the i^{th} orientational state of the m orientational trial states, of which the l^{th} state is chosen for insertion. Where co indicates that the simulation has both a cavity and an orientational bias. This P_{CO}^N is then used in the acceptance insertion criteria shown in Equation 1.58. When a deletion molecule is attempted, $m - 1$ alternate conformations for the water molecule of interest are generated, so as to calculate P_{CO}^{N} for the deletion move. This ensures detailed balance is maintained.

While cavity bias has been shown to be an effective method to increase acceptance rates of GCMC moves, the extent of its benefit will depend on the cost of the additional effort of evaluating P_c^N throughout the simulation. Efficient methods for calculating P_c^N in the simulation have been suggested[91] using a FCC packing grid that updates with a frequency dependent on the types of accepted moves within the simulations – a simple translation is unlikely to largely affect the number or size of cavities, whereas a successful insertion or deletion will. The additional orientation bias was also found to increase the GCMC insertion rates, from 0.06% to 0.81%, but it also requires many additional calculation steps for each MC of the protocol.

grand canonical integration

The binding affinity of a single water molecule can be calculated using the interacting-particle method, presented by Clark et al.[93] This has been applied to protein-ligand systems, and a general form of the equation is shown in Equation 1.60. This requires simulations performed over a range of chemical potentials, to give a range of corresponding N values.

$$N(B) = \frac{1}{1 + e^{\beta \Delta F_{trans} - B}} \tag{1.60}$$

The equation is of the form of a logistic function. This general form of the equation was presented by Ross et al., Equation 1.61,[54] where the integral is performed over the sum of multiple logistic functions. Equation 1.60 shows that ΔF_{trans} will be equal to the half-maximum point of the curve, where N(B) = 0.5. This is where the chemical potential of the system is equal to that of the ideal gas system, and the water is equally likely to be in either system, resulting in an average occupancy of 0.5. The half-maximum is determined by fitting a logistic function to the simulation results, and from that calculating the point of half-maximum.

One benefit of GCMC is the ability to study multiple particles at the same time, and for the case of water molecules in protein-ligand systems, the ability to calculate the energies of multiple water molecules would provide a mechanism with which the optimal water occupancy could be calculated stochastically. The ability to calculate the optimal occupancy is an improvement on the purely deterministic grand canonical methods discussed above. Ross et al. introduced the GC Integration (GCI) Equation, whereby the form of the single-water, Equation 1.60, can be generalised to the case of many-water systems. Previously, fitting a logistic function to the titration points of a single-water system allowed the free energy of that water to be determined. The generalised form can calculate the energy of changing the water occupancy from N_i to N_f molecules, Equation 1.61

$$\beta\Delta F_{trans}(N_i \rightarrow N_f) = N_f B_f - N_i B_i + ln\left(\frac{N_i!}{N_f!}\right) - \int_{B_i}^{B_f} N(B)dB \tag{1.61}$$

ΔF_{trans} is the energy of moving $(f-i)$ waters from an ideal gas into the system of interest. B_i and B_f are the Adams parameters which produce an average water occupancy of N_i and N_f waters respectively. The integral term is calculated by fitting multiple logistic equations to the titration of multiple waters. The $ln\left(\frac{N_i!}{N_f!}\right)$ term is a multiplicity term, which accounts for the ability for molecules to exchange within the active site. In DD simualtions, water molecules with specific atomic labels are either restrained or constrained to a particular hydration site, whereas GCMC allows for the exchange between different atomic labels to occupy different sites. As all molecules can move into the ideal gas 'off' state and sample the whole system, any water is able to insert into any water position within the site. The logarithmic factorial term is able to account for the multiplicity of both the initial and final water network considered. Equation 1.60 is equivalent to the GCI equation, for the case between zero and one water, where $N_i = 0$,$N_f = 1$;

$$\beta\Delta F_{gci}(0 \rightarrow 1) = \cancelto{1}{N_f} B_f - \cancelto{0}{N_i B_i} + \cancelto{0}{ln\left(\frac{N_i!}{N_f!}\right)} - \int_{B_i}^{B_f} N(B)dB \tag{1.62a}$$

$$\beta\Delta F_{gci} = B_f - \int_{B_i}^{B_f} N(B)dB \tag{1.62b}$$

F_{trans} has been replaced with F_{gci} and F_{single} to discern between the free energy determined from Equations 1.60 and 1.61 respectively. Integrating Equation 1.60 gives;

$$\int_{B_i}^{B_f} N(B)\mathrm{d}B = \int_{B_i}^{B_f} \frac{1}{1 + e^{\beta\Delta F_{single} - B}}\mathrm{d}B \tag{1.63a}$$

$$= \left[ln\left(e^{\beta\Delta F_{single}} + e^{B}\right)\right]_{B_i}^{B_f} \tag{1.63b}$$

$$= ln\left(\frac{e^{\beta\Delta F_{single}} + e^{B_f}}{e^{\beta\Delta F_{single}} + e^{B_i}}\right) \tag{1.63c}$$

$$= ln\left(\frac{1 + e^{B_f - \beta\Delta F_{single}}}{1 + e^{B_i - \beta\Delta F_{single}}}\right) \tag{1.63d}$$

For the single water case, both $N(B_i) = 0$ and $N(B_f) = 1$ hold. For the first condition, the denominator in Equation 1.60 must go to infinity, therefore the limit $\beta\Delta F - B_i \rightarrow \text{inf}$ holds (alternately $B_i - \beta\Delta F \rightarrow -\text{inf}$). For $N(B_f) = 1$, the denominator must go to 1, and therefore the limit $B_f - \beta\Delta F \rightarrow \text{inf}$. Substituting both of these into Equation 1.64a;

$$\int_{B_i}^{B_f} N(B)\mathrm{d}B = ln\left(\frac{\cancel{1} + e^{B_f - \beta\Delta F_{single}}}{1 + \cancelto{0}{e^{B_i - \beta\Delta F_{single}}}}\right) \tag{1.64a}$$

$$\int_{B_i}^{B_f} N(B)\mathrm{d}B = B_f - \beta\Delta F_{single} \tag{1.64b}$$

where ΔF_{gci} and ΔF_{single} are now equivalent by inspection of Equations 1.62b and 1.64b.

The form of the GCI Equation shown in Equation 1.61 does not determine standard state binding free energies, but this will be discussed and the correct form presented in Chapter 2.

The GCMC method determines ΔF_{trans}, the Helmholtz free energy to transfer water molecules from an ideal gas system into the system of interest. What we would like to be able to calculate is the Gibbs free energy of binding, ΔG_{bind}, of transferring water molecules from bulk water to the system. For a network of water molecules in a system, the equilibrium occupancy will be where the thermodynamic equilibrium is where this is at a minimum;

$$\frac{d\Delta G_{bind}(N)}{dN} = 0 \tag{1.65}$$

Where ΔG_{bind} is the binding free energy of the water molecules to the system — the metric of interest in GCMC. The Gibbs free energy of binding is the free energy of increasing the number of waters in the system, combined with the free energy of removing those water molecules from solution;

$$\Delta G_{bind} = \Delta G_{sys} - \Delta G_{sol} \tag{1.66}$$

Where ΔG_{sol} is the Gibbs free energy of insertion of the water molecules into bulk water, $N\mu_{sol}$, where μ_{sol} is the chemical potential of a water molecule in bulk water. The Gibbs free energy of the system, ΔG_{sys}, can be equated to the Helmholtz free energy of the system, ΔF_{sys}, as the effect of pressure on the Gibbs free energy under standard conditions is negligible.[94]

The Helmholtz free energy of a system, ΔF_{sys}, is the combined energy of introducing the water molecules into a coupled ideal gas system, ΔF_{ideal} before transferring the water molecules from the ideal gas to the system ΔF_{trans}. From this with Equation 1.66, the Gibbs free energy of binding can be determined;

$$\Delta G_{bind}(N) = \cancelto{0}{\Delta F_{ideal}(N)} + \Delta F_{trans}(N) - \Delta G_{sol}(N) \tag{1.67}$$

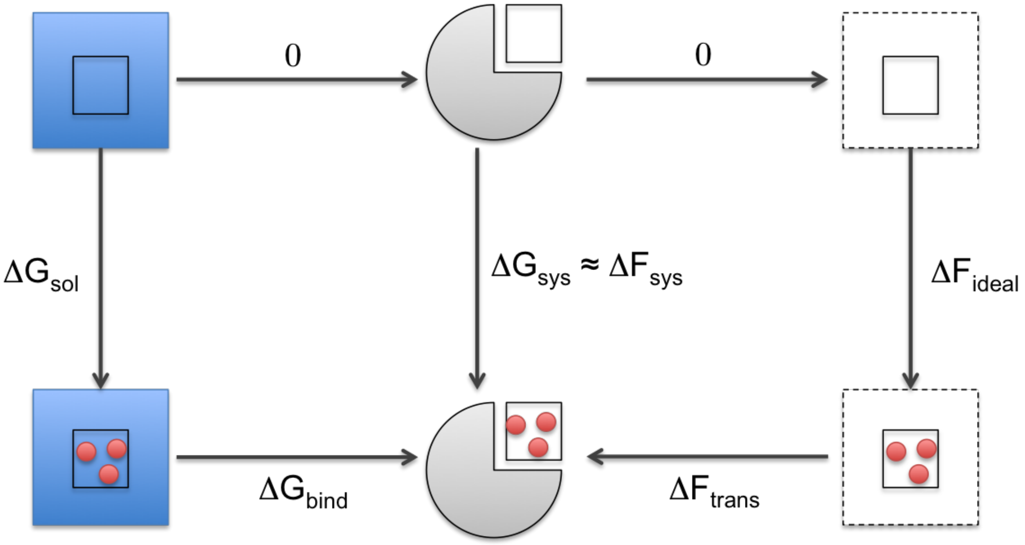


Figure 1.5: Thermodynamic cycle of the Gibbs free energy, where the system is coupled to bulk water, and the Helmholtz free energy, where the system is coupled to an ideal gas. The ΔG_{bind} can be calculated by following the alternate pathway around the cycle, using the approximation $\Delta G_{sys} \approx \Delta F_{sys}$.

In previous work, the ΔF_{ideal} was erroneously believed to have zero contribution to the energy. This error has been corrected, and the contribution of ΔF_{ideal} explicitly included, in Section 2.3.3. The effect of this mistake, and its correction, are discussed fully in Chapter 2. The overall thermodynamic cycle illustrating how the Gibbs free energy of binding can be calculated from the other thermodynamic contributions is shown in Figure 1.5.

computational implementation

Within this thesis, GCMC has only been attempted with water molecules. However, in theory the method could be used for any particle or species, and will be defined as a GC particle in this section. GCMC simulations are performed using in-house Monte Carlo biomolecular simulation program, ProtoMS.[95] GCMC differs from typical MC methods as the number of molecules, N is able to fluctuate throughout the simulation, by insertion or deletion moves. Within ProtoMS, a cubic GCMC region is chosen by the user. In practice, any shape of GCMC region could be used, but a cuboid has been used for simplicity. Within this region any molecules of the same species of the GC particle are removed during set up, and insertions and deletions are only attempted within this region. If a translation move of a GC particle results in moving the centre of mass of the particle outside the GC region, then the move will be rejected to prevent any GC water molecules from leaving the box. Theoretically these molecules are coupled to an ideal gas, however simulation of this ideal gas is not required. One method for simulating additional particles to insert into the system is to simulate an overlaying system of 'ghost particles'. Insertion and deletion moves attempt to vary the ghost molecules between an on and off state. These ghost molecules move through the system through normal Metropolis sampling methods. The ghost molecules are able to move freely as any MC move will be accepted as the species are non-interacting; therefore they take a random walk throughout the region. ProtoMS versions 2.3 and 3.0, as were used in the work by Ross et al., previously used this method of sampling ghost water molecules, but there is a tendency for this method to be slow to converge. If a water molecule has been deleted, then a vacancy will be present in the system and it is likely that the newly off water will not move far from this vacant site before it could be turned back on again. This hysteresis of ghost water molecules does not alter the results of the simulation, but slows down sampling of the site, and therefore convergence.[96]

The software has been updated such that the sampling of ghost particles is avoided. Before each insertion, the ghost particle will be assigned a random orientation and location within the GC region. The randomisation of the particle's position prevents the need for the random walk of the ghost particle. ProtoMS uses the widely-adopted method established by Norman and Filinov,[97] using three moves of molecule displacement, deletion and insertion. The displacement and rotation of GC atoms follows the typical Metropolis sampling, and the insertion and deletion moves are accepted using Adam's criteria, discussed in Section 1.4, using Equation 1.53. Of the three grand canonical specific moves, the insertion and deletion moves must be attempted with equal probabilities as to maintain the requirement of detailed balance. The third move of grand canonical sampling has no requirement to be proportional to the insertion or deletion moves, however convergence is found to be fastest when moves are sampled at a 1:1:1 ratio,[97] thus used as the default setting in ProtoMS.

1.5 Experimental comparisons

X-ray crystallography

Experimentally, protein structures can be determined by a range of methods such as X-ray crystallography, NMR, fiber diffraction and electron microscopy. While other methods are increasing in popularity, X-ray crystallography is the most commonly used, and is the only method considered herein. Elucidation of crystallographic structures is not trivial, as both protein crystallisation and solving the electron density is difficult. Protein crystallisation is difficult as proteins are inherently sensitive to biological conditions; temperature, pH, ionic strength, metal ions, inhibitors, cofactors and the presence of other small molecules.[98] Protein crystallisation is attempted by performing large-scale matrix trials to attempt to find any conditions that stimulate crystallisation, which is then iteratively improved to grow crystals of sufficient quality for X-ray studies. Some proteins are more difficult to crystallise than others, such as intrinsically disordered proteins, or membrane proteins that are largely hydrophobic and therefore generally insoluble. The bias in crystallisable proteins means that well-behaved targets are over represented, while membrane proteins, which represent 20-30% of the proteomes[99] of most organisms make up only 1% of the protein data bank (PDB).[100]

Crystallising the protein is only the first hurdle; the primary result of an X-ray experiment is the electron density, which needs to be correctly assigned to atoms of the molecule. The atomistic map of the model is then refined by optimising the fit between the expected electron density of the model (F_{calc}) to the experimental electron density (F_{obs}).[101] The level of agreement between F_{calc} and F_{obs} is measured by the R-factor, which is a measure of the global accuracy of the model. This refinement of the atomistic model can be performed by using computational algorithms such as least-squares, however the models can become trapped in local minima, and need intervention from qualified crystallographers to achieve the best agreement.[102] Along with the atomistic model, each atom is assigned a temperature factor (B-factor, which is different to the Adams' parameter B). B-factors describe the isotropic amplitude of displacement of an atom, within a range of 2 to 100 Å^2, and indicate how mobile or disordered an atom is, relative to the rest of the structure. Anisotropic B-factors are available for some structures, and can indicate a relationship between the structure of a molecule and its dynamics,[103] however the crystallographic data needs to be high quality to justify modelling anisotropic B-factors for a given structure.[104]

While much of the model refinement process is automated, human input can be required to find the model that is the best fit to the electron density. This human input can be subjective, and can result in slightly different models depending on the experimentalist. One example of this subjectivity is the case where the same high-quality electron density was given to two experienced crystallographers. However, in their resultant atomistic models, over 50% of the assigned water molecule locations differed by a distance greater than 1.0 Å.[105] As the automation process in crystallography continues to improve, the bias of the crystallographer in the results should be reduced. Efforts have been made to post-process electron densities available in the PDB in a project called PDB_REDO.[106] These re-refined structures are publicly available online[107] and have all been produced without human intervention (although the inherent design of the software will introduce some degree of human influence), but also can improve structures that were processed with older generations of software. While PDB_REDO is an interesting project; it is limited in its application to our interest in active site water molecules. PDB_REDO will attempt to reposition assigned water molecules, or remove clashing ones but does not currently support the addition of missing water molecules.

Crystallographic structures are used to assess the accuracy of computational methods of active site water locations. If active site water locations in crystallographic structures are not reliable, then this can make it difficult to determine if a computational method is functioning optimally. Efforts can be made to quantify how reliable a crystallographic water molecule is by measuring the underlying quality of the electron density in the local region. Various methods exist to assess the validity of the model, including real-space R (RSR),[108] and real-space correlation coefficient (RSCC).[109] RSR compares the calculated and observed electron density for a grid placed over the atoms of interest. RSCC is the calculated correlation coefficient of the RSR. Both of these real-space methods are limited, as both methods rely on, and are sensitive to, a choice of atomic radius as this defines its extent in the electron density. The atomic radius can either be fixed based on atom type, or a function of the atoms' B-factor. This means that both the RSR and the RSCC are strongly correlated with the metrics used for the model i.e. B-factor. As the definition of the atomic radius was not rigidly defined in the original publication, issues have arisen with results varying with differing software packages to calculate supposedly the same metric.[110]

Methods have been developed that calculate the quality of crystallographic model by considering the difference density map. $f_o - f_c$ and $2f_o - f_c$ are the difference maps (f_o and f_c the observed and calculated electron densities respectively), which can indicate regions where there is electron density with no atom assigned, or where there are atoms assigned with no supporting electron density. These difference maps can be useful in assessing the precision of a model, and the real-space difference density Z score (RSZD) is a χ^2 test for these differences to measure the normalised difference in density.[110] The real space observed density score (RSZO) is a measure of the signal-to-noise ratio in the RSZD and is a measure of the precision of the proposed model. RSZO scores regions of well defined electron density as greater than 1σ, where σ is the standard deviation in the electron density. RSZO should be more reliable than other real space methods, as the atomic radius definition is clearly defined; calculated using using B-factor, element, charge, and structure resolution, but with the B-factor being less correlated to the final metric than for RSR and RSCC.

Electron density for individual atoms (EDIA) is another metric to assess the quality of the underlying electron density for a given atom. This was initially developed and used to calculate electron densities of 2.3 million crystallographic water molecules,[111] before being generalised to all atoms in a crystal structure.[112] In the EDIA calculation, the atomic radii used are taken from a calculated table, depending on the element, charge and resolution of the structure. The table is generated from all available PDB structures, and the average B-factor is available based on an atoms element, charge and resolution, for resolutions between 0.5 and 3.0 Å in steps of 0.5 Å. The table of atomic radii used have been determined from the average B-factor for all PDBs within the given resolution. Using the average B-factor over each set of resolutions of PDBs should avoid issues with the constrained optimisation of B-factors for a given structure. The EDIA score sits on a scale between 0 - 1.2, where the higher the score, the better the electron density supports the position of the atom. An EDIA score below 0.8 suggests that there is not enough electron density to support the location of the atom, with this value chosen based on inspection of many structures and electron densities.[112]

affinity experiments

The usefulness of an organic molecule will depend on many factors from its adsorption, distribution, metabolism, excretion, toxicity (ADMET), its affinity and its specificity. While computational methods exist to attempt to model all of these metrics, only the affinity will be considered in this thesis. The affinity of a reversible ligand (L) to a protein (P) at equilibrium, can be considered;

$$P + L \rightleftharpoons PL \tag{1.68}$$

Where PL is the bound complex. The rate of association and dissociation are respectively calculated as;

$$rate_{ass} = k_{+}[P][L] \tag{1.69a}$$

$$rate_{diss} = k_{-}[PL] \tag{1.69b}$$

where [X] indicates the concentration of species X. The rate constants of association and dissociation are shown as k_+ and k_-. The association step is a second order reaction, as it depends on the concentration of two species. The rates of second order reactions are often dominated by the rate of collision, rather than the likelihood of conversion into product. The rate of collision in the case of a small molecule and a protein is determined by the size of both species and the size of their interaction surface. The association rate constant for a protein-ligand complex is therefore fairly constant, typically within the range of $10^6 - 10^7 Ms^{-1}$.[113] The dissociation step is first order, depending only on the concentration of [PL]. The dissociation rate constant, k_-, is the probability of the ligand to unbind from the complex within a given time. The equilibrium of the reaction shown in Equation 1.68 is the point at which $rate_{ass} = rate_{diss}$, where the following holds;

$$k_+[P][L] = k_-[PL] \tag{1.70}$$

and the equilibrium constant (K_{eq}),

$$K_{eq} = \frac{k_+}{k_-} = \frac{[PL]}{[P][L]} = K_d^{-1} \tag{1.71}$$

where the larger K_{eq}, more of the species are in the associated, PL, state. K_{eq} has units of M^{-1}, if the activity is neglected. Inverse molar units are unintuitive, therefore the dissociation constant K_d, which is the inverse of K_{eq} is more commonly used. Small values of K_d, which uses units of M, typically indicate a slower dissociation rate, and therefore a higher affinity of the ligand. The K_d of a reaction can be related back to the Gibbs free energy change of the reaction,

$$\Delta G^{\ominus}_{bind} = RTln(K_d) \tag{1.72}$$

where R is the gas constant and T is temperature.

The affinity of a ligand to its target can either be calculated using equilibrium, or kinetic experiments. Equilibrium assays afford the rate of the association and dissociation reactions, as a function of the concentration of one of the reactants. Recording the concentration of product as a factor of reactant concentration should result in a hyperbola correlation, from which K_d can be determined from the point of half maximum.[113] Kinetic experiments are more involved; rather than simply changing the concentration of a reactant and recording the concentration of the product as performed in equilibrium experiments, kinetic experiments involve altering the conditions of the experiment and monitoring the time taken to return to the equilibrium distribution. For a dissociation rate constant, this can be calculated by monitoring the displacement of a fluorescence labelled ligand on addition to the system of a non-labelled ligand.[113] Kinetic experiments allow for the calculation of k_+ and k_-, which in turn can be related back to K_d using Equation 1.71.

Equilibrium or kinetic experiments can be performed using optical assays. One example of an optical assay is where the fluorescence intensity is measured as a function of a reactant concentration. Proteins fluoresce due to aromatic moieties and disulfide bonds within their structure, with the strongest response occurring from tryptophan residues.[114] These naturally occurring fluorescent groups are known as intrinsic moieties. In some cases, a shift in fluorescence can be seen on ligand binding, which can be used to monitor [PL] within the experiment. If there is no tryptophan present (it accounts for ∼1.3 % of amino acids in vertibrates[115]), or ligand binding does not shift the fluorescence, then extrinsic fluorescing moieties can be used. This involves tagging a reactant with an extrinsic dye molecule, that allows for an optically measurable response. This is inconvenient as creates additional synthetic work, and makes the possibly incorrect assumption that the extrinsic moiety does not alter the binding of the ligand.

Surface plasmon resonance (SPR) methods are kinetic experiments; allowing for the measurement of rate constants, rather than just K_d. SPR is beneficial as it does not require the labelling of any of the species involved. The protein is immobilised on a sensor surface, over which a continuous flow of ligand is passed. As the ligand molecules bind to the protein, the refractive index of the surface shifts depending on both the mass of the ligand and the K_d. As it is possible to monitor the refractive index as a function of both time and ligand concentration, the rate constants can be elucidated by the method.[116] Limitations of SPR include the unquantified effect immobilising the species has on the association, and can give incorrect results if the reaction is not bi-molecular, and it can be difficult to calculate k_+ rates faster than $10^6 Ms^{-1}$ or k_- rates outside the range of $10^{-5}-1s^{-1}$.[116]

Isothermal titration calorimetry is another method able to elucidate more than just the K_d of a binding event.[117] Using two cells (one of which is a sample cell, the other, a reference cell) that are thermally coupled using a thermally conducting material within an adiabatic system, where the energy remains constant, the heat evolved from a binding reaction is measured. The heat evolved is monitored by recording the power required via a reference heater to maintain the same temperature in the two cells. This allows K_d, the stoichiometry of the reaction and the enthalpy ΔH to be directly determined. The Gibbs free energy (ΔG) can be calculated from K_d using Equation 1.72, which can in turn be combined with ΔH to indirectly calculate the entropy of binding. In addition, the thermal heat capacity (ΔC_p) of the binding can be calculated by recording the temperature dependence of the enthalpy. ITC is commonly used in the drug discovery process as elucidation of additional thermodynamic properties can be useful for rationalising Structure Activity Relationships of protein-ligand complexes.[118]

1.6 Current errors in computational modelling

"*All models are wrong, but some are useful*"[119] — George Box

The applications of computational chemistry are broad, and computational methods are able to contribute to science in many ways; from large scale *in-silico* screening of small molecule libraries, to high-level quantum mechanical simulations to elucidate a reaction pathway. While there are many examples of computational simulations correctly modelling reality, there are many occurrences where they may go wrong, and errors can occur. Errors that can arise belong to five categories: error in the force field used, incomplete sampling of the model, incorrect model generation during set up, mistakes in underlying theory, and computational bugs. These will be discussed in turn.

computational errors

Computational errors, i.e. programming errors, will exist in every software package. While never fully avoidable, they can be limited by ensuring coding best practises. Best coding practise involves constant testing of code during development, on a range of systems. If changes or new functionality is added to a package, testing of seemingly unchanged sections of the code is also required, to ensure that additions do not adversely affect other functionalities. Something that is often missed when developing software for biomolecular simulations, is the testing of the code on simplified test systems, such as Lennard-Jones fluids or bulk water, as it can often be easier to spot errors in less complex, faster to converge systems. Analytical result may be available for simple test systems, to provide reliable comparison. One such example is the high-profile disagreement between two renowned scientists, groups were observing differing phase-states of water at the same conditions.[120] The seven-year dispute was only resolved when it became apparent that one groups' simulations were occurring at a different temperature than which they believed. Issues can also arise if software is used on hardware on which it has not previously been implemented on, but this can be prevented by software such as Docker, which are linux containers that allow for consistent development platforms through virtual machines that are consitent at the operating-system level.[121]

force field errors

Within molecular modelling, a system is treated on an atomistic scale, and atoms are assigned bonded (bonds, angles, dihedral) and non-bonded (van der Waals (σ,ϵ) and electrostatic (q)) terms. A set of atom terms, known as a forcefield, are parameterised to reproduce experimental properties. Errors in a model can arise if the model is being used beyond the properties or conditions initially intended. Some atoms cause particular issues for force field parameterisations; such as charged ions, or if the fixed point charge does not capture the electronic structure of the atom. A huge range of varying parameters exist in the literature for one ion type.[122] Another issue that can occur with force fields is mistakes in the atom typing during parameterisation, due to large redundancy, and counter-intuitive differences in atom types. Methods to avoid atom types have begun with the Open Force Field Consortium via the force field format, SMIRNOFF.[123]

Other force field related issues can arise if the level of theory used in the model is not appropriate for the issue of inquiry. Higher-level quantum simulations may be more appropriate for certain questions. Polarisable forcefields exist which can better respond to electronic influences.[124] These should be particularly useful for protein-ligand binding, X-ray crystallography and other cases where correctly modelling the electrostatic properties are necessary. The additional polarisation terms require even more parameters to be optimised during force field development, and the additional terms can reduce the speed of simulation.

sampling limitations

The discord between the timescales achievable computationally, and those timescales at which biologically interesting processes occur, are frequently discussed. However, the timescales and complexity of systems modelled using computational chemistry continue to increase. One factor causing this is software improvements, with design of methods that are able to speed up the rate of sampling of a system. Metadynamics is able to encourage a system over high energy barriers using biases,[125] coarse-graining can speed up simulations by reducing the number of interacting parameters[126] and many other enhanced sampling methods can overcome energetic barriers.[127,128] These methods allow the progress of a simulation, or its effective timescale, to increase with a given amount of computational resource.

The other factor in the increase of available computational timescales is the improvements in available hardware. The increase in achievable timescales associated with Moore's Law, which states that computational power doubles every ∼18 months, owing to the increase in the number and speed of transistors in integrated circuits.[129] Moore's Law has shown to be reliable since its prediction in 1965. Another achievement of computer science is the invention of GPU's and introduction of parallel computing, initially designed for the video game industry, but has been repurposed for use in biomolecular simulations.[130] These computational advances have significantly improved the timescales achievable by simulation methods much beyond the first example of a microsecond long simulation 20 years ago.[131]

One of the major hardware developments is the specialised MD machine, Anton, that contains custom-designed GPU chips.[132,133] Anton has demonstrated its ability to simulate large systems for long timescales, and of particular note has been its demonstration of unbiased binding simulations of ligands to their protein targets.[134] While Anton is a remarkable machine that is able to lengthen simulation timescales, its usefulness will be limited as long as access to the technology is restricted.

Despite the improvements in software and hardware, computational timescales are still significantly shorter than many biological processes. Improvements in both software and hardware are likely to continue, as the field matures.

model errors

Errors can occur if the model used in a simulation is not correct. Crystallographic structures are usually the starting point for building a model for simulation. The process from a crystallographic structure to a computer simulation is not yet a black-box method, but requires human input from a computational chemist. Hydrogen atoms are not observed in crystallographic structures due to their low electron density, and must be added. This can be tricky for titratable functional groups, such as arginine, histidine, lysine, aspartic acid, and glutamic acid. Histidine requires particular thought as three protonation states exist (δ, ϵ or both protonated) and its rotameric state is also unclear as carbon and nitrogen have similar electron densities. The same is true for both asparagine and glutamine, which have isoelectronic rotamers. Constant pH simulations (CpHMD)[135] can aid this, allowing for titratable sites to be protonated correctly according to the defined pH. This can work for either protein or ligand functional groups, and can limit the assumptions that are made about the locations of protons when setting up a simulation.

Particularly mobile regions of a protein can be difficult to observe in the electron density due to blurring and can result in missing side-chains or residues. Many crystallographic structures are missing their termini. These missing residues can be built in to the structure, but the uncertainty in the atomic positions will be high, and the more residues are missing, the uncertainty will increase. Amino acids within the sequence of the crystallised structure may be non-native if there are cloning artefacts. The crystallographic conditions, such as the pH and temperature can also alter the structure of the protein from its native, solvated state. A model error that can be particularly problematic is the assignment of water molecules, or other small molecules that are present due to the experimental conditions such buffers or solvents. Owing to sampling limitations discussed previously, it is unlikely that a system will be able to diffuse far from the local minima of its starting position. This can cause errors where electron density is incorrectly assigned, assigning atoms where there is little supporting density, missing atoms where there is electron density, or assigning the wrong small molecule to the density that is available. Enhanced sampling methods can help with understanding where small molecules should be in a system. GCMC is useful for solvating protein-ligand complexes, highlighting where water molecules might have been erroneously added or missing from a structure. Saltswap is able to sample distributions of salt concentrations, to correctly account for the locations of biologically relevant ions.[136] Both of these methods attempt to correct the discrepancy between the crystallographically available structure, and the structure of biological relevance.

Two things can aid with these errors in the simulation model. Either improvements can be made to the experimental model or the computational method. Advancing methods such as neutron diffraction[137] or cryoEM,[138] and ensuring best practise assignment of those results. Computational methods can also play a role in reducing modelling errors. Methodologies that reduce the assumptions made during model building will reduce errors. GCMC removes assumptions of active site water locations, Saltswap removes assumptions of salt concentrations, CpHMD reduces assumptions of the protonation states, and enhanced sampling methods combined with sufficient simulation time that allow a system to rearrange itself if incorrectly modelled. These methodologies do however use their own assumptions, but these remove a degree of human error when decisions are made by the user.

theoretical errors

Theoretical errors, fall into two categories; accidental and intended. Accidental errors are difficult to find in the literature, as they are unlikely to be largely publicised. There is a theoretical error presented in this thesis, Chapter 2, where an energetic contribution was theoretically overlooked. The error was missed empirically due to the level of noise from the results. Accidental theoretical errors may occur when the noise of simulation results is such that the theoretical error is indiscernible. Issues such as this can be difficult to spot, but repetition of simulations, and good data for comparison, whether experimental or computational, can indicate if something is amiss. Theoretical 'errors' can also exist when a conscious decision is made to approximate a component of a simulation, whether this be as simple as applying a cutoff for non-bonded contributions, or arbitrarily choosing the free energy penalty for water molecule binding to be 7 kcal·mol^{-1}.[72] Arguably, any force-field error could also be considered as a theoretical error, if a molecular interaction has been approximated to some degree.

In conclusion, there are broadly five main issues that determine the accuracy and precision of a computer simulation. Some may be alleviated by improving computational power, improving the force field used in a simulation. Computational and theoretical errors can be difficult to spot, but their likelihood can be reduced by ensuring best practices, and using reliable data for comparison and benchmarking. Errors in the simulation model will improve with experimental developments that allow better understanding of the atomic positions of a structure or an experimental ensemble of structures. Another way to improve the model of a system is through intelligent computational methodologies such as constant pH simulations, that reduce the reliance of the results on the initial model building, through adaptively correcting the model during the simulation. GCMC is an example of this, where the enhanced sampling of active site water locations and occupancies is made possible through coupling active site water molecules with an artificial reservoir.

Many topics have been introduced here, starting with the basics of computational simulation through discussion of both molecular dynamics and Monte Carlo simulations in Section 1.1. It was illustrated how free energies can be calculated from sampled microstates of ensembles, rather then requiring the full partition function to be evaluated. Absolute free energies require evaluation involving an e^{E} term, which in practise will cause the results to vary significantly as more states are considered. Relative free energies ($e^{\Delta E}$) are significantly more viable to determine from simulation.

Practical methods involved with calculating relative free energies was introduced in Section 1.2, both in discussion of rigorous free energy methods (TI, BAR and MBAR) and the practicalities of various restraints and constraints used within simulation, which will be both used and discussed in Chapter 2.

The importance of water molecules for rational drug design has been discussed, and a selection of other published methods that are able to calculate the binding affinities of active site water molecules have been introduced in Section 1.3. Following in Section 1.4, the theoretical basis of simulating in the grand canonical ensemble is shown, as well as how GCMC can be used to locate active site water molecules. The binding affinity of GCMC water molecules can be determined by using the GCI Equation, Equation 1.61.

Validating computational methods requires comparable data, whether that be from other computational results, or through comparison to experimental data. Both comparing results to other computational methods and comparison to experimental data will be used. Both experimental methods of calculating binding affinity and X-ray crystallography are discussed in Section 1.5, which have been used to validate GCMC water placement and free energy calculations in Chapters 3 and 4. Finally, the current state of computational simulation methods was summarised, with a discussion of regions of potential errors and limitations arise in Section 1.6.

In the first results chapter, Chapter 2, RE will be introduced into GCMC simulations. RE between B values improves the reliability of binding free energies of water molecules to such a degree that an issue in the accuracy of the results is apparent. This inconsistency between binding free energies when calculated using GCMC with RE when compared to double-decoupling calculations led to a re-derivation of the GCI Equation.

While GCMC has previously been validated on a small set of systems, Chapter 3 presents a curated dataset of 105 protein-ligand complex of FDA approved drug molecules. The dataset has been used to test the performance of GCMC on systems of pharmaceutical interest — this is the largest validation of a simulation-based methodology for locating active site water molecules. Not only is the success of GCMC presented for a diverse dataset of relevant structures, but discussion focuses on the difficulty in quoting a single value for the success, and how this can lead to difficulties when comparing between different published methods that are simulated on different datasets, and analysed with different protocols.

While knowing where, and how stable, active site water molecules are is important, this is all for the primary goal of understanding the effect that water can have on ligand affinity. Knowing the location and stability of an active site water molecule is not necessarily informative as to if displacing said water molecules will have a beneficial effect on the ligand's affinity. GCAP allows for relative binding free energies of ligands to be calculated with dynamic sampling of active site water molecules. GCAP allows for ligand affinities to be accurately calculated, particularly in cases where the location of water molecules is unknown, or if the two ligands considered bind with differing water networks. The GCAP method and the results for two protein-ligand systems are demonstrated in Chapter 4.

Chapter 2

Water network binding free energies

2.1 Introduction

GAR implemented replica exchange between neighbouring B values for GCMC simulations in ProtoMS. All simulations were performed by HBM, the disagreement between GCMC and DD methods were empirically observed by HBM, and theoretically proven by GAR.

GCMC can determine the binding free energy of networks of water molecules through performing a titration where the system is simulated at a range of chemical potentials.[54] The binding free energies of water molecules is dependent on fitting multiple logistic functions to the titration results, which was introduced in Section 1.4. The logistic function is then used to determine the binding free energy of the water network using Equation 1.61. The smoother the results, the smaller the error in the fit. Figure 2.1 is the titration result of BPTI, which has a network of three water molecules in a small pocket. Figure 2.1 illustrates typical GCMC results, where it is clear that the noise in the data will result in binding free energies with large associated errors.

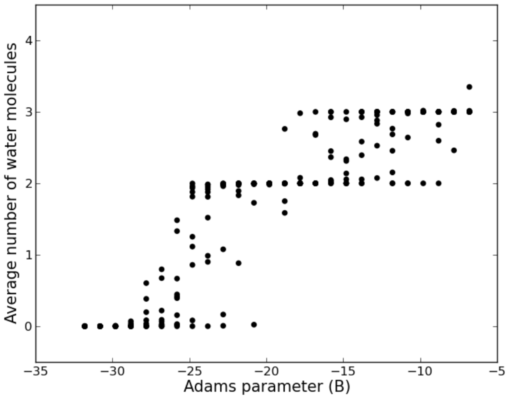


Figure 2.1: GCMC titration data for BPTI system, without replica exchange. Each point corresponds to the average number of water molecules at a given B value. The first 200,000 MC steps have been excluded as equilibration. Plot shows ten titration repeats for the system.

As the results are noisy, fitting to the data to afford a reliable binding free energy is difficult. The noise in the GCMC results is the motivation for introducing RE between neighbouring B values to the methodology.

This chapter will outline the re-validation of the GCMC methodology and theory, following the introduction of RE of B values in ProtoMS. RE between neighbouring chemical potentials reduces the variance of calculated binding free energies, without notable change to the median values. This reduction in noise has highlighted previously unobserved discrepancies between GCMC results and the gold standard method, DD. This discrepancy will be illustrated in Section 2.3.2. This discrepancy led to improvements both to the computational implementation of GCMC, and reassessment of the underlying theory, Section 2.3.3.

The theoretical developments result in the determination of an updated GCI equation, Equation 2.7, which is the major result of this chapter. Two changes have been made to the equation; the addition of a volume term and the removal of the multiplicity term. Simulations with a 'toy' system Scytalone Dehydratase will demonstrate that the inclusion of the volume term results in binding free energies that are independent of GCMC box size in Section 2.3.5. The removal of the multiplicity term will be supported using calculations with both Scytalone Dehydratase; with two water molecules considered, and BPTI, Section 2.3.6. In all cases, the results have been compared the gold standard method for water binding free energies - DD.

RE has been implemented in the GCMC method and is illustrated in Figure 2.2. Throughout the simulation attempts are made to swap system configurations between neighbouring B values. This is an enhanced sampling method, equivalent to the exchange between neighbouring λ values used in free energy calculations.[139] A background to RE methods is found in Section 1.2. RE should enhance the sampling in GCMC simulations, as simulations at higher chemical potentials, where GCMC insertions are more probable are able to interchange with lower chemical potentials, which will have lower insertion acceptance rates. An attempt to

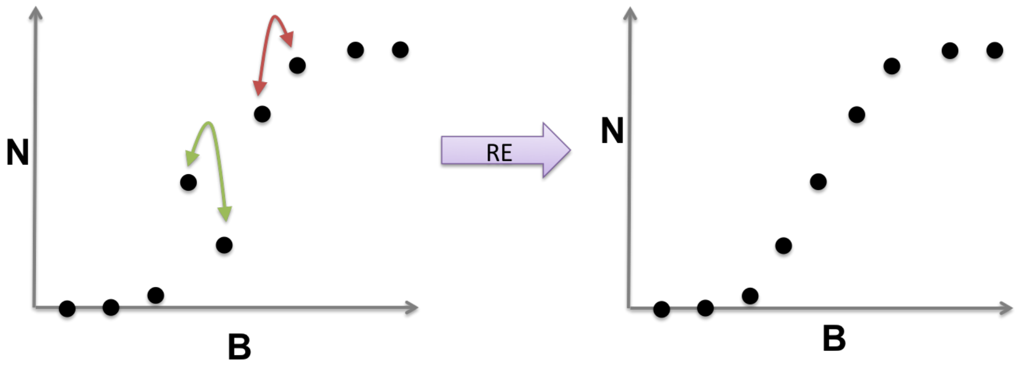


Figure 2.2: Illustration of replica exchange in B value. Two possible swaps are shown, in green where the two points are discordant and red where the two points are concordant. Both swaps will be accepted or rejected following the acceptance criterion, Equation 2.1. As the green swap is discordant, the swap will always be accepted. As the red swap is concordant, it will be swapped based on the probability derived from Equation 2.1.

swap neighbouring replicas is made every n moves, where the swap is accepted or rejected based on the following acceptance criterion:

$$P_{swap} = min[1, e^{(B_j - B_i)(N_i - Nj)}] \tag{2.1}$$

where B_x and N_x are the B value and water occupancy for the x^{th} replica respectively. As the GCMC insertion and deletion Metropolis conditions are dependent on B, N should theoretically increase with B. However in practice, owing to sampling limitations, sometimes this monotonicity condition does not hold. RE between neighbouring B values is essentially a test for the positive correlation (where the increase in one variable corresponds to the increase of the other) of the titration results. If the two neighbouring B values tested for a swap are discordant - that is the higher chemical potential replica has a lower water occupancy, then the attempt to swap the two points will always be accepted (the result is then concordant). If the two neighbouring replicas are concordant, then they may be swapped, based on the probability outlined in Equation 2.1, with the likelihood of the swap being proportional to the gradient between the two points as the gradient,

$\Delta x = \Delta y$, is in this case is $(B_j - B_i)(N_i - Nj)$. This has the effect of smoothing the results of N against B. The reduction in noise of titration data results in precise and reliable binding free energy values. In practice, RE is attempted every n MC steps, where n is typically the default output frequency of ProtoMS. At random, either the odd pairs, or the even pairs are chosen for an attempted swap, i.e. from the set [1,2,3,4] either [1,2] *and* [3,4] are attempted to swap, or [2,3].

The RE protocol is the same as that used to perform swaps in λ value in free energy simulations,[139] where the swap is dependent on the energy difference between neighbouring λ states. As calculating the energy difference between λ states is computationally expensive, swaps are attempted at the same time as simulation output as the energies are calculated anyway at this point. This means that the number of attempted swaps equals the number of results files output. With replica exchange in B, as both B and N are explicitly updated at every step of the simulation, the cost of any attempted swap is effectively free - excluding the cost of evaluating Equation 2.1 and any (message passing interface) MPI costs. RE in B could be attempted much more frequently as it is computationally cheap, but is kept at the frequency for λ for consistency. The following results indicate that this RE frequency is sufficient. *The RE protocol was implemented by GAR in ProtoMS and tested by HBM.*

2.2 Methodology

2.2.1 System set-up

For all proteins simulated, the amber14SB force-field has been used.[6] All ligands have been simulated using the gaff14 forcefield with AM1-BCC charges.

BPTI protein and its surrounding solvent system were set up by GAR from the 5PTI pdb entry.[54] The region studied is a solvated cavity where no ligand is bound. Calculations were performed on the apo structure.

SD protein structure used is from the 3STD PDB entry. The protein was scooped to a radius of 15 Å. The protonation and tautomer states of the proteins were determined using molprobity.[140] Two ligands bound to SD have been studied, ligands **1** and **3**. The 3STD PDB entry has the bound structure of ligand **2**, from which the other two ligands binding positions has been assumed by structural superimposition.

For all water simulated, the TIP4P force-field has been used.[141] Protein-ligand complexes were solvated using a half-harmonically restrained sphere of radius of 30 Å, with any crystallographic water locations retained. This includes solvating any sterically available active site regions.

2.2.2 Water binding affinities

Replica exchange

GCMC simulations were performed over a cavity of multiple waters (volume 5.0x4.0x8.0Å^3, origin: 29.0, 5.0, -2.0). 1M GCMC only equilibration, 1M full sampling equilibration and 100M production steps were performed. Various replica exchange frequencies were tested to compare to a no-RE protocol. The frequencies of attempted RE were 100,000, 200,000, 500,000 and 1,000,000. For each, a B-value range of -31.0 to 0.0 was used and was repeated 10 times. In the GCMC only equilibration moves are split equally between grand canonical insertion, deletion and sampling. When fully sampling bulk solvent, protein, GC insertion, GC deletion and GC sampling are split with a ratio of 461:39:167:167:167 respectively.

Table 2.1: Details of GCMC region used for each one-water system. The GCMC region is cuboidal. Range and increments of B values used for each set of calculations.

System	origin (x,y,z)	length (x,y,z) /Å^3	Bs
SD 1.a	24.141, 11.225, 32.916	4, 4 - 8, 4	(-26, -11, 1)
SD 1.b	27.913, 11.260, 28.713	4, 4 - 8, 4	(-26, -11, 1)
SD 3.a	24.141, 11.225, 32.916	4, 4 - 8, 4	(-10,+5,1)
SD 3.b	27.913, 11.260, 28.713	4, 4 - 8, 4	(-26, -11, 1)

2.2.3 Grand canonical integration

GCMC — single water

GCMC simulations were performed for a range of box sizes, with four repeats at each volume. The range of box sizes was generated by extending the GCMC box in 1 Å steps, over a 5 Å range along one axis. The box coordinates, and the dimension of extension are available in Table 2.1.

The protein is not sampled in these simulations, so a protein conformation from a previous fully sampling GCMC simulation where both of the waters are bound was chosen. Simulations of 20M MC moves were performed, with the first 4M steps excluded from analysis. No protein or ligand moves were sampled and bulk water was excluded for the SD simulations, with all Monte Carlo moves assigned to grand canonical insertion, deletion and grand canonical water sampling with equal probabilities. RE in B was attempted every 100,000 MC steps. As water b is expected to have a lower binding free energy with ligand **3**, water b was included as a solvent molecule in the GCMC of the water a region. For ligand **1**, the simulations were repeated both with and without the other water molecule. When present, the additional solvent molecule was sampled with an equal probability to the GC water.

Table 2.2: Details of GCMC region used for each two-water system. The GCMC region is cuboidal. Range and increments of B values used for each set of calculations.

System	origin (x,y,z)	length (x,y,z) /Å^3	Bs
SD 1.a+b	24.1, 11.2, 30.0	4, 8 - 13, 4	(-26, -11, 1)
SD 3.a+b	24.1, 11.2, 30.0	4, 8 - 13, 4	(-26, +5, 1)

GCMC — multiple waters

Calculations of the SD in complex with ligands **1** and **3** were performed, with a GCMC region covering both hydration sites *a* and *b* (volume 4.0x8.0x4.0 Å^3, origin: 24.1, 11.2, 30.0). The GC region was extended in 1 Å steps, over a 5 Å range along the y-axis. No protein or ligand sampling was performed, and bulk water was excluded. Simulations were repeated four times at each volume.

For BPTI, the GCMC results used were taken from previous simulations, where the method is outlined in Section 2.2.2.

double decoupling

For each water location found with GCMC, DD simulations were performed to determine the binding free energy of each water. DD was performed over 16 alchemical λ states, where the LJ and Coulombic terms were scaled simultaneously. Moves were split between protein, bulk water and decoupled water at a ratio of 402:98:1 respectively. The water molecules were decoupled sequentially, from weakest to strongest bound. Where the free energies of multiple waters are similar and the order of binding was unclear, calculations were repeated with a different order of decouplings. 500,000 equilibration and 40M production moves were performed for each water at each λ value. Each simulation was repeated four times. Soft-cores (soft66 in ProtoMS package)[33,35,37] were used for DD calculation with δ=0.2 and δ_c=2.0 used for the decoupled water molecule. The free energy to decouple the water from the system was determined using MBAR.

A harmonic restraint with a force constant of 2 kcal·mol^{-1}·Å^{-2} was used on the oxygen of the water being decoupled at all λ values. A gas phase correction of,

$$\Delta G^{gas}_{rest} = k_B T \ln \left(\frac{V_{sim}}{V^o} \right) \tag{2.2}$$

where

$$V_{sim} = \left(\frac{2\pi k_B T}{k} \right)^{\frac{3}{2}} \tag{2.3}$$

was applied to account for the removal of the restraint from the decoupled system.[48] This is analogous to the volume term introduced in the GCI equation, Equation 2.7. Prompted by the higher precision obtained in RE-GCMC and unlike our previous study,[54] the free energy penalty of applying the harmonic restraint in the bound simulation was calculated using Bennett's Acceptance Ratio method from 40,000 Monte Carlo simulations steps with six equally spaced λ values of the restraint—from 0 kcal·mol^{-1}·Å^{-2} to 2 kcal·mol^{-1}·Å^{-2}. No symmetry correction was applied to water molecules.

For SD, GCMC was performed at 16 equally spaced B values from -22.7 to -7.7. As the binding free energy of the water molecule with ligand **3** is unfavourable, higher B values are required to couple the water into the system; therefore for this ligand GCI was repeated for 16 B values from -12.7 to +2.3.

2.3 Results

2.3.1 Replica exchange in B

RE improves the monotonicity of GCMC titrations as well as reducing the variance in the calculated binding free energies of water molecules. As Ross et al. found the BPTI system the most difficult to converge in their orig-

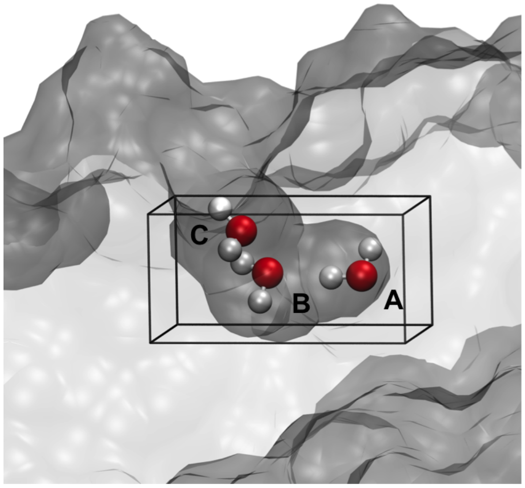


Figure 2.3: Hydrated pocket of the BPTI protein, containing three water molecules. GCMC region is indicated by a black box. PDB: 5PTI.

inal work on the GCMC method,[54] this system may benefit the most from RE. GCMC was performed on a small unliganded pocket of the protein that contains three water molecules, Figure 2.3. The frequency at which RE is attempted during a simulation is user defined, so frequencies of every one, two, five and ten-hundred thousand steps were trialled. These different RE frequencies have been compared to simulations with no RE. Ten repeats were performed at each frequency to improve statistical precision.

Kendall tau shows that the monotonicity of the results are improved with RE. The GCMC titrations from 10 results are shown in Figures 2.1 and 2.4, both without RE, and with a RE of 100,000. RE has the effect of smoothing the titration results results. The relationship between B and N should be monotonically increasing, due to the GCMC insertion and deletion acceptance tests. The Kendall rank correlation coefficient (τ) has been used to test the monotonicity of the two sets of results, where τ=1 indicates perfect positive monotonicity, τ=0.5 for random results, and τ=0 for perfect negative monotonicity. The τ of the non-RE data, shown in Figure 2.1, have a result and a standard error of 0.86

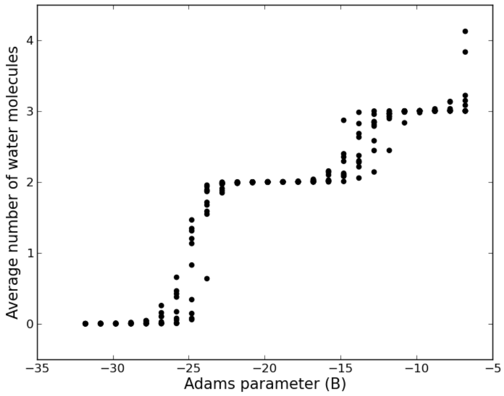


Figure 2.4: GCMC titration data for BPTI system with replica exchange in B every 100,000 steps. Each point corresponds to the average number of water molecules at a given B value. The first 200,000 MC steps have been excluded as equilibration.

(0.01) compared to 0.98 (0.00) with a RE rate of 100,000. The improvement to the monotonicity is unsurprising as the RE acceptance test, Equation 2.1, will favour results that are monotonic. Replica exchange in B is able to reduce the variance in $\langle N \rangle$ for a given B value between simulations, which results in GCMC titrations that are significantly smoother (Figure 2.4), as demonstrated by their improved Kendall τ correlation coefficient. As the function $N(B)$ is smoother, the analytical fitting of logistic functions to the titration is more reliable, providing binding free energies with a tighter distribution using the GCI Equation. The sum of logistic functions fit to the titration data take the form of;

$$N(B) = \sum_{i=1}^{m} \frac{n_i}{1 + e^{w_{0i} - w_i B}} \tag{2.4}$$

where m is the user-defined number of steps in the titration data, and n_i is the number of water molecules coupled in a given step, with an inflection point of w_{0i} and steepness of w_i. Both n_i and w_i are positive to ensure monotonicity of the function.

RE reduces the variance in calculated binding free energies for BPTI.

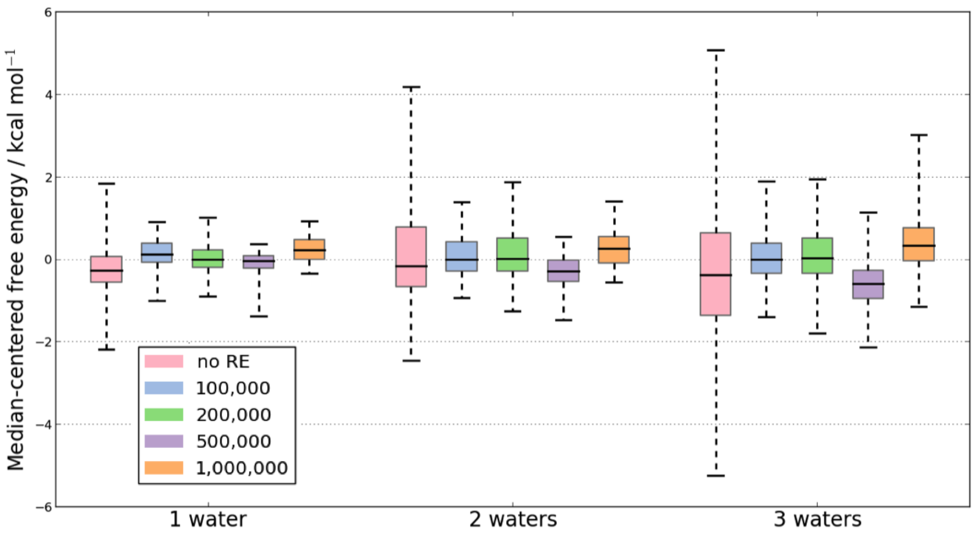


Figure 2.5: Boxplot of the median-centered free energies for each protocol, where errors have been calculated over 1000 bootstrapping samples of 10 repeats. In each case it is the free energy difference between an empty GCMC region, to a one, two and three water network, respectively. Replica exchange with GCI produces free energies that have a consistently tighter distribution than GCI free energies calculated without replica exchange.

The binding free energies of the water molecules in the BPTI system shown in Figure 2.5, are calculated by simulating at a range of B values and calculating the average water occupancy at those values. The median-centred binding free energies of each RE protocol, for each of the three waters is shown in Figure 2.5. The box-plots were generated by bootstrap sampling the titration data and calculating the binding free energy of each sample. A bootstrap sample consisted of one randomly sampled N value from the set of 10 repeats for each of the 32 B values and the titration curve was estimated as previously described. It is clear that both the range and inter-quartile range of the results are improved by RE. Including RE in the simulation reduces the variance of GCI binding free energies calculated. Based on these results no RE frequency appears to perform better than any other frequency, therefore a RE frequency of 100,000 has been chosen to further illustrate improvements to the results as it is the frequency at which results are printed. This reduction of variance will prove to be vital improvement in empirical results that will lead to the re-assessment of the form of the GCI equation. The acceptance rate for B RE swaps was 90 % for all RE frequencies attempted. This shows that the protocol is efficient, and that the replicas are well spaced for this system. The acceptance rate for exchanges was consistent for all of the RE frequencies considered.

2.3.2 Comparison of RE-GCMC results with DD

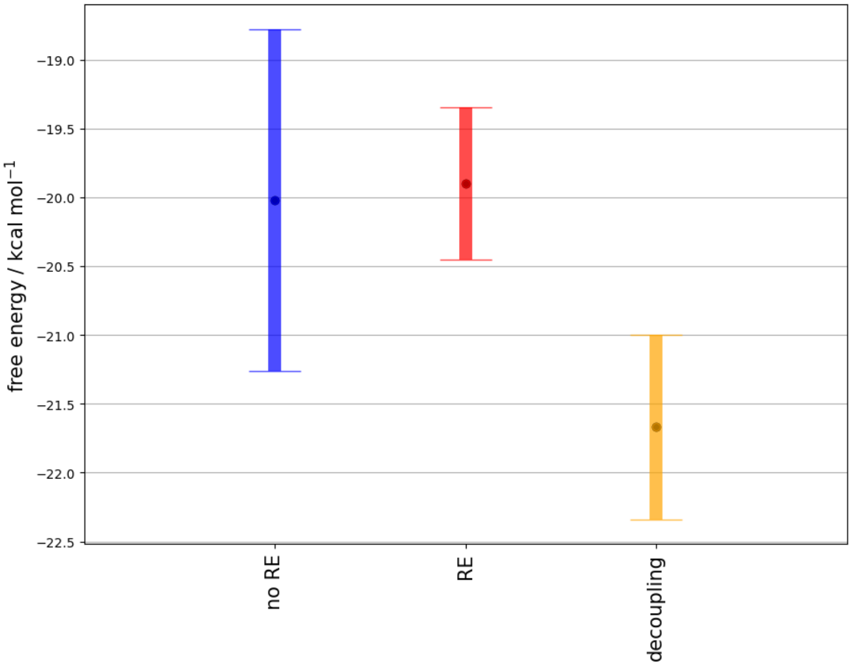


Figure 2.6: The binding free energies of the three water network in BPTI calculated using different methods. To highlight the intrinsic uncertainty of each method, the coloured bars indicate one standard deviation, as opposed to the standard error, over all repeats. Results are calculated using a protocol without RE (blue), with a RE frequency of 100,000 (red) and with DD comparison (orange), outlined in Section 2.3.

GCMC simulations without RE have sufficiently large errors that they erroneously appear consistent with DD. RE reduces the variance and indicates a disagreement between GCMC and DD. The methods used for DD and GCMC are such that the binding free energy of the three water network should be the same for both methods. For simulations without RE, the standard deviation of the results are large enough to indicate that the results without RE are statistically indistinguishable from the DD results, Figure 2.6. With the introduction of RE, the median binding free energy of the GCMC results does not notably deviate, however the variance is reduced such that it is clear there is a discrepancy between this and the gold standard DD results. The improvement of the method reveals an error in the determination of binding free energies via GCMC, which was previously masked by the noise of the simulation. This has led to the reassessment of the GCI equation, Section 2.3.3.

RE in B affords binding affinities with errors comparable to DD. In addition to revealing the discrepancy between the two methods, RE has improved the GCMC method such that the reproducibility of the simulations, i.e. the standard deviation between repeats is comparable to that achieved from DD. This means that GCMC is not only preferable due to its ability to calculate free energies of multiple waters simultaneously, without requiring hydration site information, but is also able to produce results as reliable as DD. Both methods, when the computational expense of additional GCMC moves and restraint calculations for DD are approximately comparable.

2.3.3 GCI equation

The mathematical derivation in this section was performed by GAR.

With the improvements in reliability of binding free energies evaluated using the GCI Equation, a volume dependence — that is a dependence on the calculated binding free energy of water molecules on the volume of the GCMC region is apparent. The volume dependence led to re-evaluation of the GCI Equation, which will be outlined in this section, before the problem is demonstrated in Section 2.3.5.

As GCMC is inconsistent with DD methods, and observed to have a volume dependence, the GCI equation is re-evaluated. The binding free energy of water molecules in the SD system was calculated with a range of GCMC volumes, which revealed that the binding energy of a water molecule was dependent on the GCMC volume. These results will be presented shortly, but first the source of the issue — the neglect of the Helmholtz free energy contribution from the ideal gas — is presented. This was erroneously overlooked when the noise of the simulation was large. This noise has been reduced significantly with the addition of B value RE, and further clarified by simplifying the simulation by not sampling the protein-ligand environment.

The GCI equation, as stated by Ross et al. is shown below:

$$\beta \Delta F_{trans}(N_i \rightarrow N_f) = N_f B_f - N_i B_i + ln\left(\frac{N_i!}{N_f!}\right) - \int_{B_i}^{B_f} N(B) dB \tag{2.5}$$

details of which are discussed in Section 1.4. This allows the transfer energy of $f - i$ waters from an ideal gas into the system to be calculated. From this, the relative binding free energy of water molecules can be determined, by accounting for the transfer free energy of those water molecules into bulk (μ_{sol}).

The volume correction can be understood by thinking about the proportion of insertion attempts that will be feasible. The volume dependence of the GCMC results using the above equation will be presented in the following sections. This empirical dependence illustrates that a volume term in the GCI equation is required to correct for this. Figure 2.7 shows two hypothetical model systems, where in both there are two sites in the system, of which only one is a hydration site and the other (grey) is not. The difference between these two systems is the volume of the GC region, illustrated with a red dashed line. Both systems are identical, except for the GC region over which insertions and deletions are attempted, and the binding free energy of the water molecule should be identical for both systems. Considering model A, all attempted insertions will occur on the feasible position, therefore all the attempted insertions will be feasible, and accepted based on a probability, where feasible means that the energies associated with the insertion will be finite. If the GC volume of this system is doubled to cover an inaccessible hydration site, only half of the attempted insertions will be feasible, as

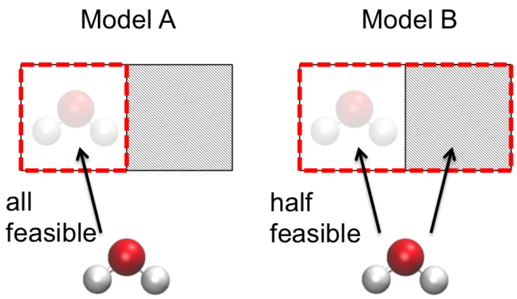


Figure 2.7: A two-site model system, where one site is a hydration site, and the other is not accessible to water (i.e. occupied by protein or ligand in a real system). The boundary of the GC region, in which GC insertions are attempted, is illustrated with a dashed red line. For model A, the GC region only covers the hydration site, whereas for model B the GC region covers both the hydration site and the inaccessible site. Here, feasible is used to indicate insertions that involve finite energy difference and therefore will be accepted with some probability. The grey, inaccessible site will result in infinite energies and therefore always be rejected.

the second site will give infinite energies and therefore an insertion move into this region will always be rejected. This means that for both systems, when simulated at the same chemical potential, system B will have fewer insertions accepted, and therefore a lower average water occupancy. The rate of accepted deletion moves is not dependent on the volume of the GC box. While it is possible that the volume of the GC region could be accounted for in the GC acceptance rates, it was found that the free energy results could be corrected with a *post hoc* correction.

Inclusion of the Helmholtz free energy for the ideal gas phase remedies the observed volume dependence. The equation for Helmholtz free energy of an ideal gas is shown in Equation 2.6a, such that the free energy difference to change the number of molecules in the gas is Equation 2.6b.

$$F_{ideal}(N) = k_BTln\left[\frac{1}{N!}\left(\frac{V_{gas}}{\Lambda^3}\right)^N\right] \tag{2.6a}$$

$$F_{ideal}(N_i \rightarrow N_f) = k_BTln\left[\left(\frac{N_i!}{N_f!}\right)(N_f - N_i)\left(\frac{V_{gas}}{\Lambda^3}\right)^N\right] \tag{2.6b}$$

If this is introduced to the GCI equation using the thermodynamic cycle shown in Figure 1.5, and the equality $\mu_{sol} = \mu'_{sol} + k_BTln(\frac{\Lambda^3}{V^{\ominus}})$ is used, the GCI equation becomes:

$$\beta\Delta G^{\ominus}_{bind}(N_i \rightarrow N_f) = N_fB_f - N_iB_i - \beta\mu'_{sol} - ln(\frac{V_{sys}}{V^{\ominus}}) - \int_{B_i}^{B_f} N(B)dB \tag{2.7}$$

Where the volume of the ideal gas, V_{gas} will be equal to the volume of the GC system V_{sys}, such that the . $V^{\ominus}$ is the volume of a water molecule in bulk water, 30.0 Å^3. This volume term is able to correct for the affect of the volume on the GC insertion rates. This term is analogous to the volume correction introduced by Gilson et al. that corrects for the energetic penalty of constraining or restraining a molecule that is being decoupled in a DD simulation and yield standard free energies.[48]

The multiplicity term in the bound state is equivalent to the multiplicity in the ideal gas state. Initially the multiplicity term, $ln(\frac{N_i!}{N_f!})$ was introduced to account for the inherent degeneracy present in the GCMC method – any GC water molecule can occupy any hydration site in the protein – a degeneracy that is not present in DD simulations, where each water molecule is constrained to its own site and no exchange is allowed. This term cancels when the Helmholtz free energy of the ideal gas is considered. While it is correct that there is a degeneracy for inserting GCMC water molecules into a system with multiple sites, this degeneracy is also present in the ideal gas phase of the thermodynamic cycle as shown previously in Figure 1.5, and therefore the effect cancels within the thermodynamic cycle, and does not need to be considered. This will be empirically supported by performing a set of simulations where the binding free energies of two waters are considered individually, and together using the GC method, Section 2.3.6. The following section will look at example systems to demonstrate the consistency between GCMC and DD, for single-water and multiple-water systems.

2.3.4 Equilibrium B value

The equation for B is shown again below, Equation 1.56, which was previously introduced in Section 1.4.

$$B = \mu\beta + ln\frac{V_{sys}}{\Lambda^3} \tag{1.56}$$

A network of water molecules is at equilibrium when the binding free energy is at a minimum.

$$\frac{d\Delta G^{\ominus}_{bind}(N)}{dN} = 0 \tag{1.65}$$

The Gibbs binding free energy can be determined from these following terms, however, ΔF_{ideal} is now recognised to contribute, as discussed in Section 1.4.

$$\Delta G^{\ominus}_{bind}(N) = \Delta F_{ideal}(N) + \Delta F_{trans}(N) - \Delta G_{sol}(N) \tag{1.67}$$

Where the mixing of Gibbs and Helmholtz free energies is due to the approximation $\Delta F_{sys} \approx \Delta G_{sys}$. In the thermodynamic limit;

$$\Delta F_{trans}(N) = \int_0^N \mu'_{sys}(N)dN \tag{2.8}$$

Where ΔF_{trans} is the Helmholtz free energy to transfer N water molecules from an ideal gas reservoir to the GCMC region. Substituting Equations 2.8, 2.6 and $\Delta G_{sol} = N\mu_{sol}$ into Equation 1.67 gives;

$$\Delta G_{bind}(N) = \int_0^N \mu'_{sys}(N)dN - k_B T ln\left[\frac{1}{N!}\left(\frac{V}{\Lambda^3}\right)^N\right] - N\mu_{sol} \tag{2.9}$$

Using Sterling's approximation, allows Equation 2.9 to be differentiated with respect to N.

$$\Delta G_{bind}(N) = \int_0^N \mu'_{sys}(N)dN - k_BT\left[Nln\left(\frac{V}{\Lambda^3}\right)^N + Nln(N) - N\right] - N\mu_{sol} \tag{2.10}$$

$$\frac{d\Delta G_{bind}(N)}{dN} = \mu'_{sys} - k_BT\left[ln\left(\frac{V}{\Lambda^3}\right) + ln(N)\right] - \mu_{sol} = 0 \tag{2.11}$$

Using $\mu'_{sol} = \mu_{sol} + k_BTln(\rho_{sol}\Lambda^3)$, and $\frac{N}{V} = \rho_{sys}$ it is possible to equate the chemical potential of the system, μ'_{sys}, and bulk solvent, μ'_{sol}, at the point of equilibrium.

$$\mu'_{sys} - k_BTln(\rho_{sys}) = \mu'_{sol} - k_BTln(\rho_{sol}) \tag{2.12}$$

The equilibrium water occupancy can now be simulated directly, rather than deterministically. Previously, neglecting the Helmholtz free energy of the ideal gas phase, led to the understanding that the excess chemical potentials of the system and solvent led to equilibrium. This meant that it was only possible to simulate at a range of chemical potentials, and calculate which satisfied Equation 1.65. However, as it is the chemical potentials, rather than the excess chemical potentials, that are equal at equilibrium, it is possible to determine the correct chemical potential, or B value *a priori*. From Equation 1.56 and $\mu'_{sol} = \mu_{sol} + k_BTln(\rho_{sol}\Lambda^3)$ is trivial to determine;

$$B_{eq} = \beta\mu'_{sol} + ln\left(\frac{V_{sys}}{V^{\ominus}}\right) \tag{2.13}$$

Simulations can be run only at the equilibrium B value, rather than requiring a full titration. The correct B value can be determined before simulating, using μ'_{sol}, which is a constant for a given water model, and using the volume of the GCMC region, V_{sys}, which is user-defined. This means that only one B value need be simulated to see the equilibrium location and occupancy of the water molecules, as opposed to the range of chemical potentials previously required.[54] This reduces the computational expense of the simulations. To determine

ΔF_{trans}, and therefore $\Delta G^{\ominus}_{bind}$, the function $N(B)$ is required to be integrated, so a simulation of a range of B values is required to explicitly calculate the water binding free energies. Knowing the equilibrium B value is still helpful for the simulations where a range of chemical potentials are simulated for several reasons. The range of B values to simulate can be determined based on B_{eq}, as if only favourable water molecules are of interest then only B values below the equilibrium value are required, which aids a more logical choice of simulation parameters. The equilibrium B value can also provide a sanity check, as the N that satisfies the minimum in $\Delta G^{\ominus}_{bind}(N)$ should correspond to the N simulated at B_{eq}. This analysis is now performed automatically.

The main advantage of the ability to determine the equilibrium B value for simulations is in the GCAP method, where GCMC is coupled to alchemical ligand perturbations, and will be discussed in Chapter 4.

2.3.5 Single-water system

Scytalone Dehydratase

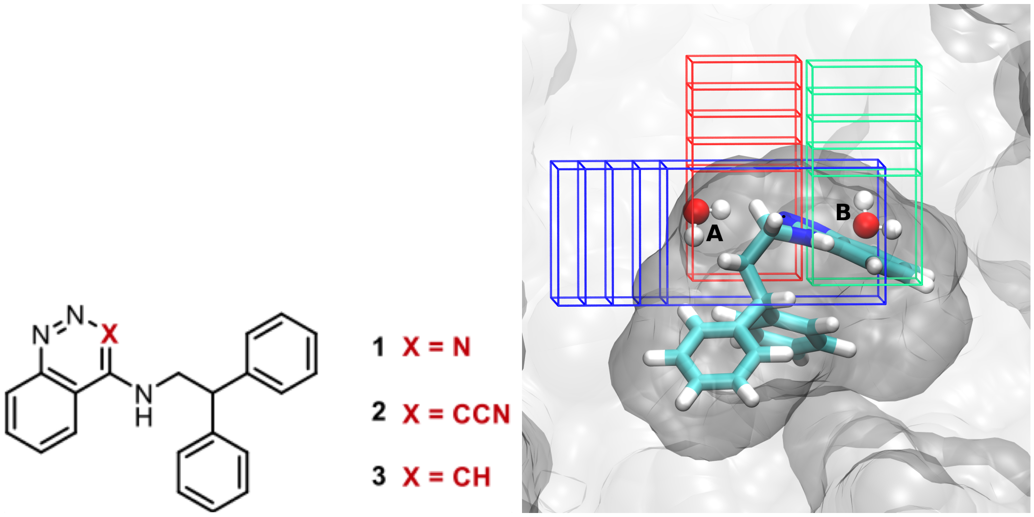
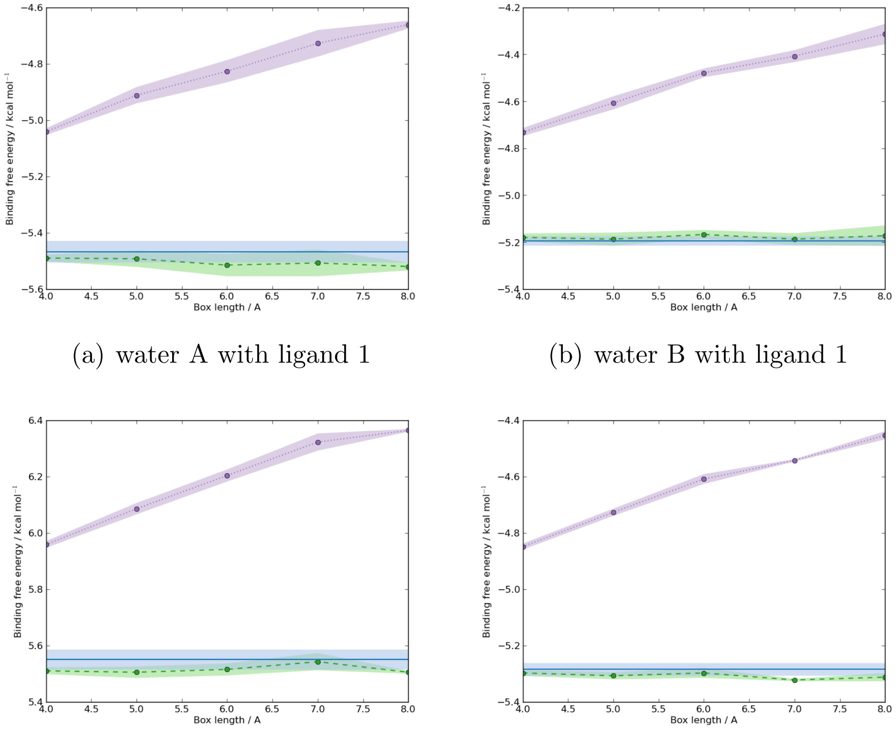


Figure 2.8: Structure of SD ligands, of which ligands **1** and **3** are considered herein. Ligand **3** bound to SD, with water A and B present. The active site of SD is shown with a transparent grey surface. The incrementally increasing GCMC boxes for each calculation are shown; A (red), B (green) and for the box encompassing both waters (blue). Each box repeatedly increased in 1 Å increments. The increasing volume of the GCMC region covers protein, not accessible to water.

GCMC calculations were performed on two single-water sites of the protein SD in complex with two ligands, **1** and **3**. For each system the calculation has been repeated with an increasing length of GCMC box, which increases the volume of the GCMC system (V_{sys}). These GCMC boxes are shown in Figure 2.8, with the red and green GCMC boxes for waters A (red) and B (green) respectively. For these simulations, simplifications were made to the regions of system that will be sampled as converged, precise results are more important for this validation than reliable experimental reproduction. No bulk water was simulated, and the protein and ligand system were treated as rigid. Only the two active site water molecules were sampled within these simulations.

When Equation 2.7 is used to calculate water binding free energies, the results are independent of the GCMC box volume. Figure 2.9 shows the binding free energy of each water in each system when calculated with DD, and both the old and new versions of the GCI Equation. The result obtained using conventional DD methods is shown with a solid line. The GCMC results illustrate a clear linear increase in binding free energy with an increasing GCMC volume when calculated with the previous GCI equation. This was overlooked in the GCMC method before, as the implementation of RE of B-values significantly improved consistency between repeats of the same system. As the protein and ligand are non-sampling, the error between repeats is reduced, allowing the volume effect of the method to be identified above the noise of previous calculations.

Changing the GCMC box volume changes the proportion of feasible insertions that are attempted, and therefore the insertion acceptance rate, which shifts the GCMC titration curve. Increasing the GC volume reduces the probability of attempting an insertion in the site of water binding. This causes a decrease in successful insertion moves, and therefore results in a lower average water occupancy for a given B value. Deletion moves are not proportional to the volume of the box, and so do not affect the result. Figure 2.10 shows the effect of the lower average water occupancy, where the titration curves are shifted to higher B values as a consequence of the increasing box volume. The binding free energy of the molecule is calculated from the integration of the fit, and therefore the right-shifted titration results in a weaker binding free energy. No successful water insertions have been made into the region of extension due to steric clashes with the protein, indicating that this is not a consequence of locating an alternate water location. This clearly illustrates a dependency of the binding free energy calculated on the volume of the GC region, when calculated using the previous GC result, Equation 1.61. This is a consequence of neglecting the Helmholtz free energy of the ideal gas phase. The updated GCI equation, Equation 2.7, contains the term $-k_BTln(\frac{V_{box}}{V^{\ominus}})$, which is able to correct for this artefact. Using the updated equations, the results from the simulations are shown in Figure 2.9 by the



(a) water A with ligand 1

(b) water B with ligand 1

(c) water A with ligand 3

(d) water B with ligand 3

Figure 2.9: Binding free energy of waters in SD. Dotted line (purple) - GCMC results using Equation 2.5 - without volume correction. Dashed line (green) - GCMC result using Equation 2.7 - with volume correction. Solid line (blue) - DD result. For each, the shaded region show one standard error calculated from four repeats.

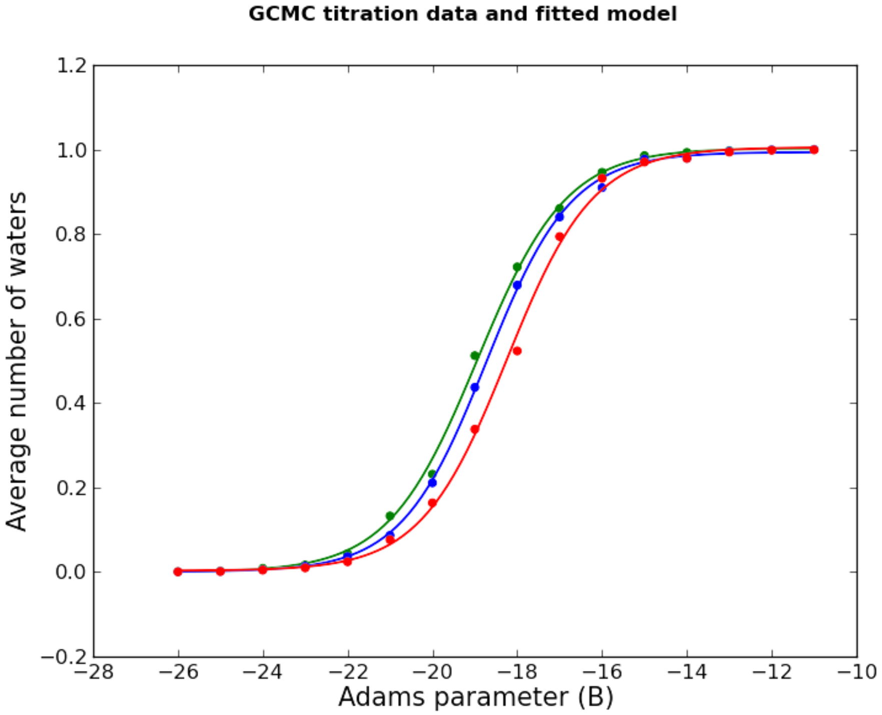


Figure 2.10: Plot of titration results for water A with ligand **1** bound to SD. Green, blue and red are results from GCMC box lengths of 4, 6 and 8 Å respectively. As the volume of the box is increased, the titration curve shifts to higher B values. This corresponds to lower binding free energies, calculated using Equation 1.61.

dotted line. This shows that the new theoretical result provides water binding free energies by the GCMC method that are both independent of box size and consistent with DD results.

Figure 2.11 shows the thermodynamic cycle of removing both water molecules A and B with both ligands, **1** and **3**. The GCMC binding free energies have been calculated using the new GCI equation, Equation 2.7. DD results have been corrected for the restraint correction used on each water molecule, discussed in Section 2.2.3. The GCI results are within 0.1 kcal·mol^{-1}of the DD results, and the results and standard errors are available in Table A.2. These results show that the

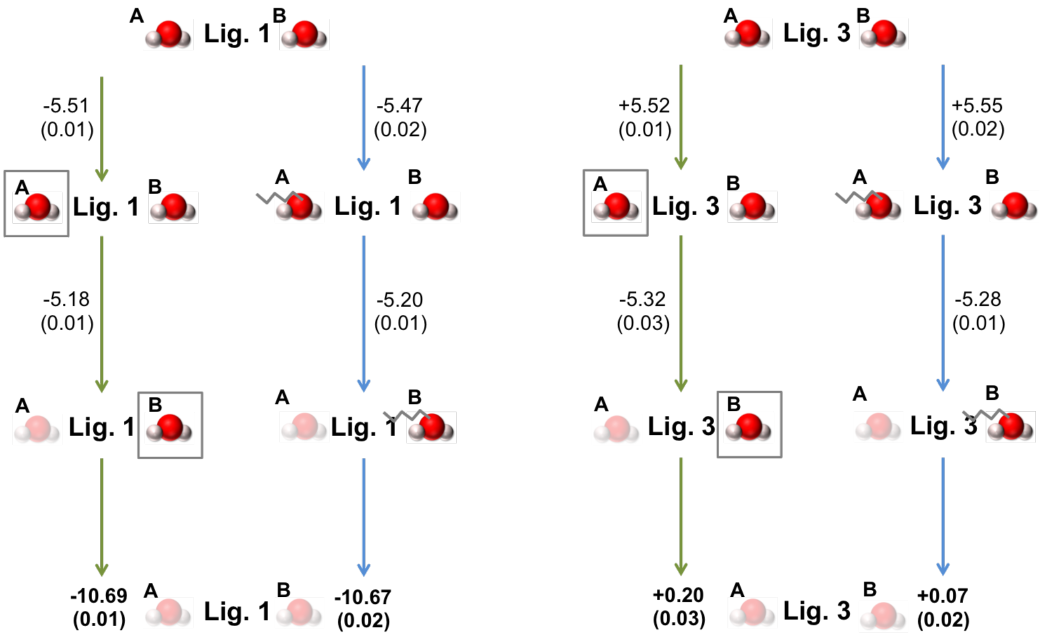


Figure 2.11: Thermodynamic pathway of the two waters considered for each ligand with SD. A box and a green arrow indicates a GCMC simulation, and a spring or a blue arrow indicates a restrained DD simulation. Energies are shown in kcal·mol^{-1}. Errors are standard deviations from four repeats.

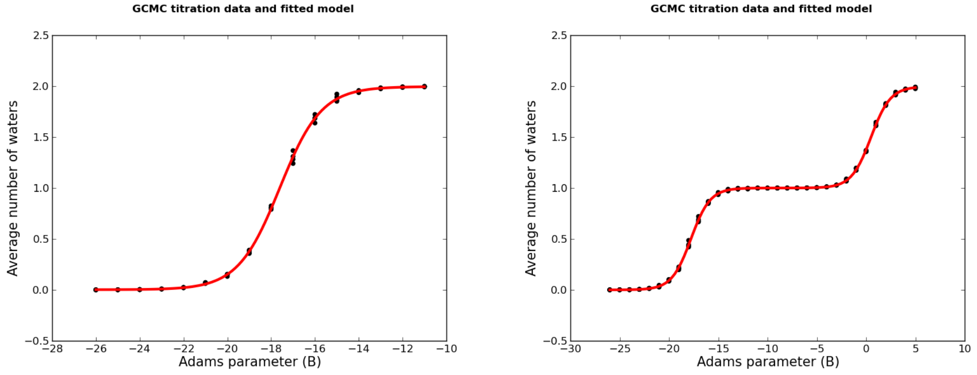
introduced volume term is needed, and the new GCI equation is the correct form.

2.3.6 Multiple-water systems

Tests of the GCI equation for a two, single water have shown that the new form is correct so far as the inclusion of the volume term. However, as the multiplicity term is zero in the old GCI Equation for the case of an occupancy change of one water, it cannot indicate whether or not the multiplicity term is correct. Simulations involving multiple waters are needed to clarify if the exclusion of the multiplicity in the GCI equation is correct in the general form. Calculations were performed for SD and BPTI.

Scytalone Dehydratase

In SD, the same two water molecules, A and B, with both ligands, **1** and **3**, were considered as above, Section 2.3.5, however one larger GCMC region was used to cover both hydration sites, shown in blue boxes (Figure 2.8) to calculate both of their binding free energies simultaneously. As before, the system sampling was limited only to these two waters. Five volumes of GCMC region were tested, increasing incrementally by 1 Å in length.



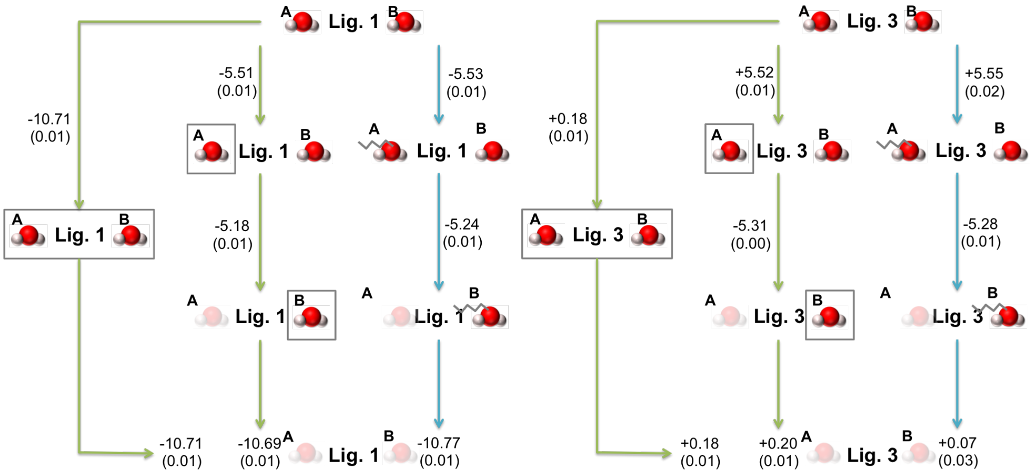
(a) ligand 1

(b) ligand 3

Figure 2.12: GCMC titration two-water networks in SD, with ligands **1** and **3**. Fitting is calculated with four repeats, with calculations performed with a box length of 8 Å.

GCMC has been used to calculate the binding affinity of two active site water molecules for SD with ligands 1 and 3. Figure 2.12 shows the GCI titration for the two water system (waters A and B) when calculated in a single simulation, using the smallest of the GCMC regions that covers both sites (blue - Figure 2.8) for SD bound to both ligands **1** and **2**. With ligand **1**, the waters couple into the system simultaneously, and therefore it is not possible to decompose the energy of the two-water network to the two individual waters; however, the binding affinities can be assumed to be similar as they couple into the system at the same B values. With ligand **3**, the binding free energies of the two waters in the system are different, and therefore enter the system at different B

values. As the titration for ligand **3** occurs over two steps, the binding free energy of the two water network can be decomposed to the two molecules. As before with the single-water calculations, a box volume effect is observed for the calculations over two-water network. The volume correction term in the new GCI equation is able to remove the dependency, as with the one-water systems, Figure A.7. This further supports that the volume correction term is required, and is consistent with varying ΔN.



(a) ligand 1

(b) ligand 3

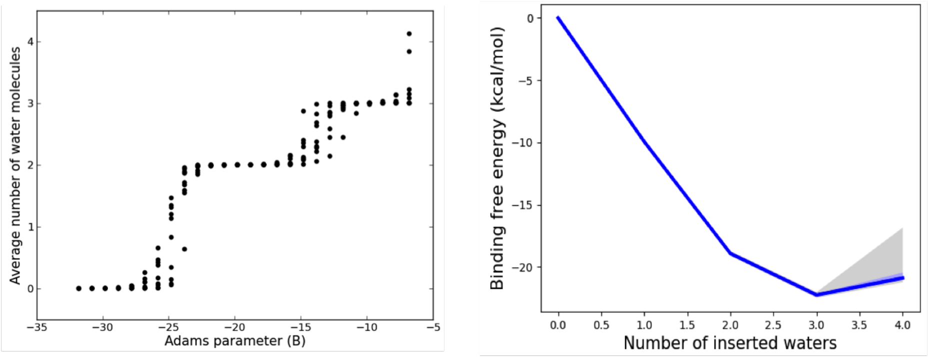
Figure 2.13: Full thermodynamic cycle of GCMC and DD results for each SD ligand. The two legs on the right are the same results as in Figure 2.10. The left hand leg shows the result when GCMC is performed using a large box over the two waters simultaneously, calculated using the new GCI equation, Equation 2.7. Errors are standard deviations from four repeats.

Decomposing the energetic contributions of each water molecule supports that the multiplicity term should be excluded from the GCI Equation. The binding free energy of a two-water network should be equal to the sum of the free energies of each independent water. This means that the free energy of the networks can be compared to the energies calculated in Section 2.3.5. Figure 2.13 show the binding free energy of the two-water network, as calculated with the new GCI equation. The results are consistent with both the single water GCMC and the DD results. As this is a network of two waters, the previous GCI formulation would have a 0.4 kcal·mol^{-1}contribution from the multiplicity term. As the DD and GCMC results all agree to within 0.1 kcal·mol^{-1}this supports the exclusion of the multiplicity term from the GCI equation as being theoretically correct. This is a result of the multiplicity in the system being equal to the multiplicity in the ideal gas reservoir, and therefore cancelling. The results of both the single and multiple water SD water networks indicate that the updated form of the GCI equation is now firmly consistent with the gold standard method - DD using a simplified model system. This analysis has been made possible both through the implementation of RE and the associated gains in reproducibility, and by the simplification of the test system by removing many degrees of freedom from the simulations.

BPTI

As the two water molecules A and B are separated and the system has been simplified by removing sampling of protein, ligand and non-GCMC water molecules, the more complex system of BPTI, where there is a hydrogen bonded network of water molecules has also been considered. The titration of the BPTI pocket is shown in Figure 2.14(a), where two water molecules (B and C) couple into the region as a dimer, followed by a more weakly bound water molecule (A), where the labelling of water molecules is shown in Figure 2.3. The binding free energy of the water networks were calculated, Figure 2.14(b), which shows that the optimal water occupancy of the pocket is three. Clustering of the GCMC water positions was performed and the positions are shown in the BPTI cavity in Figure 2.3. These three water molecules are all within 0.8 Å of their locations in the crystal structure (PDB: 5PTI).

The binding free energy of the three water BPTI network is consistent when calculated by GCMC and DD. The GCMC titration, Figure 2.14(a) shows that the first two waters enter simultaneously as a dimer (waters A and B), followed by the third water, (C). For rigorous DD simulations, the water



(a) titration

(b) binding free energy

Figure 2.14: The titration curve and binding free energy of water networks in the BPTI system, using RE. A minimum binding free energy is found with a water occupancy of 3. The grey region indicates the 95% confidence interval of the standard error. The titration shows that the first two water molecules enter the system as a dimer, followed by a third water molecule at a higher B value.

molecules should be decoupled in turn, in the order of weakest bound to most tightly bound, however any order should result in the same overall network energy. Performing double decoupling in order of weakest-to-tightest bound water molecule is a common protocol, as if a weakly bound water is remaining in the system once a more favourably bound water molecule has been decoupled, the weakly bound water will be likely to adopt the more favourable site. As waters A and B have similar binding free energies, the calculations have been performed twice, once for each order of DD (A then B and B then A). The free energies of each DD calculation and the GCMC results are shown in Figure 2.15. As the GCMC titration finds the binding of the A-B dimer in a single step, it is not possible to decompose the binding free energy to each individual water, and therefore this is not shown. The DD results of the dimer find a different binding free energy of the water molecule depending if it is calculated in the presence or absence of the other dimer member, however the binding free energy of the pair of waters is consistent. The GCMC results for the A-B dimer is consistent within error to both other sets of calculations, which supports the form of the new GCI equation, with the volume correction and the multiplicity accounted for correctly.

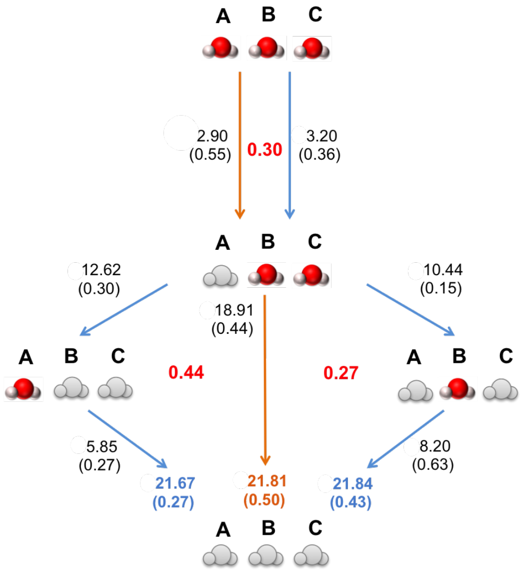


Figure 2.15: Thermodynamic cycle of the water network in the BPTI cavity. Energies shown are the free energy of removing the indicated water from the system, in units of kcal·mol^{-1}. Results are shown for DD (blue) and GCMC (orange). The two routes of DD indicate the two orders in which the water molecules in the dimer are decoupled. The GCI results have been calculated using the new GCI equation. Red numbers indicate thermodynamic cycle closures. Errors are standard deviations from four repeats.

Water A is the weakest bound of the three water molecules in the network, found to insert at the highest B values in the GCMC simulations. Analysis with the updated GCI equation found the water to bind with a free energy of -2.90 kcal·mol^{-1}. This is within error of the DD result; -3.20 kcal·mol^{-1}. The binding free energy for the dimers, a and b are also consistent between the updated GCI equation and the two DD pathways, where the free energies are -18.91, -18.47 and -18.64 kcal·mol^{-1}respectively. The thermodynamic closure of all three cycles shown in Figure 2.15. The closure is smaller than the standard deviations of the simulations. This rigorously illustrates the consistency of the updated GCI equation to DD results, and supports the need for the ideal gas phase Helmholtz free energy contribution to be considered.

2.4 Conclusion

RE is able to significantly improve the errors associated with calculating binding free energies of water molecules using the GCI Equation. The improvements arise due to the increase in monotonicity of GCMC titration plots, to which fitting is performed to calculate the binding free energies. The smaller the error of the logistic fit, the smaller the error in the calculated binding free energy is. The free energies calculated using RE-GCMC are comparable to free energies calculated using DD.

The reduction of errors with the introduction of RE-GCMC reveals a discrepancy of results when compared to the gold-standard DD methods. The discrepancy is shown clearly in Figure 2.6, where GCMC with no RE has larger errors that overlap with DD. When RE is used with GCMC, the median result stays the same and the error reduces and the overlap with DD is lost. The discrepancy prompted a re-evaluation of the theory of GCMC, and it was discovered that the Helmholtz free energy of the ideal gas had been erroneously neglected. Introduction of the ideal gas Helmholtz free energy term to the GCI Equation resulted in two changes; the addition of a volume correction, and the removal of the multiplicity terms. These changes have been derived mathematically, and verified with the use of model systems in Section 2.3.3, and supported by empirical testing for two systems in Sections 2.3.5 and 2.3.6.

A consequence of these theoretical improvements is the derivation of B_{eq}, the B value at which the system is in dynamic equilibrium with bulk water. Previously, a range of B values was needed to generate a titration curve from which equilibrium could be established, by finding the minimum in the Gibbs free energy for the system. B_{eq} removes the need for this, and the equilibrium can be directly simulated.

As the binding free energies of water molecules calculated using the GCMC methodology have been shown to be consistent with other methods, the following chapter will look at the precision in the placement of water molecules in the active site. For the binding free energies calculated to be reliable, the location of the water molecule must be realistic. GCMC will be used to locate hydration sites in a dataset of 105 protein-ligand complexes, where the dataset has been generated using structures that are both high-quality and of pharmaceutically relevant molecules. Discussion will focus on the effect that simulation protocol and analysis methodology can have on the apparent success rate, and the consequence that this can have when comparing between different published results.

Chapter 3

Active site water placement in targets of pharmaceutical interest

3.1 Introduction

This chapter has been completed with significant contribution from MLS. MLS and HBM triaged the dataset and optimised the simulation protocol. MLS set up 25 structures of the dataset, w. All simulations, and analysis herein was performed by HBM.

The experimental limitations of determining active site water locations have been discussed in Section 1.5, and various computational methods that try to determine the locations are discussed in Section 1.3. Computational methods for locating active site water molecules can be useful for drug design, where it may not be efficient to generate a crystallographic structure of every complex of interest. While the previous chapter validated the GCMC determined binding free energies against other computational results, this chapter will validate the locations of GCMC water molecules against a dataset of experimental structures. A dataset of 105 structures has been curated, against which the success of GCMC will be tested.

Various datasets of protein structures exist, by way of benchmarking different methods for different applications. The iridium dataset is curated by OpenEye scientific, and classifies structures based on how trustworthy the experimental data are,[142] with particular focus on the crystallographic assignment of the ligand for use in docking tests. The Astex diverse set is another generated dataset, consisting of 85 structures, which has been collected due to their interest in drug design.[143] Currently, the ProtoMS implementation of GCMC has been applied to various targets; SD, BPTI (two different pockets), MUP-I, Chk-1, HIV1-protease, ribonuclease A, GluR2, trypsin, and glutathione S-T.[54] However, the method should be validated on a larger set of systems, from which statistics of the success rates can be reliably extracted. Here we have curated a dataset of 105 protein-ligand complexes, which are of good experimental quality, pharmaceutically relevant ligands, and contain water molecules. To our best knowledge, this is the largest validation set of a simulation-based water placement methodology. Full details of the curation performed are outlined in Section 3.2.1.

Different methods of crystallographic water placement have been tested and validated on a range of crystal structures, using differing criteria to assess their success. Proposed water molecules are generally assumed correct if they are located within a given distance cutoff to a crystallographic water molecule. The success rate of a method will vary as the considered cutoff is changed; the larger the cutoff, the higher the success rate will be. To validate the performance of GCMC in water placement, various cutoff distances will be considered to illustrate the accuracy of the method. The fact of using a cutoff means that randomly placing water molecules within a region, with no intelligent consideration of chemistry, will also reproduce some of the crystallographic water molecules correctly by virtue of chance. The larger the cutoff used the more likely randomly placed water molecules will be to successfully find a crystallographic water position. The success rates of GCMC water placements for various cutoffs will be presented alongside the success rates of random water placements, to act as an illustrative baseline.

While efforts have been made to select high-quality crystallographic structures, there is still a degree of uncertainty in the data. Water molecules that are diffuse, or weakly bound may not be possible to resolve in any quality of structure. Four metrics; the Z_{obs}, EDIA, B-factor and B$_{norm}$ scores will be used to inspect the underlying electron density of assigned water molecules. Even if the electron density is clear, there can also be a bias from the crystallographer, due to decisions that they make during the refinement process. Another factor to consider is the experimental crystallisation conditions, which may introduce ions or small molecules that are not present in the biological conditions of the protein.[144] In addition, the xray diffraction conditions can also be non-biological, and the majority of our dataset has been resolved at <100 K. Despite this, crystallographic data is the best comparison available for many structures, and therefore will be used to benchmark GCMC, while taking due consideration for the degree of the experimental accuracy.

3.2 Methodology

3.2.1 Dataset generation

The dataset was generated by collating FDA approved drugs that have a protein-bound structure within the PDB. The drug molecules were filtered using the criteria outlined in Table 3.1, resulting in 1554 PDB structures, covering 279 FDA approved drugs.

Table 3.1: Ligand requirements used for FDA dataset generation.

Carbon count	>5
Phosphorous count	=0
Molecular weight	100 → 750
Rotatable bonds	< 9
Ring size	< 9

Of these 1554 structures, the data were further filtered to include only structures released since 2000, with a resolution better than 2.5 Å, of homo sapien, viral or bacterial origin. The rejection of structures older than 2000 is due to more recent improvement in the software used in assignments of crystal structures. Crystal structures were excluded if they contained no water molecules. Complexes were excluded if they were covalent binders, contained co-binding molecules, such as organic solvent in close proximity to the ligand, or metal ions not covered within ProtoMS software. Structures with any missing resides in the active site, or more than 3 missing consecutive residues distal to the active site were removed. No single drug molecule or protein was allowed within the final data set more than 5 times each so as to ensure the dataset is diverse. The resulting data set has 105 complexes of 80 unique proteins and 72 unique drugs, with no repeated protein-ligand pairs. Details of the targets, ligands, pdb codes, publication years and resolutions of the dataset are in Section A.2.

As it is only the location of water molecules that are to be considered here, it is possible to only simulate at the equilibrium B value only, defined previously in Equation 2.13. Simulating only at B_{eq} avoids the need to perform the more computationally expensive full titration. As Chapter 2 demonstrated the improvements in sampling when RE between B states is included in the protocol, three additional B values proximal to B_{eq} will also be simulated. Water locations and other analysis will only be performed on the B_{eq} replica. For uniformity across all systems in the dataset, a cubic GCMC region that is a minimum distance of 4 Å to all ligand heavy atoms.

3.2.2 System set-up

FDA dataset

All 105 proteins used in the FDA dataset were set up using the following protocol. The structures used are shown in Table A.1. Where a protein is replicated in the dataset, the setup was performed all structures independently. The protonation and tautomer states of the proteins were determined using Maestro.[145] A scoop of 30 Å was used, with full amino acid sampling within the inner 15 Å and the rest of the protein held rigid, for the sampling simulations. For the fixed simulations, the whole protein is held rigid.

Proteins

For all proteins simulated, the amber14SB force-field has been used.[6]

Ligands

For all ligands, the gaff14 forcefield has been used with AM1-BCC charges. All 105 ligands used in the FDA dataset were set up and protonated, and tautomer state chosen, using maestro.[145] The structures used are shown in Table A.1. Where a ligand is replicated in the dataset, the setup was performed on all ligands independently.

Solvation

For all water simulated, the TIP4P force-field has been used.[141] Protein-ligand complexes were solvated using a half-harmonically restrained sphere of radius of 30 Å, with all crystallographic water molecules were removed. This includes solvating any sterically available active site regions.

3.2.3 Simulation protocol

For every protein-ligand complex in the FDA dataset, two simulations were performed; one sampling, and one fixed. GCMC has been performed using four B values; B_{eq}-1, B_{eq}-0.5, B_{eq}, B_{eq}+0.5. A B spacing of 0.5 has been used to ensure good exchange between replicas for the full dataset, which is demonstrated in Section A.1. A GCMC box of 4 Å padding around ligand heavy atoms was used.

Sampling simulations consisted of 10 M GCMC only equilibration, 10 M full sampling equilibration and 40 M full sampling production steps. Full sampling consists of half of the simulation moves sampling the system, with the other half performing GCMC moves. System sampling is shared between bulk water molecules, protein residues and the ligand at a ratio of 1:5:50. GCMC sampling is split equally between insertion, deletion and GCMC water sampling.

Fixed simulations consist of 10 M GCMC only equilibration steps, followed by 20 M GCMC only production steps. No protein, ligand or solvent is sampled. The number of GCMC moves attempted will be the same as the sampling simulation, within the limit of stochastic sampling.

EDIA scores were calculated using the proteins plus web server.[146] Zobs scores were calculated using edstats in the CCP4 software suite.[147]

3.3 Results

The success of GCMC in placing crystallographic water molecules will be considered — that is the percentage of crystallographic water sites that are reproduced to within a given distance cutoff. GCMC water locations from the simulation are clustered using hierarchical linkage clustering, where a cluster is defined as requiring all its members to have a maximum average cartesian distance of 3.0 Å, that is that the average distance of all members of a cluster is less than 3.0 Å. MLS has improved the clustering algorithm by applying an arbitrarily large distance to GCMC water locations that appear in the same frame. The large distance condition prevents two water molecules that are observed simultaneously in the same frame being erroneously placed in the same cluster, and consequently sets the maximum cluster occupancy at 100%, which was not true for previous applications of the algorithm.

Each water site will have an occupancy, which is the number of water molecules from the simulation that are put in the cluster. While a full titration is needed to calculate the binding affinity of a network of water molecules, some qualitative assumptions can be made from simulations performed at B_{eq}. If a water site is occupied for 50% of the simulation, then the water molecule is equally stable in this site and in bulk water, which means that its binding free energy is 0.0 kcal·mol^{-1}as they are equally likely to occupy both bulk water and the hydration site. Water molecules with higher occupancy can be generally be assumed to be more tightly bound than lower occupancy water molecules.

3.3.1 Success rates

One protein-ligand complex from the FDA dataset, zanamivir bound to neuraminidase (PDB: 3B7E) will be used to introduce the methods of analysis, and the issues that can arise with the analysis. The success rates will then be applied to the entire dataset. Figure 3.1 shows the crystal structure of 3B7E, with the GCMC region illustrated with a grey line, and crystallographic water locations

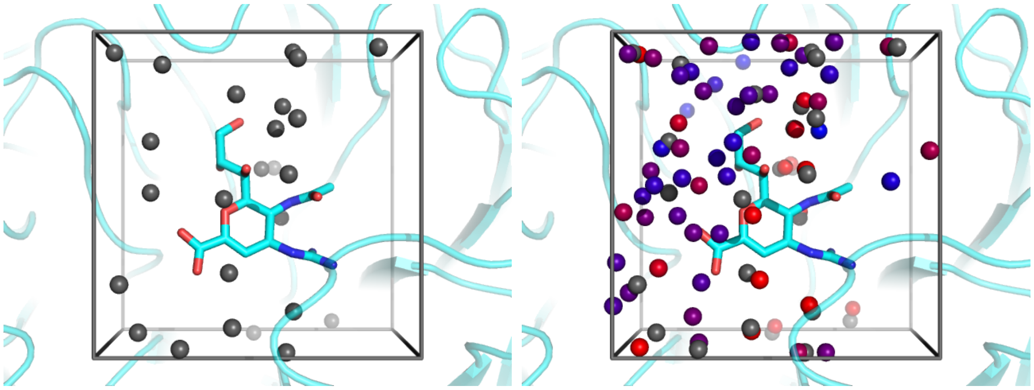


Figure 3.1: Crystallographic structure of zanamivir bound to neuraminidase (PDB: 3B7E), with the GCMC region indicated by a grey box. Crystallographic water locations within the GCMC region are shown by grey spheres. All GCMC cluster centres (right) are shown, coloured blue (low) - red (high) occupancy.

shown as grey spheres. On the right hand side, overlaid onto the crystal structure are the cluster centres determined from GCMC simulation. These cluster centres are coloured according to their occupancy — that is the amount of time they are seen in the simulation, or the number of water molecules from the simulation that are placed into that cluster — with blue indicating low occupancy water molecules, through to red, high occupancy water molecules.

Some of the GCMC water molecules clearly overlap with the crystallographic sites, some crystallographic water molecules are close to GCMC clusters that are low occupancy, and some are a distance from the closest GCMC site. To decide if a crystallographic site is correctly located comes with several considerations; is the closest GCMC site close enough to be considered correct? Is GCMC site occupied enough to be considered to correctly observed? Both of these issues will be discussed in this section, where both the cartesian distance cutoff, and the GCMC occupancy cutoff that are used to measure success will affect the result. Another point to note is that cluster centres can be closer together than typical water-water distances, which occurs where the water density is diffuse, and mul-

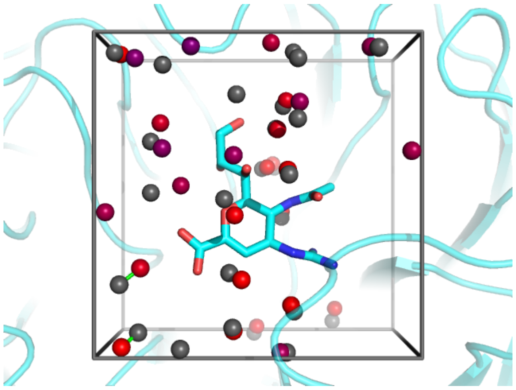


Figure 3.2: GCMC cluster centres with occupancies > 50% with zanamivir bound to neuraminidase (PDB: 3B7E). Green lines indicate distances of 1.0 - 1.5 Å between crystallographic water sites and GCMC cluster centres and all other labels are consistent with Figure 3.1.

tiple, low-occupancy cluster centres are used to fit the density. Figure 3.2 shows the system, but with only GCMC cluster locations that have occupancies > 50%. Here, GCMC cluster centres are not unrealistically close together.

Removing the low occupancy sites, reveals several GCMC sites that are not observed by a > 50% occupied cluster. For some crystallographic sites, the distance to the closest GCMC cluster center is increased when the low occupancy sites are removed. Crystallographic-GCMC distances that are between 1.0 - 1.5 Å are highlighted by a green dash. Whether these sites should be considered as correctly identified is a matter of opinion, and the distance cutoff used by different published methods varies. What distance cutoff is considered will change the apparent success of the method. Zanamivir bound to neuraminidase (PDB:3B7E) has been used to introduce the issue of classifying the method as successful for a single structure, but the overall results from the dataset of 105 protein-ligand complexes will now be presented.

GCMC water molecules with occupancy of 50% or greater, have a success rate of 51% and 67% at 1.0 and 1.4 Å respectively. Deciding if a crystallographic water molecule has been correctly identified will depend on both the distance to a GCMC water site, and the occupancy of that site. Figure

3.3 illustrates this, where the accuracy of GCMC has been recorded for various distance cut-offs and GCMC water occupancies.

The accuracy of GCMC is dependent both on the cutoff used, and the occupancy of GCMC water molecules considered. Both of the following trends are to be expected; the longer the cut off, the more crystallographic sites will be correctly predicted, and the higher the GCMC occupancy required, the fewer sites are correctly identified. The higher reported accuracy with longer cutoff can also be observed for the other methods shown.

Other published methods perform competitively, but it is difficult to compare methods with different methodologies that have been applied to different datasets. Figure 3.3 contains published success rates for other water placement methods, compiled by MLS. These data are based on the test set used in the publication, and not our dataset of 105 structures. If methods have quoted their success rates at different cutoffs, then they are all shown. AcquaAlta quoted the success rate with two different datasets, at the same cutoff, which are both shown. Considering only GCMC water molecules with occupancy greater than 50% (red, Figure 3.3), the GCMC method has higher accuracy than several published methods. Many other methods fall between the threshold of any GCMC occupancy and 50%, indicating that GCMC is locating the sites, but only transiently in the simulation. Other published methods have been shown for comparison, but much of this Chapter will focus on the difficulty of comparing different water placement methods by looking at the variability in success rates that can be achieved by minor changes to the protocol.

Randomly placing water molecules within a system will correctly identify some crystallographic water molecules by chance. For a baseline comparison 'random' water locations have been generated. These involve the identification of water sites using the ProtoMS set up tools, whereby water sites are naïvely identified for a starting conformation. A pre-equilibrated water box is overlaid with the complex and water molecules are removed if they overlap with

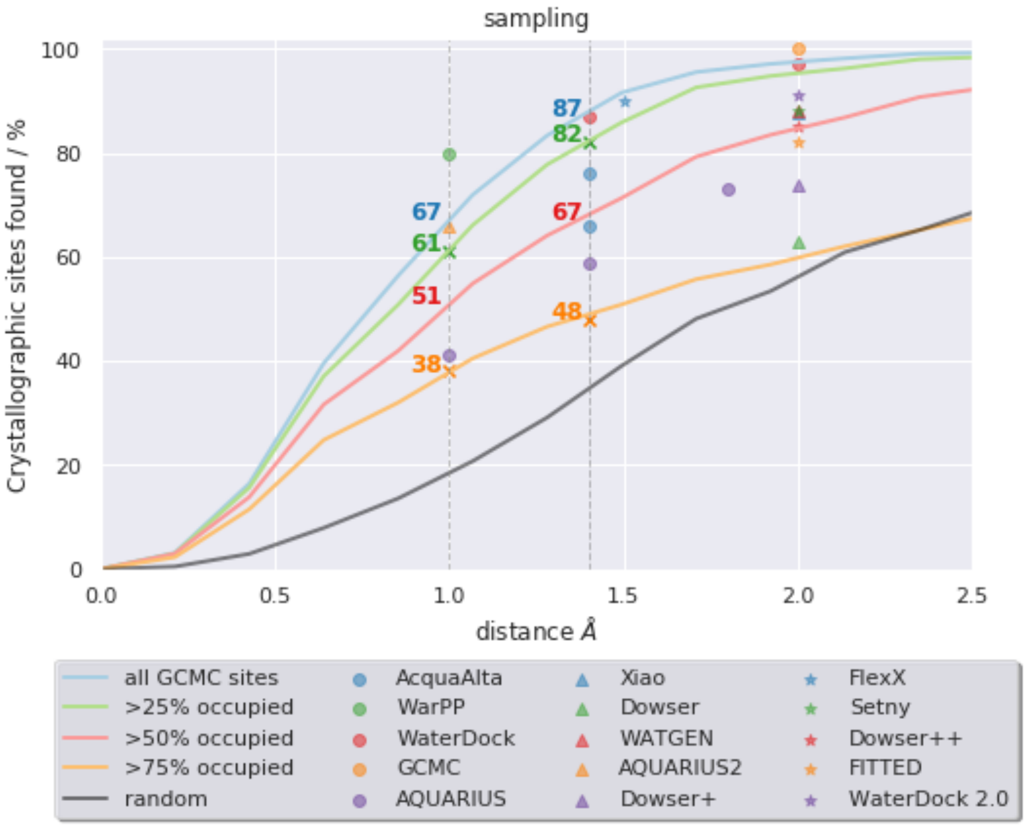


Figure 3.3: Accuracy of GCMC at different cutoffs. Results based on random solvation (black) are shown for a comparable baseline. GCMC results are calculated for 632 active site crystallographic waters in 105 structures. Other published methods are shown by markers; however, all have been calculated using different protocols and test sets. A dashed black line is shown at 1.0 and 1.4 Å, annotated with the GCMC percentage success rate for each occupancy threshold at that distance.

any atoms of the system. A water molecule is considered to be overlapping if the vdW interaction energy of the water's oxygen atom to the nearest atom of the system is greater than 20.0 kcal·mol^{-1}. The random results indicate that it is possible to correctly identify some crystallographic sites effectively by chance. 20% of crystallographic water molecules are identified at a cutoff of 1.0 Å, and this value increases as the cutoff increases. At 63% at a 2.0 Å cutoff, Dowser is only 5% better than the random result of 57%, suggesting that shorter cutoffs should be used to identify if water placement methods are accurate. As all of the computational methods perform better than the random, naïve solvation, this suggests that any of the published water placement methods would be advantageous to use in the setup of protein-ligand simulation. The success rate of randomly placing water molecules is approaching 60% at 2.0 Å, which is very high, and from here on, the commonly used distances of 1.0 Å and 1.4 Å, where the random results have less success, will be considered.

These results suggest that other water placement methods are outperforming GCMC. However there are several major caveats to consider: GCMC is sampling the protein-ligand environment while most other methods do not, and 'correctness' is determined against crystallographic water locations, that will come with their own limitations e.g. trusting that the electron density has been correctly assigned. These two caveats will be explored in the following two sections, looking at rigid receptor results, and analysis of the underlying crystal structure quality.

3.3.2 Rigid receptor results

The methodology of many of the other methods shown use a rigid molecule approximation during water placement, which will increase the quoted accuracy. Of the other computational methods shown above in Figure 3.3, most treat the protein-ligand environment as rigid. Many of the methods are knowledge based, where the locations of water molecules are assigned based on knowledge of crystallographic water molecules in other structures. The method that performs the best at 1.0 Å is the knowledge based method WarPP,[82] with a success rate of 80% for 1500 complexes, all of which are structures of 1.5 Å or better. Of all the methods shown, only GCMC,[54] which uses a previous implementation of ProtoMS following a different protocol, and Setny[148] sample the protein-ligand environment. If the environment is sampled, then it is likely that the system will move away from the crystallographic starting structure (a feature of sampling), and therefore reduce the number of sites that are correctly identified based on cartesian analysis.

GCMC on the dataset has been repeated where the surrounding environment is kept rigid. Sampling of the complex allows the system to relax, and sample alternate conformations. The majority of the structures in the dataset have been crystallised at 70 K, and while GCMC simulations are performed at 300 K, the system is likely to adopt different conformations at the higher temperature. For this dataset, sampling involves MC trial moves of the ligand, bulk water and full sampling of protein residues (both side-chain and backbone) within 15 Å, while residues at a distance of 15 - 30 Å are held rigid. For example, if an amino acid side chain that forms a hydrogen bond with a water rotates, it may 'pull' the water molecule along with it. This would result in the water sitting in a different position, and therefore being assigned incorrect based on a cartesian assessment, but nevertheless correct based on the interactions maintained. For this reason, GCMC simulations have been repeated for the dataset of 105 FDA approved drug-protein complexes, where the complex is held rigid while only GCMC water molecules are sampled. No MC moves are assigned to protein, ligand or bulk water in the simulation. Only insertions, deletions and translations of GCMC waters within the GCMC region will be attempted.

Simulations where the system is kept as rigid have improved success rates. In docking methodologies, whereby a small ligand is docked to a receptor, the components may be held rigid, the ligand may be flexible, or both components may be flexible. As flexibility is introduced into the model, the docking more accurately captures the induced fit motion of protein ligand complementarity.[149] However, as flexibility is introduced, the likelihood of reproducing the native binding mode, if the rigid receptor is correctly oriented to the native state, is reduced.[150] This is known as rigid receptor theory.[151] If GCMC insertions and deletions are considered as the repeated 'docking' of water molecules into the active site, then the same rigid receptor theory should hold for GCMC. If sampling of the protein-ligand complex results in the shift in a chemical group that covers a crystallographic water position, then insertions will no longer be possible to access during the simulation, which in turn would reduce the possible success rate of GCMC. GCMC results are being compared to the single snapshot of a complex that crystallography provides, which means that a rigid complex simulation, that is unable to move away from the single snapshot to which they are compared, is more likely to generate results in agreement with the X-ray water assignment. As the sampling simulations consist of 50% system sampling and 50% GCMC sampling, the rigid simulations consist of half of the number of MC moves, but 100% GCMC sampling. This means that the number of GCMC moves attempted in both sampling and rigid simulations is consistent.

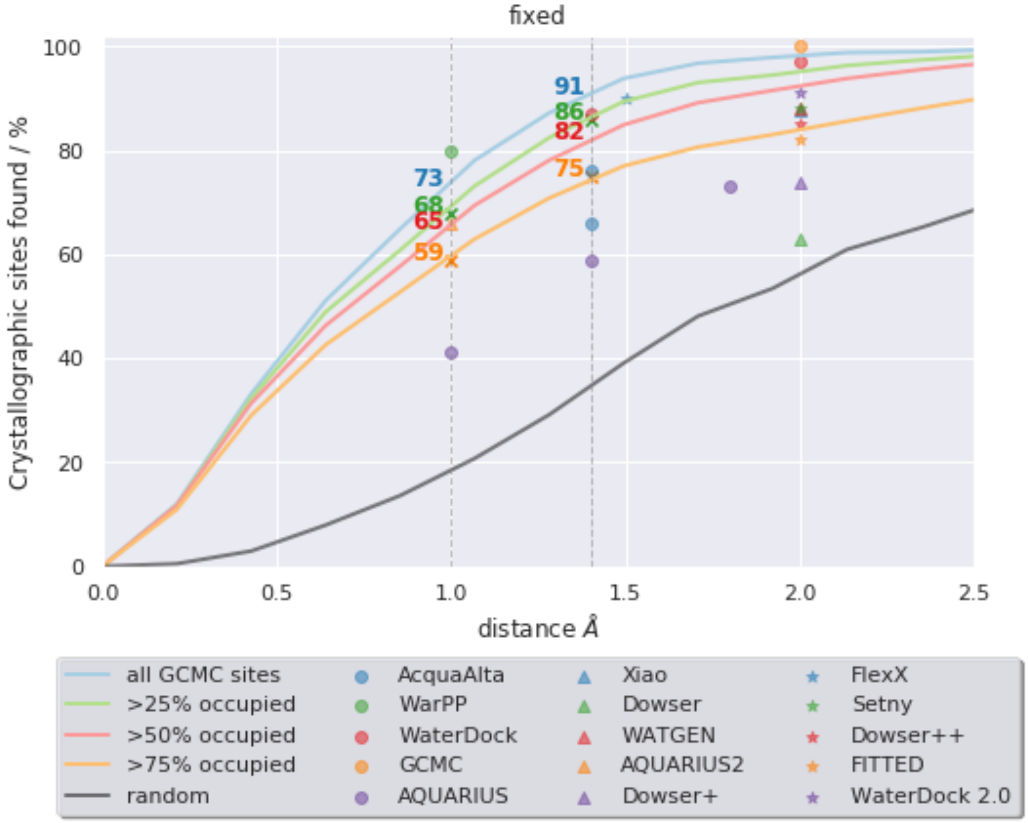


Figure 3.4: Accuracy of GCMC at different cutoffs, from fixed environment simulations. The format is the same as in Figure 3.3

Fixing the surrounding environment in GCMC simulations improves the success rate. Figure 3.4 shows the results at different cutoffs, with different minimum GCMC cluster occupancies for the fixed simulation results. For all occupancies and cutoffs, the success rates are higher than for the sampling results. As the success rates are higher for the fixed simulations, this supports the rigid receptor argument for GCMC simulations. The differences between different occupancy thresholds are reduced in the fixed results compared to those with sampling. At 1 Å distance cutoff, there is a 31% reduction in the success rates if water molecules with <75% occupancy are excluded, while for the rigid results, the success rate is reduced by only 14%. For these results, the fixed receptor data a fairer comparison with many of the other modelling methods, although significant caveats still exist; the results have been determined on different datasets of differing sizes, with different classifications of water molecules. Classifications vary in several ways, with some methods only considering conserved water sites[54] and others requiring multiple hydrogen-bonding contacts.[111] For the fixed receptor results, looking at GCMC sites of any size (blue line) only WarPP,[82] and GCMC[54] (for a small dataset of structures) have a higher success rate. Considering GCMC sites of 50% occupancy or higher, WaterDock[152] and FlexX[153] score higher, although the results are fairly similar. WaterDock has been tested on a dataset of 37 structures, covering an estimated 12 targets, which is a smaller, less diverse dataset than used here. FlexX has been tested on a large dataset of 200 structures of 120 targets, but only water molecules that form a hydrogen bond were considered. Based on these differences in dataset and analysis, GCMC is performing similarly to the best other methods.

While fixing the environment increases the success rates, sampling the system affords other benefits. Keeping the environment fixed during GCMC simulations has benefits; the simulation is faster, as time is saved by reducing the number of MC moves. For a faster simulation, the success rates improves; for 50% occupancy at 1.0 Å, the success rate increases by 15% (50% with sampling, 65% with fixed). This fixing of the system allows GCMC to be more fairly compared to other available methods, where the majority do not alter the environment from the initial crystallographic starting structure. While the improved success rate seems beneficial, it is not the only metric of success. Sampling the protein and ligand conformations provides information on multiple conformations of the systems. We are able to observe different ligand binding sites (this will be discussed in Section 3.3.6) which can be extremely useful knowledge in drug design. In addition, the structures used herein have resolved crystal structures, whereas in a real drug design project, the exact protein-ligand complex of interest may not be available, and methods such as homology modelling or docking may be required. This would likely increase the need for sampling of the protein-ligand environment during GCMC simulation. The fundamental goal is not just to predict water molecule locations in complexes, but to do so in a manner that aids the design of high affinity molecules. For example, seeing multiple binding modes or multiple conformations of a key protein residue, or multiple networks of water molecules that correspond to these different conformations is often more important than the absolute cartesian agreement with crystallographic water locations. This is particularly true when the relevance of a crystal structure to the structure of biological relevance is considered.

3.3.3 Types of water molecule

Water molecules have been classified based on their crystallographic contacts. The 632 active site crystallographic water molecules have been classified — as bridging, ligand, protein or solvent — based on 2.4-3.4 Å cutoffs to protein or ligand polar heavy atoms (nitrogen, oxygen, sulphur). Bridging water molecules are within hydrogen bonding distance to polar atoms in both the protein and the ligand. Ligand and protein water molecules are within hydrogen bonding distance to either the ligand or the protein, while solvent water molecules are considered bulk-like, as they are not within hydrogen bonding distance to any polar atoms of the complex. This classification is performed on the crystallographic water molecules, in reference to the crystallographic location of water molecules. As the classification is performed on the crystal structure, no account is taken of the flexibility of the system, and the possibility that the classification of these water molecules may change through the simulation. Nittinger et al. only consider water molecules with two or more possible hydrogen bonds to protein or ligand atoms. The Nittinger et al. classification of water molecules has been reproduced as closely as possible. As hydrogen atoms are not assigned in the clustering of water molecules, the Nittinger et al. classification performed here, checks for two H-bonding contacts, as defined before, to either protein or ligand. The requirement for two H-bonding contacts changes the proportions of types of waters, with significantly more water molecules classified as solvent when the requirement of two contacts is used. These are shown in Figure 3.5.

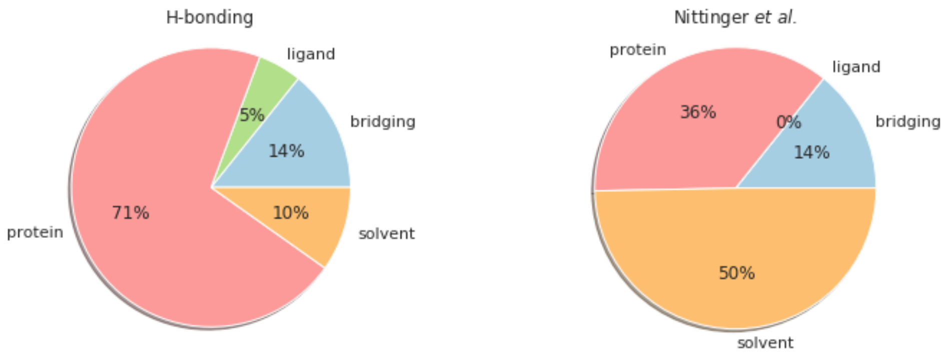


Figure 3.5: Classification of 632 crystallographic water molecules included in the dataset. A contact is defined as a 2.4-3.4 Å distance to a polar atom (nitrogen, oxygen, sulphur). H-bonding classifies water molecules based on a single contact, while Nittinger et al. requires a water site to have two H-bonding contacts, to be classified as bridging, protein or ligand.

Water molecules that directly interact with the ligand or are bridging will be the most important to predict correctly for drug design. Different types of water molecules will be of different importance to drug design. Both bridging and ligand bound water molecules will be in the first solvation shell of the ligand, which accounts for 19% of the crystallographic water molecules (14% by Nittinger classification). These water molecules are the most likely to be perturbed by incremental changes to the ligand and also the primary candidates for displacement during drug design. The correct identification of these water molecules are arguably the most important to be correctly predicted. 10% of water molecules have no contacts to either protein or ligand, and are therefore considered as bulk. The percentage of bulk water molecules increases significantly when the requirement for two hydrogen bonding contacts is used. These water molecules may either be in the second solvation shell or have a crystallographic packing contact which have not been considered herein, or may have one hydrogen bonding contact if defined by the Nittinger classification. As solvent water molecules by the H-bonding classification do not have direct contact with the protein-ligand complex, the water molecule is likely to be more mobile and more disordered, and more difficult to correctly predict. With the Nittinger et al. classification, no water molecules are ligand-bound, as any ligand-bound water molecules also have a protein contact, and are therefore classified as bridging. If a ligand-bound water molecule has two simultaneous contacts with the ligand, it also forms a protein contact, and therefore will be classified as protein bound.

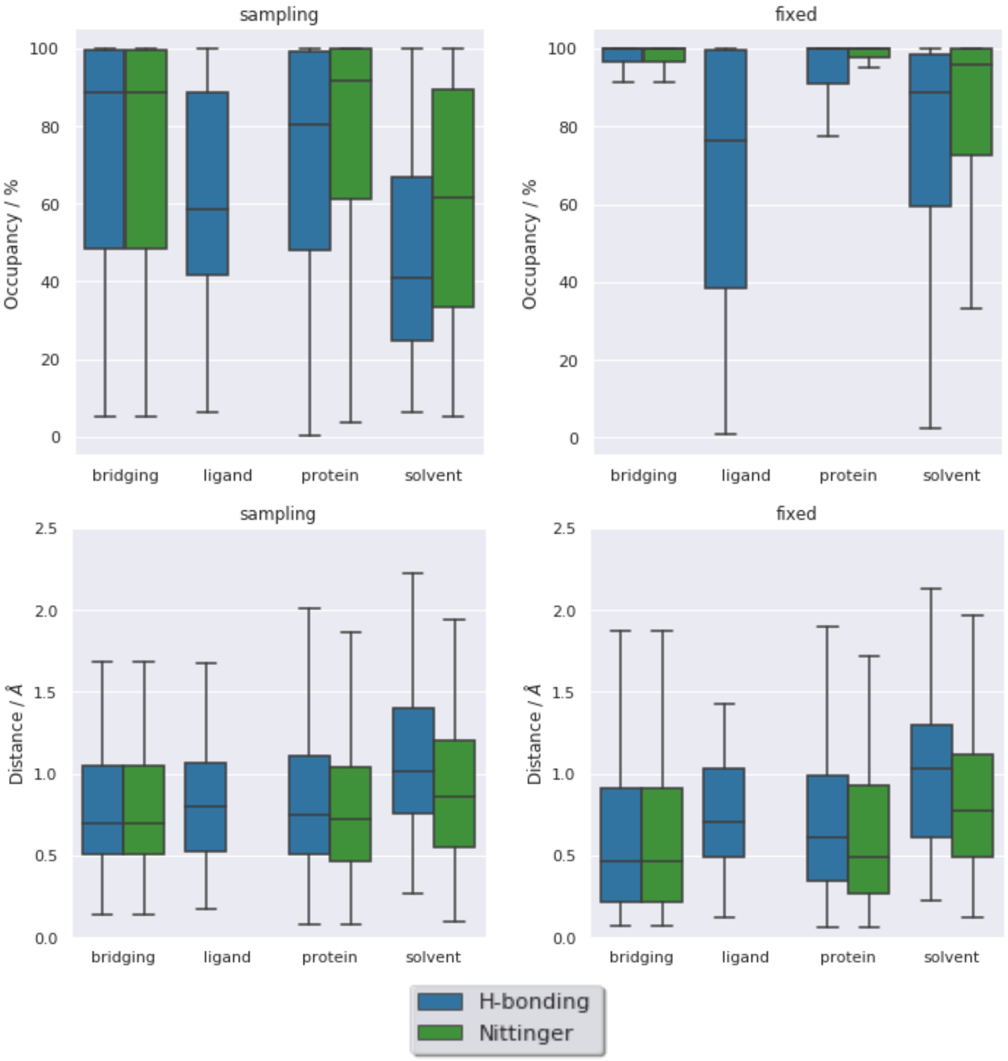


Figure 3.6: Boxplots for both occupancies of GCMC water molecules, and distances from crystallographic sites for each type of water molecule, following both methods of classification (H-bonding and Nittinger) for both sampling and fixed simulations. Median and interquartile ranges are shown, with whiskers indicating the rest of the distribution, excluding outliers. No water molecules are classified as ligand-bound when using the Nittinger et al. definition.

Different types of GCMC water molecules have differing occupan-

cies and distances from crystallographic sites. Shown in Figure 3.6, for sampling simulations, the median water occupancies are higher for bridging and protein water molecules than for ligand bound and solvent water molecules, with broad distributions of occupancies observed for all water molecule types. For the fixed simulations, the distribution in occupancies for ligand and solvent water molecules are much broader than the distribution for bridging or protein bound water molecules. This suggests that these ligand and solvent water molecules are more disordered in the fixed environment simulations, while the bridging and protein-bound water molecules are more localised with higher occupancy throughout the simulation. A water molecule with an occupancy of 50% would have a binding free energy of 0.0 kcal·mol^{-1}. The occupancy results indicate how many of each type of water molecule will be affected when the dataset are filtered for different occupancy cutoffs, Figure 3.3. The whiskers for distributions of distances of sampling and fixed results are fairly similar for most types of water molecule, but the median distance result is lower for the fixed simulations for bridging, ligand-bound and protein-bound water molecules, and equivalent for solvent type water molecules. The distribution of distances for bridging water molecules appears the most significantly lowered when comparing fixed results to sampling. The lower median distance explains the higher success rates achieved for fixed simulations, relative to sampling simulations, shown in Figures 3.3 and 3.4. The majority of ligand-bound water molecules in fixed simulations are within a 1.4 Å distance cutoff, despite having a broader distributions. For both sets of simulations, solvent water molecules have lower occupancies, and larger distances to crystallographic water molecules. The solvent water molecules are those for which GCMC performs the worst, which is unsurprising, as the lack of local structure will make these water molecules the most difficult to resolve in the crystal structure and the least likely to be correctly identified during simulation without a directional hydrogen bonding group from protein or ligand. Solvent-type water molecules are the least likely to be used in drug design as they are diffuse and therefore not appropriate to target for displacement. For this reason, the success rates have been recalculated for each type of water molecule, and also for all water molecules excluding solvent water molecules, Figure 3.7.

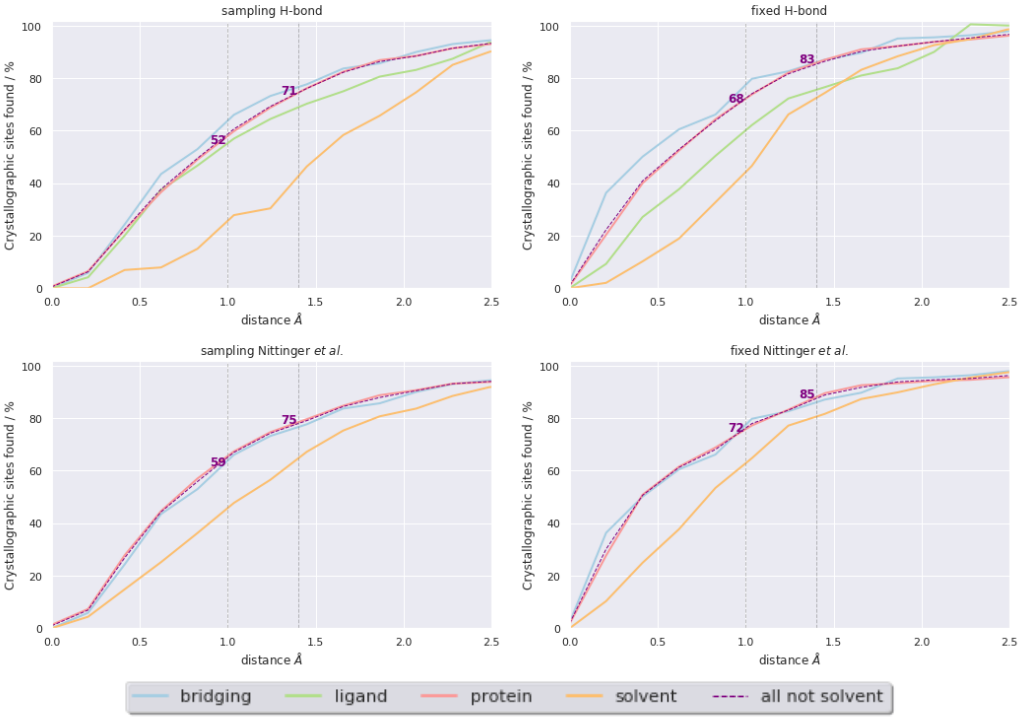


Figure 3.7: Percentage of crystallographic water molecules correctly identified by GCMC (with an occupancy cut off of 50%) at different cutoffs, broken down by classification. Results are calculated for 632 active site, crystallographic waters in 105 structures. A dashed black line is shown at 1.0 and 1.4 Å, annotated with the GCMC percentage success rate for each classification at that distance. A dashed purple line shows the success rate when solvent water molecules are excluded from the analysis, with the success rate quoted at 1.0 and 1.4 Å.

The success rates of GCMC improves if solvent water molecules are excluded from the calculation. The success rates as shown in Figures 3.3 and 3.4 have been broken down based on the classification of the water molecules, shown in Figure 3.7. In agreement with Figure 3.6 the success rates are similar for protein, ligand and bridging water molecules, while GCMC performs worse for bulk solvent water molecules. If solvent water molecules are excluded from the accuracy calculation, the success rate for GCMC increases from 50% to 52 or 59% - fixed system increases from 65% to 68 or 72% - for a 1 Å cutoff, both for H-bonding and Nittinger classification, respectively. WarPP is a knowledge based method, with its success rate of 80% quoted for water molecules that form two hydrogen bonds. The GCMC fixed simulation using the Nittinger et al. classification — as comparable to their classification as possible — gives a result of 72% which is close to theirs. The difference in success rates could exist for a variety of reasons. One explanation is that different datasets have been used, or that their knowledge-based method does not require parameterisation through use of force-fields, which may be a source of error in the GCMC method. It is possible that our success rate could improve with more optimisation on the ligand hydrogen atom locations, as the ligand-bound water molecules lower the average slightly for fixed simulation. The 9% increase in success rate for sampling simulations using the Nittinger et al. classification to exclude solvent-type water molecules indicates how much the success rate can vary with a slight change in protocol. The fluctuation in success rate will be discussed further when crystal-structure quality is also taken into consideration in Section 3.3.4. The variation of success rate shown just for the GCMC simulations, based on the cutoff, occupancy considered and classification of water molecules included in the dataset, illustrates how unreliable it is to compare directly success rates quoted in the literature for different methods, using different datasets, with different protocols and performed by different researchers.

Despite having a lower overall success rate, sampling simulations correctly locate 10% of the dataset that is missed by fixed sampling. Fixed sampling performs relatively worse for ligand bound water molecules. Figure 3.8 shows the distribution between sampling simulation and fixed simulation distance for all crystallographic water molecules. The water molecules have been coloured based on their classification in the crystal structure. Figure 3.8 b) contains the same data as a), but focussed on water molecules that are found to within 2.0 Å of the X-ray location by both methods, and has been divided into quadrants, to distinguish between water molecules that are found to within a 1.0 Å cutoff, with the distributions of types of water molecules across these quadrants shown in c). 40% of water molecules are correctly located by both sampling and fixed backbone methods, and 25% of water molecules are missed by both methods. Sampling simulation are able to predict 10% of water molecules that are

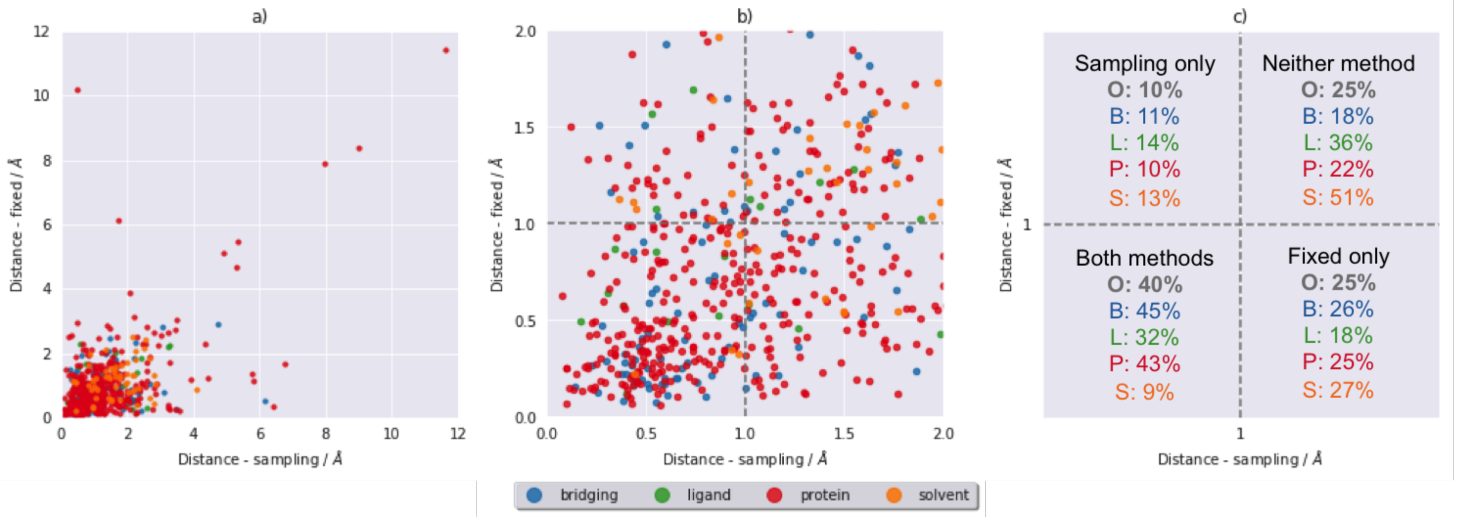


Figure 3.8: Distance of GCMC water molecules (>50% occupancy) to crystallographic water, for sampling simulations against fixed simulations. Each datapoint corresponds to a crystallographic water molecule, with the distance shown indicating the distance to the closest GCMC cluster center of occupancy greater than 50%. Crystallographic water molecules are sorted by contact type. a) shows all 632 crystallographic water molecules. b) shows the same data as figure a), excluding water molecules that were not located to within 2.0 Å for both sampling and fixed. Grey dashed lines are shown at a distance of 1.0 Å for both sampling and fixed, forming quadrants. The lower left quadrant indicates water molecules correctly identified by GCMC in both sampling and fixed simulations and the upper left quadrant shows crystallographic water molecules missed by both simulations. The upper right quadrant indicates water molecules that are identified only by sampling, and the lower right quadrant indicates those found only by fixed simulation. c) Shows the proportion of water molecules within each quadrant for the entire dataset overall (O), and for each classification of water molecule.

missed by fixed simulations, with a marginally larger proportion of these being ligand bound water molecules. 25% of water molecules are found only in the fixed simulations, where a slightly lower proportion of these are ligand bound water molecules are found in these simulations, however these distances may not be significant. Fixed simulations likely perform better as the environment is unable to move away from the crystallographic starting point. The proportionally larger likelihood of sampling simulations to predict ligand bound water molecules than fixed simulations may possibly be due to the opportunity for hydrogen atoms to sample, and achieve more optimal hydrogen bonding contacts with these water molecules. This is simply one explanation, and it is unclear why the lack of sampling of ligand hydrogens would be more detrimental to water placement than the lack of sampling of protein hydrogens involved in hydrogen bonding. Potentially, a method where protein backbone and ligand scaffold atoms are rigid, with side chains and functional groups and hydrogen atoms able to sample could improve the success rate relative to fixed simulations.

3.3.4 Quality of crystallographic data

Quality filters used to generate the dataset use metrics that describe the overall quality of the crystal structure, rather than atomistic metrics. All of the structures in the dataset have been published more recently than the year 2000, with a resolution of 2.5 Å or better, which attempts to ensure that the structures are good quality. These however, are an assessment of the overall quality of the structures, but do not give an assessment of the quality of the electron density for a given atom. Discussion of errors that can arise in crystallography is given in Section 1.5. This section will look at different measures to assess the quality of local water electron density, and if this linked to the likelihood GCMC to correctly locate that water. This section shows results taken from the fixed GCMC simulations, the same trends are observed in the sampling GCMC results, but have been excluded for berevity.

Different metrics are available to assess the atomistic disorder or

atomistic electron density. The B-factor is a measure of the uncertainty in the position of an individual atom in a crystal structure.[154] B-factors often locally converged based on the surrounding environment, and can vary between different crystal structures, so a normalised B-factors (B$_{norm}$) will also be included in the analysis, where the B-factor of an atom has been divided by the mean B-factor of it's crystal structure.[155] EDIA and Z_{obs} are both assessments of the electron density in proximity to an atom, to identify if there is sufficient density to support the atom assignment, and are discussed in more detail in Section 1.5. The scales of the different metrics differ, but a low score for both EDIA and Z_{obs} indicate that there is less electron density. The opposite is true for B-factors and B_{norm}, where a higher score indicates greater uncertainty in the location of the atom.

Some correlation can be observed between the different atomic metrics. Figure 3.9 shows the distribution of all of these factors; EDIA, Z_{obs}, B-factor and B$_{norm}$ for 569 water molecules in the dataset (632 water molecules, where 63 solvent molecules have been removed based on H-bonding classification). Both EDIA and Z_{obs} are calculated using an atom's B-factor to estimate the atom's radius. EDIA uses a look-up table calculated by averaging B-factors from many structures to assess structure quality, while Z_{obs} uses the B-factor directly. EDIA and Z_{obs} use of B-factors explains why they are both correlated with B-factor (ρ =-0.58 and -0.69 respectively), and why the correlation is stronger for Z_{obs} than for EDIA, as Z_{obs} uses the B-factor directly in the calculation, while EDIA uses an average B-factor, given the resolution, atom type and atom charge. Z_{obs} and B-factor are the most correlated pair of metrics considered. Both EDIA and Z_{obs} are measures of whether there is sufficient electron density for a given atom, albeit calculated in a different way. The two methods are correlated ($\rho = 0.61$, where ρ is Pearson correlation coefficient), however the EDIA method has an upper limit on the EDIA value of 1.2, which may explain the non-linear correlation observed. The maximum EDIA score of 1.2 is imposed, rather than having a physical meaning. If the EDIA limit of 1.2 were removed, a more linear distribution may exist with a higher Pearson coefficient. The scatter plot of B-factor against B$_{norm}$ has a ρ correlation of 0.48. Regions of highly correlated points can be seen in B-factor against

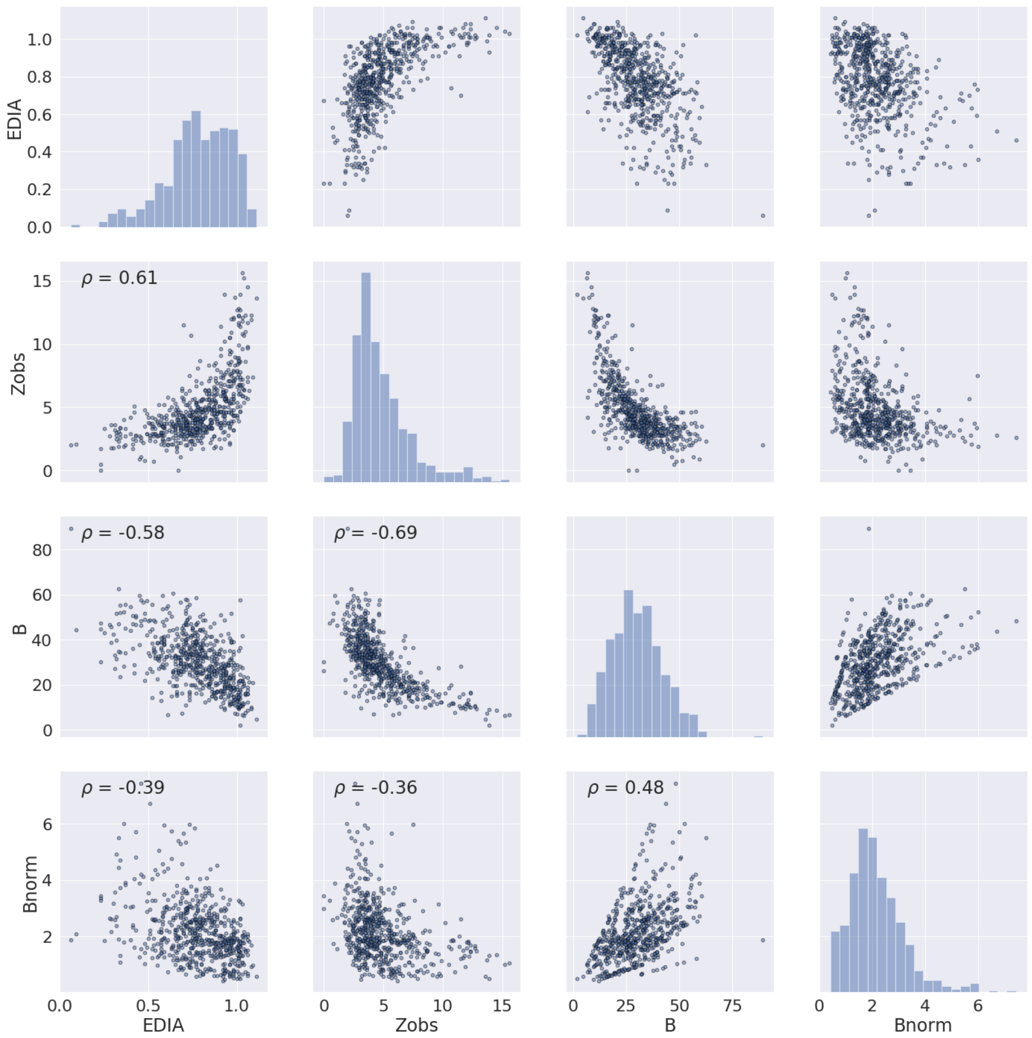


Figure 3.9: Correlation plot for different crystallographic measures considered; EDIA, Z_{obs}, B-factor and B$_{norm}$. Diagonal histogram plots show the distribution of each metric, and non-diagonal scatter plots show the correlations between different metrics. Data points are transparent, which means that regions that appear darker correspond to multiple water molecules. Annotations show the Pearson correlation coefficient (ρ) for all combinations of metrics.

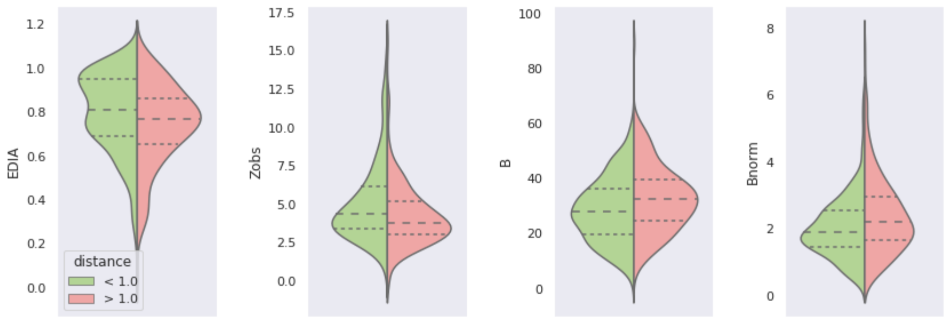


Figure 3.10: Violin plots showing the distribution of metrics considered, for water molecules found (green) or missed (red) to within 1 Å cutoff in fixed GCMC simulations. Results are shown for EDIA, Z_{obs}, B-factor and B$_{norm}$. Note the different axis for each metric. Median values are shown by large dashed lines, while interquartile ranges are indicated by the small dashed lines.

B$_{norm}$ plot, which arise when water molecules from the same structure have all been normalised by the same average B-factor. B$_{norm}$ is frequently considered a more unbiased metric to use than B-factors,[155] however while this in no way indicates which is more reliable, the non-perfect correlation indicates that differences will arise depending on which value is considered. The diagonal histogram plot of B_{norm} shows that the modal value is around 2, showing that water molecules in crystal structures are likely to be more disordered than the average atom of the system. Water molecules are therefore more likely to be disordered, and less likely to be correctly assigned than other atoms in protein structures.

GCMC is more likely to miss water molecules that have less electron density and are more disordered. Distributions of the different metrics of water molecules that have been found or missed are shown in Figure 3.10. This highlights some differences in the water molecules found and missed. The distribution of both EDIA and Z_{obs} values of crystallographic water molecules that are missed by GCMC is shifted to lower values than for those water molecules found. Crystallographic water molecules with less electron density to support their placement in the crystal structure are more likely to be missed by GCMC, whether EDIA or Z_{obs} are used to assess the electron density. The shift in electron density may suggest that some water molecules that are missed by GCMC may have been incorrectly assigned in the electron density. Looking at the distribution of both B-factors and B_{norm} show that water molecules missed by GCMC have marginally higher values than those found. B-factors are a measure of the uncertainty in an atom's position, and therefore water molecules with higher B-factors or B_{norm} are less clearly resolved. B-factors differ to EDIA or Z_{obs}, as they do not suggest if a water molecule is truly there or not, but may suggest water molecules where a longer, more lenient distance cutoff would be needed to locate it by computational methods.

Different published methods have used inconsistent criteria to select appropriate water molecules for testing. Validations of other computational methods have excluded some crystallographic water molecules from their test sets based on various assessments of the crystallographic water molecules. GCMC[54] considers water molecules that are 'conserved', i.e. that are observed in multiple crystal structures of the same target. WarPP[82] in addition to only considering water molecules that have two hydrogen bonds, only assesses water molecules with an EDIA score of >0.24 (a previous publication suggested an EDIA >0.8 to suggest sufficient electron density[112]). If the success rate of GCMC with the FDA dataset is analysed as similarly to the WarPP results as possible — using a distance cutoff of 1.0 Å, a minimum EDIA score of 0.24, GCMC sites that are over 50% occupied for a fixed simulation for water molecules that form at least two polar contacts with protein or ligand at hydrogen bonding distance — then the success rate is 72 %. Figure 3.10 illustrates a difference between different metrics and the likelihood of GCMC to correctly predict the site. Owing to the difference in distributions of water molecules found and missed, if water molecules with poorly scoring metrics are excluded from the dataset, the apparent success rate of the method will increase. While EDIA scores of 0.24 and 0.8, and a Z_{obs} score of 1, have been suggested as suitable cutoffs, these values seem to be a ball-park suggestion, rather than arising from a physical measure. In addition, there is no cutoff for B-factors, or B_{norm} to distinguish a water molecule as ordered or disordered. For this reason, no specific value of any metric has been chosen, instead the success rate of the method has been plotted against a range of each metric, Figure 3.11.

Filtering the dataset based on crystallographic metrics improves the success rate. Where each of EDIA and Z_{obs} have a minimum of zero, all 632 crystallographic water molecules will be considered, and the success rates are consistent with those shown in Figure 3.11. As the minimum for each increases, water molecules will be excluded from the determination of the success rate, and a consequence of excluding water molecules with low scores, the success rate of the method is increased. For EDIA there is a notable increase in the success rate at around 0.8. The success rate increases by 13% (fixed) and 19% (sampling) for EDIA, and 16% (fixed) and 14% (sampling) for Z_{obs} depending on the cutoff applied. B-factor and B$_{norm}$ differ to EDIA and Z_{obs}, as they are measures of atomic uncertainty, where a higher value indicates more doubt. The plot for B-factor and B_{norm} need to be considered in the opposite direction, where the maximum B-factor/B$_{norm}$ will include the whole dataset (53% and 68% success rates for sampling and fixed respectively), and as the plot moves to the left, the dataset becomes more selective, removing water molecules with high B-factors. If a smaller, maximum B-factor/B$_{norm}$ is considered (only considering water molecules with more certainty in their position) then the success rate will increase. The gain in success rate across the range of maximum B values is 14% (fixed) and 13% (sampling) for B-factor, and 5% (fixed) and 14% (sampling) for B$_{norm}$.

The changing success rates indicates how unreliable it would be to compare two different water placement methods. Analysis into the effect that different cutoffs of differing metrics have on the apparent success rate does not indicate suitable crystallographic quality to be used, it does highlight the large degree to which a success rate can vary based on water molecules analysed. It is difficult to compare two computational methods. It can be hard to understand whether a difference in success rate between methods is due one method being *better* or purely down to differences in protocol. Using no cutoff for any of the metrics

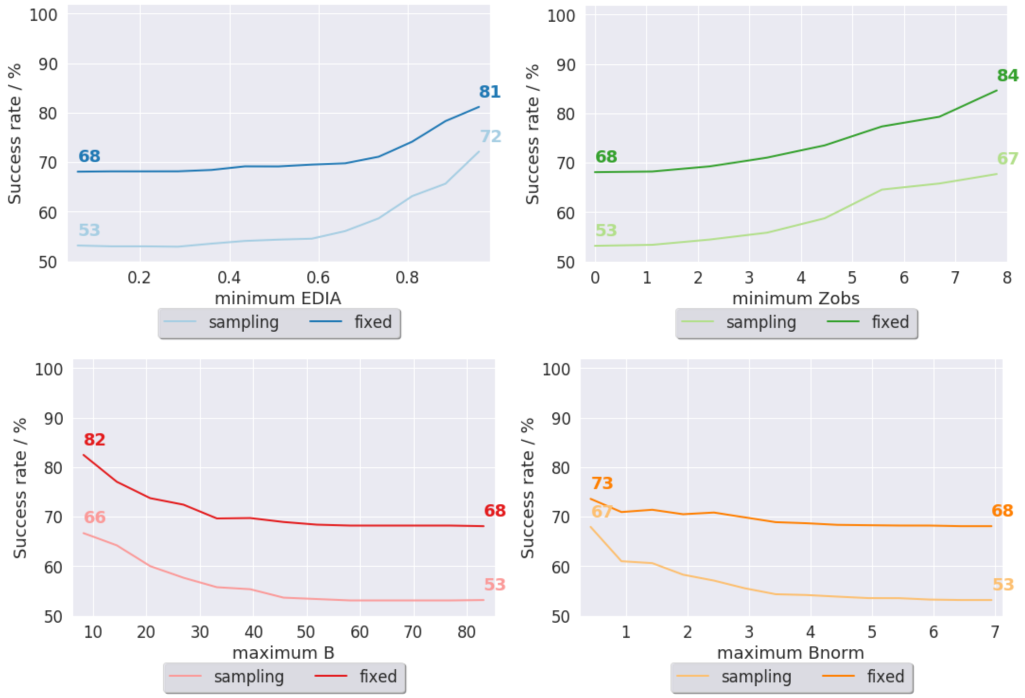


Figure 3.11: Success rate for both fixed and sampling simulations, when the data set of 569 water molecules (632 water molecules, where 63 solvent molecules have been removed based on H-bonding classification) is filtered to remove low electron density water molecules (EDIA or Z_{obs}), or to remove high uncertainty (B-factor or B_{norm}) water molecules. The success rate calculated using a 1.0 Å distance cutoff, with GCMC cluster centres of occupancy greater than 50%. The x-axis indicates the cutoff used for the minimum, or maximum value of the metric used the filter the dataset. All points plotted correspond to a dataset of at least 50 of the initial crystallographic water molecules.

seems naïve — water molecules with little electron density, or much disorder, are being used to benchmark a method. Excluding water molecules, however, begins to lean towards cherry-picking of results, biasing a quoted success rate to higher values. All of the values plotted in Figure 3.11 are determined from a dataset of a minimum of 50 water molecules, however as this is an average of 0.5 water molecules per structure analysed (50 waters from 100 structures), this too feels overly selective. Ultimately, choosing any cutoff or quality filter on the results will alter the success rate quoted. What seems to be more important is assessing all methods to the same criteria to ensure a fair comparison.

Looking at crystallographic metrics helps somewhat to suggest which water molecules may be artefacts, but they do not indicate where a crystallographic water molecule has not been assigned, but would fit the electron density. As the success rate increases when filtering out low-scoring water molecules by different metrics, this suggests that some water molecules may not be real, and are an artefact of the crystallographer and the refinement process. These metrics are only recorded for water molecules that are present, and do nothing to indicate where water molecules may be present experimentally, but have not been assigned. Understanding if hydration sites are being missed as they are partially occupied, or if they are disordered, is also difficult to quantify. Experimental locations of water molecules are the only information against which it is possible to test the success of computational methods, but this does not mean that the experimental data is without flaws.

3.3.5 Additional hydration sites

GCMC predicts many more hydration sites than are observed in the crystallographic structure, as illustrated in Figure 3.1. These additional sites can arise for a range of reasons. Additional sites may be due to low occupancy GCMC clusters, where the occupancy of the GCMC cluster is such that the electron density would be too low to resolve. This may also occur for disordered water molecules, where the electron density will be blurred, and may not be assigned to a water molecule. Another reason for the additional GCMC hydration sites is that, as the GCMC region is cuboidal, and defined automatically across the dataset to have a 4 Å padding around the heavy atoms of the ligand. Both the shape of the GCMC region, and as the region has been uniformly defined irrespective of the nature of the binding site, for many systems, the GCMC box extends over bulk-like water. Bulk like water molecules are unlikely to be resolved in the crystal structure due to their disorder, but will be identified by GCMC, which explains many additional hydration sites that are distal to either protein or ligand. The final explanation for the additional water sites as identified by GCMC is the temperature at which both the simulation and the crystallography is performed. The majority of the dataset has been resolved at 70 K, while all the GCMC simulations are performed at 300 K. It is likely that this temperature difference means that entropically bound water molecules are more likely to be observed in the low temperature crystal structures, while enthalpically bound water molecules will be stabilised at ambient room temperature, but this is difficult to quantify and hard to test. One possibility to probe the differences in bound water molecules at various temperatures would be to study structures that have been resolved at both low and high temperatures, or for complexes where neutron diffraction data is available.

3.3.6 Hydrogen bonding networks - 2RIN

One of the main advantages that GCMC has is the ability to sample networks of molecules. Above, the analysis has been performed on cluster centres derived from simulations. These results show that, with any analysis, fixed simulations have higher success rates than the equivalent simulations that sample the protein-ligand environment. While the success rates — the likelihood of correctly predicting the locations of crystallographic water molecules — are lower with sampling simulations, despite the additional computational expense, there are other advantages of sampling simulations. The fixed simulations assume that the starting location is correct. The starting location may not be at a minimum for multiple reasons — unclear electron density, ambiguous assignment of the density, or if the complex has been generated using a computational technique such as docking. Allowing the protein ligand to sample allows the system to move away from high-energy conformations, and sample an ensemble of conformations for a complex.

One example of the dataset of 105 will be considered to illustrate the benefits of including system sampling within GCMC simulations; ABC-transporter choline binding protein with acetylcholine (PDB: 2RIN). Acetylcholine is a small molecule that has one crystallographic water molecule within the GCMC region simulated. The water molecule is protein-bound (based on both H-bonding or Nittinger et al. classifications), and correctly located by both fixed and sampling GCMC simulation, to 0.35 and 0.62 Å respectively. GCMC however, predicts more hydration sites than are resolved in the crystal structures, as discussed in Section 3.3.5.

Fixed GCMC

Fixed simulations of 2RIN reproduces the crystallographic sites. There are four GCMC cluster centres from the fixed GCMC simulation, three of which have occupancies of 100%, and one that is only 1.5% occupied. All three high occupancy water molecules correspond to crystallographic water molecules, one inside the GCMC region, while the other two crystallographic water sites are 0.8 Å outside of the GCMC region, shown in Figure 3.12. The success rate of this system in isolation is 100%.

Sampling GCMC

Sampling simulations result in more cluster sites, with the crystallographic water molecules reproduced as in the fixed simulations. In contrast to the fixed simulation results, the cluster locations from the sampling simulations are shown in Figure 3.13. Instead of four GCMC water clusters, there are now 13 cluster locations from the sampling results. As before, three of these sites are 100% occupied (IDs 1 - 3), and are within 1 Å of the three high occupancy clusters from the fixed simulation and the crystallographic water molecules. Five sites (IDs 4 - 7) are partially occupied, with occupancies of 81, 73, 35 and 35%

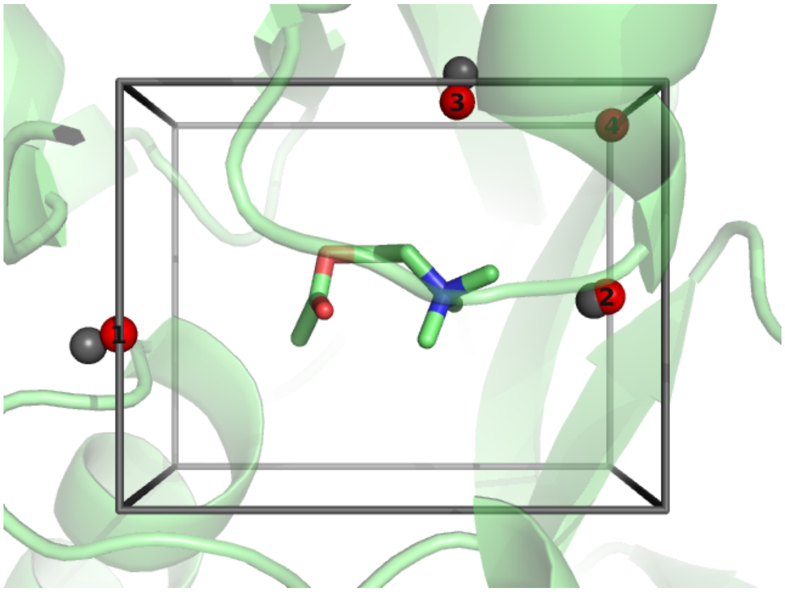


Figure 3.12: Acetylcholine bound to ABC-transporter choline binding protein. GCMC cluster centres from the fixed simulations are shown. GCMC region is shown by black line, protein shown in green cartoon, acetylcholine coloured; carbon - green, nitrogen - blue, oxygen - red. Crystallographic water locations are shown by yellow spheres. GCMC cluster locations are labelled, and coloured using a spectrum of blue (low occupancy) to red (high occupancy). Labels are the cluster IDs.

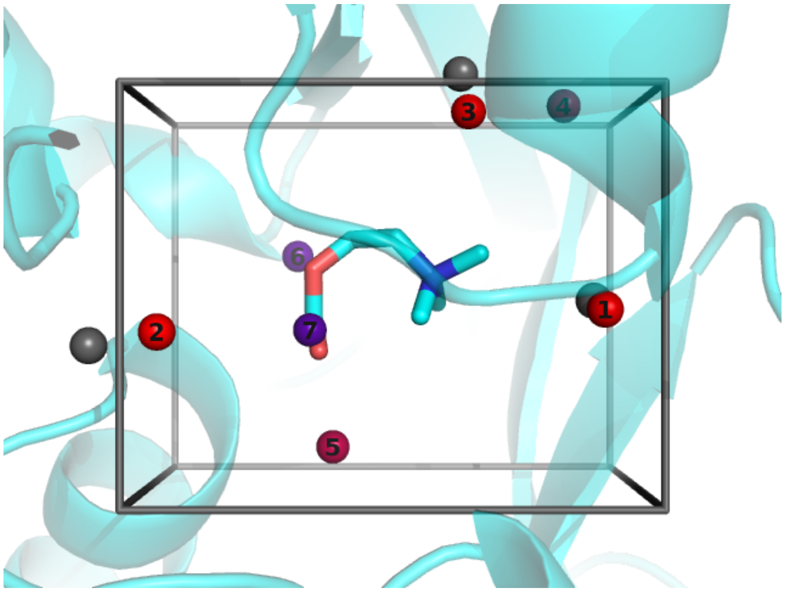


Figure 3.13: Acetylcholine bound to ABC-transporter choline binding protein. GCMC cluster centres from the sampling simulations are shown. Colours are the same as Figure 3.12. The protein-ligand conformation shown is the starting conformation, while the cluster locations are determined from the dynamic simulation.

respectively. Clusters with occupancies lower than 25% have been excluded for clarity.

One cluster site has been identified that is clashing with the starting location of the ligand. The most notable aspect of Figure 3.13 after the number of cluster centres, is the position of cluster 7, which is clashing with the ligand. In the fixed simulation, the ligand will not move from the position shown, so no attempts to insert the water molecule at this position would be accepted. The conformation shown however, is the starting conformation, taken from the crystallographic structure. The position of cluster 7 indicates that the ligand has moved sufficiently for a water molecule to now occupy this site.

The correlation of a pair of water molecules can be calculated by counting

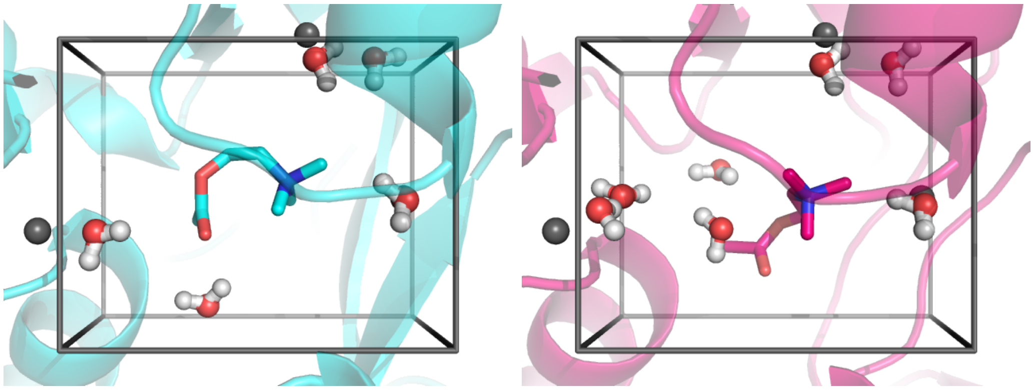


Figure 3.14: Two representative frames of acetylcholine bound to ABC-transporter choline binding protein. GCMC cluster centres from the fixed simulations are shown as spheres, with low occupancy sites removed, while water molecules from the representative frames are shown as sticks. Two binding modes of the ligand are observed, with one similar to the crystallographic position (left) and the other novel (right), which would not be observed in the fixed simulations. The observed ligand flip reveals two hydration sites, 6 and 7, while displacing the water molecule at cluster site 5.

the frames in which they are observed together, and comparing this to how many frames would be expected by chance. If a pair of water molecules are both 50% occupied, if they were non-correlated, they would be expected to be observed together 25% (50% x 50%) of the simulation. If they are observed together significantly more, or significantly less than this, then they can be considered correlated or anti-correlated. Cluster sites 5 and 7 are 3.3 Å apart, which could be a long hydrogen bond, however they are -17.2% anti-correlated (8.5% observed - 25.7% random). Two representative frames of the simulation are shown in Figure 3.14 which demonstrates the anti-correlation of GCMC clusters 5 and 7. The acetylcholine flips during the simulation, and an alternate binding conformation is observed. One of these conformations is complementary to GCMC cluster 5, while the other stabilises GCMC cluster 7. Sampling of the 2RIN system is able to reveal two possible binding modes, both of which are in agreement with the crystallographic water locations, but with distinct water networks. While the results are more complex, and the system is able to move away from the crystallographic structure, with lower success rates, sampling GCMC simulations have a significant advantage. Both of these ligand conformations and water networks would be of importance if part of a drug design effort, and would not be seen by the fixed simulations, or any other published water location methodology that does not sample the protein-ligand environment.

3.4 Conclusion

This chapter looks at the rate at which GCMC simulations correctly reproduce crystallographic water locations. The locations of water molecules have been predicted for a dataset of 105 complexes of FDA drug molecules to protein targets. The dataset has been generated with attempts to ensure good crystallographic comparison — 2000 or more recent release date, 2.5 Å resolution or better and no structures with obvious crystallographic contacts. The placement of water molecules has been attempted while both sampling the protein-ligand environment and while holding it rigid.

A success rate of 59% and 72% has been demonstrated for sampling and fixed simulations respectively, at a 1.0 Å cutoff, for GCMC water molecules with greater than 50% occupancy, for crystallographic water molecules with at least two polar contacts to protein or ligand. The major result of this Chapter is not the success rate itself, although this is gratifying, but rather the demonstration of how variable the quoted success rate can be, depending on the protocol implemented. The results vary depending on the occupancy of GCMC water molecules considered, the types of water molecules included in the dataset (and how those types of water molecules are defined) and if various crystallographic metrics are used to filter out lower-quality crystallographic water molecules. For a 1.0 Å cutoff for sampling simulations, success rates vary between 37% (GCMC sites greater than 75% occupied for all 632 crystallographic sites) to 72% (GCMC sites greater than 50% occupied, for h-bonding water molecules with an EDIA score greater than 0.9). This indicates that care needs to be taken when comparing published results of differing methods, as much variation exists in the published protocols.

Here, the accuracy of GCMC has been tested on correctly locating water molecules for a large dataset. In the following chapter, GCMC has been combined with ligand free energy calculations to allow for dynamic adaptation of the active site water network along the along the alchemical pathway, known as GCAP. GCAP will be demonstrated on two systems, SD and A_{2A}, both of which have limited crystallographic data where crystal structures are not available for all protein-ligand complexes considered, and some of those which are available are poorly resolved.

Chapter 4

Ligand binding free energies with displaceable water molecules

4.1 Introduction

This chapter has been completed with help from CCA and GAR. GAR is responsible for the initial implementation of GCAP in ProtoMS. CCA has written the surface-GCAP analysis script that performs the calculation of 2D-MBAR. All simulations, and analysis herein was performed by HBM.

This chapter will look at the development of grand canonical alchemical perturbations (GCAP), whereby GCMC sampling of active site water molecules is performed during relative ligand binding free energy calculations. This allows for congeneric ligands to be simulated accurately, despite having differing active site water networks.

Rational drug design often involves making stepwise modifications to a known ligand to improve the affinity of the molecule. The relative binding affinity of two ligands can be calculated by performing the perturbation of one ligand, into the other in both bulk solvent, and in the bound ligand environment. Issues can arise if the bound environment of the two ligands considered differ. If the change in ligand causes a perturbation to the active site water network, then the change in ligand will interfere with this. The water network may be unable to adapt appropriately within the timescale of the simulation if the bound water molecules are unable to exchange with bulk, or if there is a kinetic barrier to water unbinding. This issue will be particularly notable for occluded binding sites. If the water network is unable to adapt, then either one or both ligands may be simulated in a non-native state, which can introduce errors into the simulation. Relative binding free energies simulated in the NVT or NPT ensemble may incorrectly indicate that one ligand is more favourably bound than another, if the water network is complementary to that ligand.[5]

Simulating relative ligand free energy perturbations in the μVT ensemble avoids this issue, as the sampling of active site water molecules is enhanced via grand canonical sampling. The networks of the two ligands considered do not need to be the same — or even need to be known *a priori* — as the active site water molecules are able to adapt across the alchemical pathway.

4.1.1 Grand canonical alchemical perturbations

GCAP is the methodology whereby relative ligand free energies can be calculated, in combination with GCMC to correctly model the active site hydration state of the ligands, for every λ intermediate. This allows for the correct, equilibrium hydration state to be modelled for both ligands, as well as all intermediate λ states. As with GCMC simulations, GCAP can be performed at a range of B values, resulting in a two-dimensional simulation, over a range of B and λ values. This results in a two-dimensional binding free energy surface, and hence will be referred to as surface-GCAP. Alternatively, if only B_{eq} is simulated, each λ window is dynamically hydrated to an extent appropriate for equilibrium with bulk water, and the result is a one-dimensional free energy curve along λ (single-GCAP). As GCAP is able to alter the hydration of the grand canonical region of the simulation, this allows for the relative free energy of two ligands with differing water occupancies to be determined in a single free energy simulation. GCMC has been used previously to study changing water networks for an absolute binding free energy calculation.[156] Unlike previous work, we are simulating fully in the μVT ensemble, in contrast to only periods of μVT equilibration. Further, we show here how simulations using multiple B values can be used to construct self-consistent thermodynamic cycles for sets of ligands, with the full benefits of replica-exchange in both B and λ.

For single-GCAP simulations, as only λ is varied and B is constant at the equilibrium B_{eq} value, the relative free energy of two ligands can be determined using classical free energy approaches, discussed in Section 1.2. As with running GCMC at a single B value, single-GCAP is only able to determine the equilibrium number and location of water molecules, but not the binding affinities of the water network. RE may be performed between simulations at different λ values to aid convergence.[23,139]

In surface-GCAP simulations, a range of both λ and B values are simulated. An illustration of the surface-GCAP simulations is shown in Figure 4.1. The surface-GCAP simulations are aided by replica exchange (RE) in both dimensions; λ and B.[157] The relative binding free energy of the ligands in their equilibrium hydration states, as well as the number of water molecules, their locations and the binding free energy of the water networks can all be determined from surface-GCAP. MBAR is trivially applied to two dimensions, allowing for all available states of the simulation to contribute in calculating the relative free energy of the two ligands and their associated water networks.[22] This is calculated by using the reduced potential function, $u_i(\mathbf{x})$, Equation 4.1 with the MBAR estimator.

$$u_i(\mathbf{x}) = \beta \left[U_i(\mathbf{x}) + \mu_i N(\mathbf{x})\right] \tag{4.1}$$

i is the index over all states, U_i is the potential energy according to the i^{th} Hamiltonian, μ_i is the chemical potential of the i^{th} state and N is the occupancy of water molecules of state $\mathbf{x}$. This 2D-MBAR allows the free energy of the ligand perturbation to be calculated from the entire surface-GCAP simulations, using statistically optimal contributions from all simulated states. Surface-GCAP is advantageous over single-GCAP, as it is able to calculate the binding free energy of networks of water molecules for any perturbed state of the ligands, while also benefitting from the convergence advantage of RE in B.[157] The computational resources required by single-GCAP is determined by the specified number of λ windows. Surface-GCAP requires the equivalent resources multiplied by the number of B values simulated.

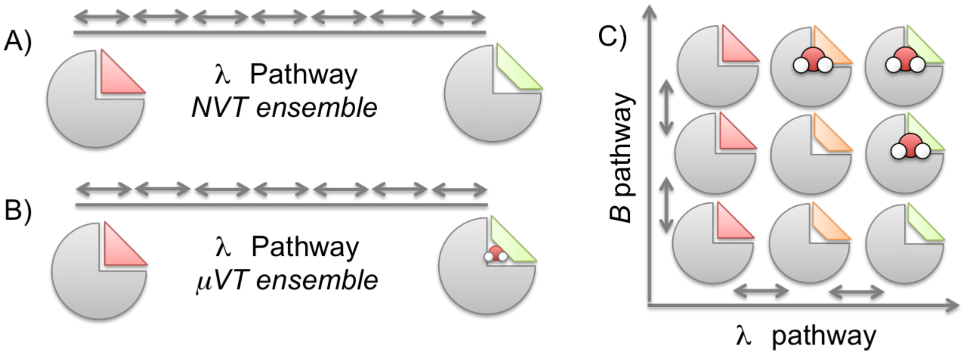


Figure 4.1: Relative ligand free energy methods, where one ligand (red) is perturbed to another (green) across a λ coordinate. A) A typical relative ligand free energy simulation where the perturbation is performed in an NVT ensemble and the hydration state of the protein-ligand system is unable to adapt to the perturbation. B) Single-GCAP. The same perturbation in the grand canonical ensemble, where insertion and deletion moves allows the water occupancy to vary across the λ pathway. The equilibrium chemical potential (B_{eq}) solvates each ligand in dynamic equilibrium with bulk water. C) Surface-GCAP. Both a λ pathway and range of B values are simulated, generating a two-dimensional network, with RE between neighbouring states. The relative free energy between different B and λ values can be determined from the surface, using MBAR. Free energies of water networks can be calculated by using the GCI equation at a given λ value. Calculating relative ligand binding affinities requires a corresponding bulk water ligand perturbation. The bulk leg contributions are included in the calculation, but excluded from this graphic for clarity.

Two systems will be used to present this method; Scytalone Dehydratase (SD), used previously in Chapter 2, and a water soluble form of adenosine A_{2A}receptor (A_{2A}). SD has been used previously as a test systems for free energy methods; there are three similar ligands on a common scaffold, with significantly different binding free energies.[3] These differences have been suggested to be due to the favourable displacement of an active site water molecule.[2] Michel et al. used this system to demonstrate stepwise free energy calculations, whereby the ligand perturbation is performed, followed by DD of water molecules in the system.[5] Their method will be reproduced herein for comparison to the GCAP methodologies.

The GCMC region for SD will be a 4x4x4 Å^3 cubic box focussed on the single potential water site illustrated in Figure 4.2.

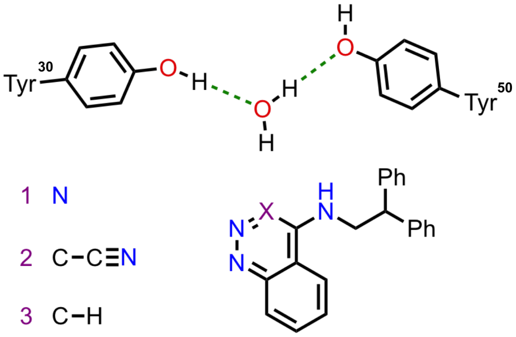


Figure 4.2: Representation of the SD ligand binding site, with the structures of ligands **1**, **2** and **3** shown. The potential active site water location is shown, with hydrogen bonds (green dash) to two active site tyrosine residues. Ligand **2** is the only compound for which a crystallographic structure is available (PDB:3STD), in which there is no water molecule present. The binding modes of ligands **1** and **3** have been assumed to be the same as ligand 2. The presence of a water molecule with the smaller ligands **1** and **3** has been studied by Michel et al.[5]

For A_{2A}there are twelve antagonists in the dataset of 1,2,4-triazine derivatives published,[4] where various aromatic substitutions have been made to either ring A or ring B, shown in Figure 4.4. Ligand names, R group numbering and ring labelling are consistent with the previously published work.[4] Of the twelve ligands, three have been selected for free energy calculations here; ligands **E**, **F** and **G**, Figure 4.4. These were chosen as both **E** and **G** are the only holo-crystal structures available (**E** PDB:3UZC, **G** PDB:3UZA), the differences between the ligands are all located on ring A, and the relative free energies calculated from both the K_i and K_D data are consistent to within 1 kcal·mol^{-1}, which is the level of accuracy for which we would aim computationally. More details of the comparison of K_i and K_D are outlined in Section A.4. Any experimental $\Delta G^{\ominus}$s reported herein will correspond to the K_D results, calculated by surface plasmon resonance (SPR) binding analysis. The crystallographic structures of ligands **E** and **G** are both 3.3 Å

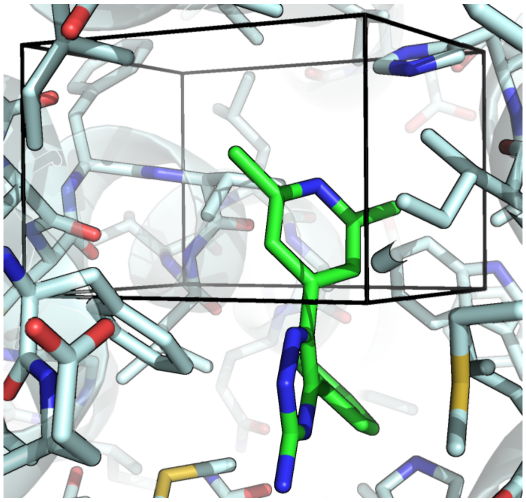


Figure 4.3: Active site of A_{2A} ligand **G** (PDB:3UZA). Protein residues (light blue) Ligand **G** (green) shown as sticks, with nitrogen (blue), oxygen (red) and sulfur (yellow) shown. The GCMC box region, shown as black lines, covers ring A of ligand **G**, and the active site cavity near ring A. No water molecules are shown, as there are no resolved water molecules in the crystal structure.

resolution, respectively. As these structures are low resolution, no crystallographic water molecules have been resolved. While the lack of crystallographic water locations makes the validation of GCMC more difficult, it illustrates a system where water placement methodologies can be of most help.

As the ligand perturbations are all on ring A, a GCMC region covering a protein pocket near ring A will be sampled. As there are no crystallographic waters it is unclear how many hydration sites this box will cover, but it will likely be more complex than the single water site considered for SD. The cavity near ring B will be naïvely solvated using ProtoMS[95] during the system set up.

4.1.2 Free energy surfaces

To create the free energy surfaces, PMFs are calculated along B using GCI, and along λ using rigorous free energy methods. These are combined to generate a free energy surface using least-squares fitting. For PMFs along B, free energy values

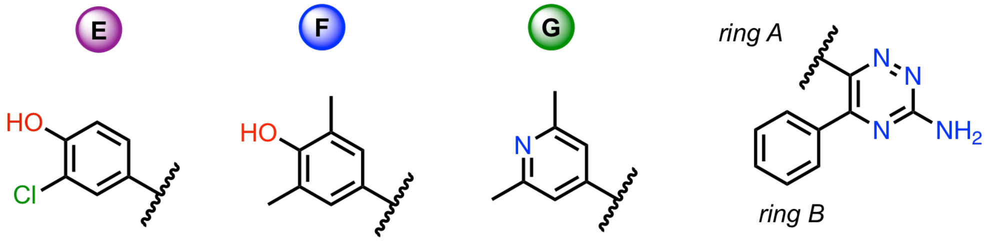


Figure 4.4: Three A_{2A} ligands that will be considered herein. All of the substitutions are to ring A in the molecule.

for states with non-integer occupancies are determined by linear interpolation of the binding free energy curves output by GCI.

In principle it is possible to calculate free energy surfaces directly using MBAR. The free energies produced between states with differing B values will include contributions from changes in chemical potential however that are not physically meaningful in the context of the binding free energies of interest in this work. The above approach produces consistent Helmholtz free energy surfaces using GCI. MBAR free energy differences between states at the same B value are consistent with NVT free energy cycles.

In all cases, $\lambda = 0$ corresponds to the larger ligand, and $\lambda = 1$ to the smaller.

4.2 Methodology

4.2.1 System set-up

Proteins

For all proteins simulated, the amber14SB force-field has been used.[6]

SD protein structure used is from the 3STD PDB entry. The protein was scooped to a radius of 15 Å. The protonation and tautomer states of the proteins were determined using molprobity.[140] In Chapter 2 two hydration sites in the SD active site were considered, water molecules A and B. Here, the water site A will be used to define the GCMC region, as it is close to the site of the changes on the ligands.

A_{2A} protein structure used is from the 3UZA PDB entry. For A_{2A}a scoop of 20 Å was used, with side chain and backbone sampling in the inner 16 Å, and side chain only beyond that. The protonation and tautomer states of the protein were determined using Maestro.[145] A_{2A} has an active site His_{278} residue; this was ϵ protonated during the set up. Owing to its proximity to the GCMC region, the single-GCAP simulations were repeated for the δ protonated His_{278}. GCMC results can be dependent on the tautomer and rotamer of histidine used in a simulation.[158]

Ligands

For all ligands, the gaff14 forcefield has been used with AM1-BCC charges.

Three similar ligands bound to SD have been studied, ligands **1**, **2** and **3**. The 3STD PDB entry has the bound structure of ligand 2, from which the other two ligands binding positions has been assumed.

For A_{2A} the PDB file of the complex containing ligand **G** is used (PDB:3UZA). Models of the other ligands (**1** and **3** for SD, and **E** and **F** for A_{2A}) studied were generated from these scaffolds. As the perturbation from ligand **E** to ligand **G** involves both the addition and removal of functional groups, it has been performed in two steps, via the intermediate, where the C-OH group of ligand **E** has been perturbed to a N atom, but the meta groups are unchanged, shown in Figure 4.5.

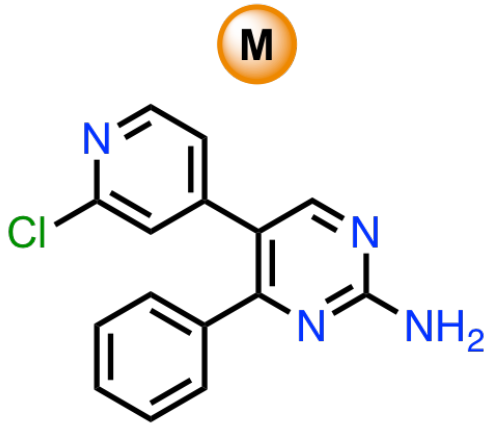


Figure 4.5: Ligand **M** (for mutant), not included in the published dataset,[4] but used as the mid-point for the **E** — **G** leg, as this perturbation requires both the growing and shrinking of different R groups. It is more straightforward to calculate the relative free energy of both **E** — **M** and **G** — **M** and use this to calculate the **E** — **G** leg. **M** was calculated to have lower affinity than any of **E**, **F** or **G**.

Solvation

For all water simulated, the TIP4P force-field has been used.[141] Protein-ligand complexes were solvated using a half-harmonically restrained sphere of radius of 30 Å, with any crystallographic water locations retained, apart from the FDA dataset where all crystallographic water molecules were removed. This includes solvating any sterically available active site regions. For the free simulation legs, each ligand is solvated in a cubic box with a padding distance of 10 Å between ligand and box edge. For grand canonical simulations, water molecules within the GCMC region are removed prior to the simulation.

4.3 Simulation protocol

4.3.1 Ligand binding affinities

For any simulation performed with either multiple λ windows or multiple B values (or both), replica exchange between neighbouring B and λ values was attempted every 100,000 moves. For consistency with previous publications, a non-bonded interaction cutoff of 10.0 Å was used for SD, and a cutoff of 15.0 Å for A_{2A}simulations was used.

Single-topology alchemical transformations were performed on pairs of SD ligands. Perturbations were performed in two stages; considering the perturbation as taking place from a large molecule to a small, the electrostatic parameters first perturbed, followed by the van der Waals (vdW) interactions. Each simulation is split across 16 equally spaced λ windows. These perturbations are performed both in the bound state and for the ligand in bulk solvent. 5M MC equilibration steps are performed, followed by 40M production steps. The ratio of MC moves for each system is shown in Table 4.1.

GCMC has been shown previously to be consistent with double decoupling methods for calculating binding free energies of water molecules.[54,157] To validate the thermodynamic consistency of GCAP, the SD system was simulated in the bound state both with and without the active site water molecule. In addition, DD has been performed on the active site water location in SD with all three ligands, consistent with the method described by Ross et al.[157] These simulations generate the thermodynamic cycle shown in Figure 4.7, that allows for a comparison to the GCAP results, in addition to the experimental data.

4.3.2 GCMC

For SD and A_{2A}, a region of the active site was defined using a GCMC box over a region of the active site. For SD, this is a small box over a single active site water molecule and for A_{2A}, a box covering the active site cavity near ring A

Table 4.1: MC move ratios for each simulation performed. A — indicates that no GCMC type moves were performed.

Simulation	solvent	protein	solute	GC insertion	GC deletion	GC sampling	n equilibration / M	n production / M
SD AP	280	218	2	—	—	—	5	40
SD DD	280	218	2	—	—	—	5	40
SD GCAP	280	218	2	167	167	167	5*	80
A_{2A} GCAP	376	118	7	167	167	167	10*	120
A_{2A} naïve	376	118	7	—	—	—	10	60

Table 4.2: Details of GCMC region used for each system. The GCMC region is cuboidal. B_{eq} is calculated from the GCMC volume using Equation 2.13

System	origin (x,y,z)	length (x,y,z) /Å	Volume /Å^3	B_{eq}
SD	24.141, 11.225, 32.916	4.000, 4.000, 4.000	64.	-9.70
A_{2A}	-44.253, 0.565, -47.602	9.784, 6.533, 7.844	501.4	-7.65

was used, shown in Figure 4.3. GCMC region details are available in Table 4.2. The simulation consists of an initial GCMC equilibration of 5M MC moves, with a 1:1:1 ratio of insertion, deletion and GC water sampling moves. Following this, 5M equilibration and 80M production MC steps are attempted on the entire system with the sampling ratios shown in Table 4.1.

For SD, GCMC was performed at 16 equally spaced B values from -22.7 to -7.7. As the binding free energy of the water molecule with ligand **3** is unfavourable, higher B values are required to couple the water into the system; therefore for this ligand GCI was repeated for 16 B values from -12.7 to +2.3.

4.3.3 GCAP

The GCAP simulations followed the single-topology set up outlined above. These simulations were performed for the pairs of SD ligands, and pairs of A_{2A} ligands. The MC move ratios are the same as for the alchemical pertubation simulations, but with additional grand canonical MC moves. Details of move ratios are available in Table 4.1. For SD, surface-GCAP simulations were performed with the B values shown in Table A.3. For A_{2A}, surface-GCAP was performed with 10 equally spaced B values between -21.654 to -3.654 inclusive, so as cover the B_{eq} value, while also titrating down to the B value where the water occupancy is zero. Single-GCAP simulations were also performed on each system, at their respective B_{eq} values (SD: -9.70, A_{2A}: -7.65).

Table 4.3: B value ranges for surface-GCAP simulations, where B_{min} and B_{max} are inclusive. Interval shows the distance between neighbouring B values and N_B is the number of B values simulated.

System	B$_{min}$	B$_{max}$	Interval	N$_B$
SD lig **1 + 3**	-19.7	-3.7	1	19
SD lig **1 + 2**	-18.7	-9.7	1	10
SD lig **2 + 3**	-12.7	-3.7	1	10
A$_{2A}$ (all pairs)	-21.15	-7.65	1.5	10

4.4 Results

GCAP simulations have been performed on two systems — SD and A_{2A}. SD is a well-studied system,[5,54] where a small change in the ligand results in large differences in affinity due to the displacement of an active site water molecule.[3] As only one water is displaced, it is possible to validate the GCAP method using sequential steps of NVT alchemical perturbations and DD. To explore GCAP for a multi water system, a series of 1,2,4-triazine derivatives A_{2A} antagonists have been reported.[4] These A_{2A} antagonists have a range of ligand binding free energies, and previous studies have suggested that differences in affinity may arise from different active site water networks.[159,160] Using three of these ligands, **E**, **F** and **G**, shown in Figure 4.4, a thermodynamic cycle has been created, and the relative binding free energy has been calculated using both the single-GCAP and surface-GCAP methodology.

4.4.1 Scytalone dehydratase

For simple cases such as SD, where the water occupancy of the system is changing only by one for a set of ligands, a thermodynamic cycle can be constructed, as was illustrated by Michel et al.[5] Their thermodynamic cycle for SD has been reproduced using our open-source software package, ProtoMS as a comparison for the GCAP simulations, Figure 4.7.[95]

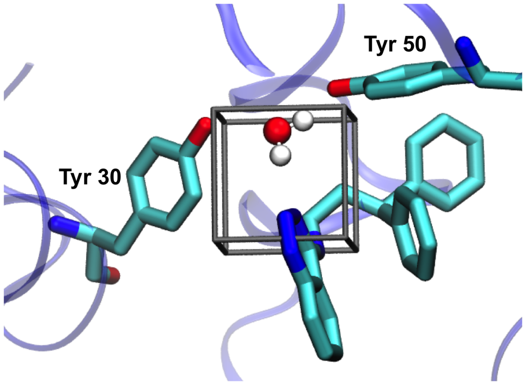


Figure 4.6: Ligand **1** binding to the active site of SD (PDB:3STD), with the GCMC region shown by a black box. Key tyrosine residues are shown. Water position is calculated from GCMC simulations.

The relative binding affinities of the ligands, and DD of the water molecule, has been performed for SD, following the protocol of Michel et al. The two triangular cycles correspond to single-topology transformations between the three ligands both in the absence and presence of the water (grey and blue cycles respectively), calculated with typical DD and alchemical perturbation simulations. The vertical legs correspond to the free energy of removing the water in each of the protein-ligand complexes, calculated by DD. A positive energy indicates a favorably bound water molecule, as it requires energy to remove the water from the system. Where the energy of the water is unfavorable, it would not be expected to be present in the bound ligand complex. These water binding free energies therefore indicate that the water is expected to be present with ligand **1**, and not with ligands **2** or **3**. For two of the cycles, the thermodynamic cycle closure is larger than the combined errors of its legs. This occurs for both cycles that involve the hydrated ligand 2, where for the ligand alchemical perturbation the water molecule is not restrained within the simulation. As the water molecule is unrestrained it is displaced into an apolar cavity of the protein, $\sim$ 8 Å from its starting position when ligand **2** is bound. As the water is restrained in the decoupling simulations, the two end points of the legs may differ. This difference

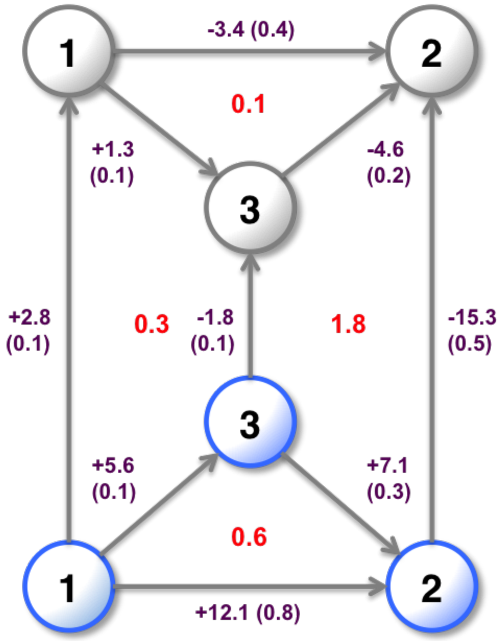


Figure 4.7: Relative binding free energies in kcal·mol^{-1} of ligands **1**, **2** and **3**, with (blue) and without (grey) the active site water molecule (shown in Figure 4.2) present. Free energies calculated using MBAR. No GCMC or GCAP simulations were used to generate this map. Vertical legs correspond to the free energy of decoupling the water from the system. This cycle is taken from Michel et al., recalculated with similar conditions where possible using the ProtoMS software package and using amber14SB and gaff14 force fields. Standard errors from four independent repeats are shown in brackets, and thermodynamic cycle closures in red. The calculated binding free energies include the free energies from the equivalent bulk-water simulations.

may be responsible for the poor closure of the cycles, however the cycles that will be presented in Figure 4.8 all close to within error.

The stepwise combination of NVT ligand perturbation with water DD correctly reproduces the rank order of affinity of the ligands. The relative binding free energy of two ligands with different water occupancies can be calculated by adding the free energies of steps between these two states. Multiple pathways exist between the states, which can result in a range of relative free energies for each pair of ligands. This has been simplified to a single set of relative binding free energies by choosing the pathway with fewest steps between two states as this represents minimal computational effort. Where there are two pathways with the same number of steps, the pathway with the smaller combined statistical error has been chosen. Figure 4.8, cycle A shows the optimum calculated free energies of binding for the ligands at their preferred hydration states. These simulations are able to correctly rank the relative binding free energies of the three ligands. However, two of the three legs are further than one standard error from the experimental result.

Calculating binding affinities using NVT ligand perturbations and water DD simulations is labour intensive, and scales poorly with additional water molecules. Multiple alchemical perturbations and DD simulations are required to generate these results, which is only feasible as the water occupancy is being varied by one. To generate a thermodynamic map for an n water network in a protein site would require n DD simulations to decouple each of the waters sequentially, or $n!$ simulations if all the different possible orders of annihilation of water molecules are considered.

The relative binding free energies of the three SD ligands have been calculated using both single and surface GCAP. GCMC has been shown to be preferable to decoupling methods as the location of the hydration sites are not needed and the binding free energy of n waters can be determined in a single simulation series, whilst also capturing cooperative binding effects in water net-

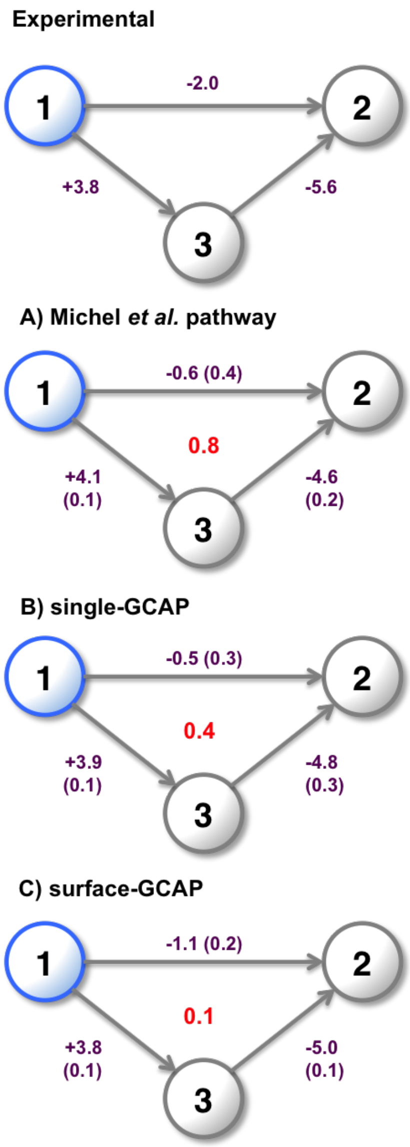


Figure 4.8: Relative binding free energies of the three SD ligands in kcal·mol^{-1}. Blue indicates a ligand expected to maintain the water in the active site. The experimental binding free energies[3] are shown along with cycle A) generated from Figure 4.7, using MBAR for ligand perturbations and water perturbations. Cycle B) calculated using single-GCAP, and cycle C) calculated using surface-GCAP. Standard errors from four repeats are shown in brackets and overall cycle closures in red.

works.[54,157] GCAP is able to perform a ligand transformation (either single or dual topology, but only single is used here), with GCMC being used at each λ value of the transformation. This allows the correct water occupancy to be adopted at each λ value. This means that the thermodynamic free energy difference between two ligands – despite any differences in their respective water occupancies or locations – can be calculated within a single simulation series.

Single-GCAP

Single-GCAP simulations are able to correctly rank order the three SD ligands. As it is possible to perform GCMC simulations at B_{eq} to predict the equilibrium water occupancy and locations, it is also possible to perform GCAP at one B value per λ value. However, this loses the sampling benefits gained from replica exchange in B in improving the precision of the results.[157] The binding affinities of water networks are also unavailable when reducing the simulation to a single B value. The results for single-GCAP simulations are shown in Figure 4.8, cycle B. The free energies calculated are consistent to within error of those calculated by separate MBAR and DD simulations (cycle A), and with smaller errors per leg.

Surface-GCAP

The relative binding free energies of the ligands calculated using surface-GCAP are shown in Figure 4.8, cycle C.

Surface-GCAP simulations are able to reproduce the SD ligand binding affinities with the best experimental agreement and smallest associated errors of the three methods considered. As described in the methods, simulations at multiple B and λ values are performed with additional RE moves. MBAR is used to estimate the free energy difference between the ligands with their optimal hydration states. These free energy results are in good agreement with both with the experimental results and the simulation results in cycle A. The surface-GCAP simulations perform the best of the three computational methods at reproducing the experimental results, although all methods are consistent to within error. The standard deviation for each simulation leg is the smallest, and the cycle closure is very small at 0.1 kcal·mol^{-1}.

Changes in the water occupancy are observed in the vdW legs of the simulation. For SD, changes in water occupancy were observed during the van der Waals (vdW) legs of the free energy calculations, when the R group of the ligand is reduced or grown in size. For this reason the free-energy surface generated by the vdW leg of the surface-GCAP is shown in Figure 4.9, for the ligand **1** ($\lambda = 1$) to **3** ($\lambda = 0$) simulation. The perturbation between ligands **1** and **3** corresponds to the change from an aromatic nitrogen (ligand 1) to an aromatic CH group (ligand 3). Examples of both electrostatic and vdW surfaces for all three pairs of ligands are available in Section A.3, Figure A.3. In all cases, $\lambda = 0$ corresponds to the larger ligand, and $\lambda = 1$ to the smaller.

The free energy surfaces are generated combining thermodynamic integration at each B value, and PMFs at each λ value. To create the free energy surfaces, PMFs are calculated along B using GCI, and along λ using thermodynamic integration. These are combined to generate a free energy surface using least-squares fitting. For PMFs along B, free energy values for states with non-integer occupancies are determined by linear interpolation of the binding free energy curves output by GCI. In principle it is possible to calculate free energy surfaces directly using MBAR. The free energies produced between states with differing B values will include contributions from changes in chemical potential however that are not physically meaningful in the context of the binding free energies of interest in this work. The above approach produces consistent Helmholtz free energy surfaces using GCI. MBAR free energy differences between states at the same B value are consistent with NVT free energy cycles.

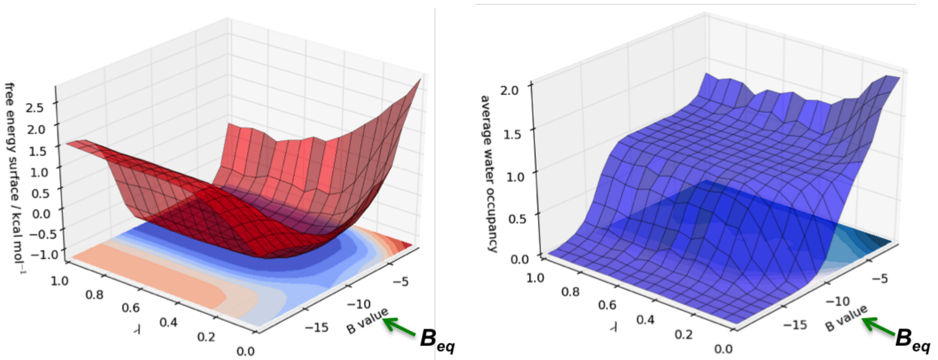


Figure 4.9: The binding free energy surface (red) and the GCMC water occupancy (blue) for the vdW leg of the surface-GCAP simulations of SD ligands, **1** (λ=1) and **3** (λ=0). Note that at λ=0 is not ligand **3** as the electrostatics have been perturbed to those of ligand 1. The free energy of the vdW perturbation of the bulk-solvent leg has been subtracted from the bound-leg surface to afford the relative binding free energy surface. From this relative binding free energy surface, the difference in free energy at the minima at λ=0 and 1, along with the equivalent energy of the electrostatic leg give the relative binding free energy of the two ligands.

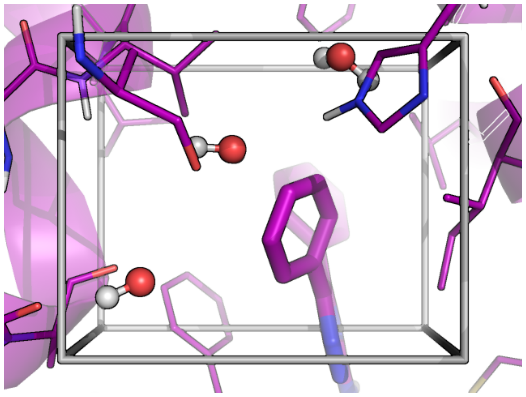


Figure 4.10: The initial placement of water molecules in the naïve solvation simulations. This naïve solvation is used with all ligands, but an unsubstituted scaffold is shown for clarity. The GCMC box is not included in the naïve simulations, but is shown in light grey for ease of comparison to Figure 4.12.

4.4.2 $\mathbf{A}_{2A}$

Naïve solvation

For comparison to other available methods, the A_{2A} simulations were also performed with a naïve solvation. The naïve simulation refers to the system being set up using ProtoMS set up tools, where the system is solvated based the available pocket volume and simulated with the NVT ensemble. The set-up places three water molecules within the GCMC region, shown in Figure 4.10. The water molecules will be sampled with solvent MC steps.

A GCMC region of A_{2A}that covers the ring A cavity has been used in the GCAP simulations. As before with SD, a free energy cycle between three A_{2A} ligands has been tested using the single- and surface-GCAP methodologies. With SD, a particular known water site of interest was chosen as the focused GCMC region. With A_{2A}, no water molecules are present in either of the two available crystal structures, although previous computational studies have highlighted hydration sites near rings A and B, that can vary between different ligands. A_{2A} ligands were treated as if no prior information were available, and a

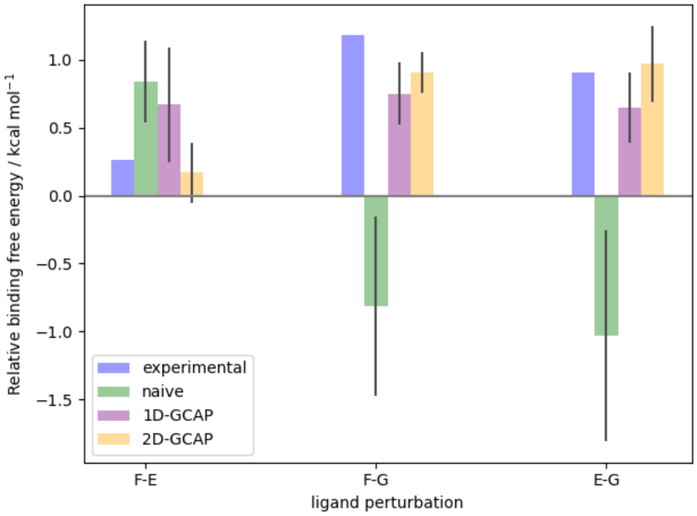


Figure 4.11: Relative binding free energies of A_{2A} ligand pairs in kcal·mol^{-1}. Results shown are experimental (blue), naïve (green) single-GCAP (purple) and surface-GCAP (orange). Error bars shown are standard errors from three repeats of each leg.

GCMC region was chosen to cover the active site cavity near ring A and the sites of alchemical perturbation, shown in Figure 4.3. The GCMC region is ∼8 times larger than for SD, and the number of water sites encapsulated in this region is higher than for the single water case of SD.

Nav̈e solvation (NVT ensemble) simulations do not correctly rank order the ligands, while both single-GCAP and surface-GCAP do, with surface-GCAP affording the best experimental agreement and smallest associated errors. The relative binding free energies of the pairs of ligands have been calculated using both single- and surface-GCAP, Figure 4.11. Both methods correctly rank order the ligands, with surface-GCAP results producing better experimental agreement, and smaller standard errors, for all legs. The thermodynamic cycles for these calculations are shown in Figure 4.13, where surface-GCAP also has better thermodynamic closure. In contrast to this, the relative free energies have also been calculated using a naïve solvation — where the water molecules have been placed in the system using default set up tools based on available pocket volume, and simulated with an NVT ensemble, Figure 4.11. The naïve solvation places three waters within the GCMC region, illustrated in Figure 4.10. Where the GCAP methods were able to rank order the ligands, the naïve calculations do not. The naïvely solvated simulations predict ligand **G** to be the most tightly bound, when experimentally it is the weakest binder. This shows the errors that can occur if relative binding free energies are calculated without proper consideration of the effect of the perturbation on the active site water network. The difference between the naïve cycle and the GCAP cycles is that no assumption has been made about the network of water molecules in the region of the ligand perturbation. The grand canonical ensemble allows the region to be dynamically solvated, and adaptively change as the ligand perturbs. This also allows us to predict the hydration sites for the various ligands, shown in Figure 4.12. As there are no available crystallographic water molecules, these cannot be experimentally validated.

GCMC cluster locations indicate the hydration sites of each of the three ligands considered with A_{2A}. The clustered water locations, and their occupancies are shown for all three ligands in Figure 4.12, labelled a - d, with hydrogen bonding contacts shown with yellow dashed lines, determined using pymol.[161] Water site a is deep in the pocket, and is stable and conserved for all three ligands. For ligands **E** and **F**, a water molecule b is able to bridge between their hydroxy group and the water site a. Water site b is 100% occupied for ligand **E**, but is only observed in ~40% of the simulation with ligand F. The destabilisation of this water molecule is likely due to the local methyl substitution on ligand F. With ligand **E**, as water site b is stable, a third site, water c is observed in ~30% of the simulation. This water is able to form two donating hydrogen bonds with backbone carbonyl groups. With ligand **G**, the substituted phenyl group of ligands **E** and **F** is replaced with a substituted pyridine group. The conserved water site a is observed, in addition to water site d, which was not observed with ligands **E** or **F**. Water site d bridges between two protein residues, rather than directly hydrogen bonding with the pyridine group.

Surface-GCAP results can decompose the energetic contribution the

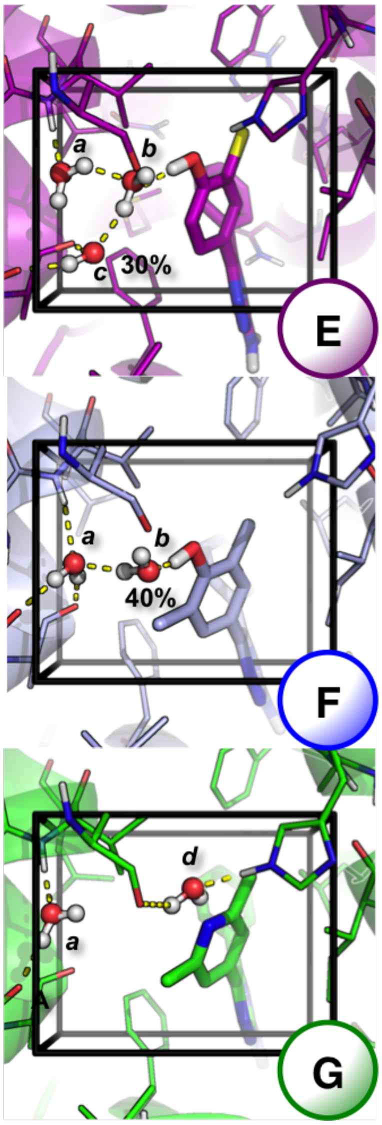


Figure 4.12: GCMC water locations top to bottom for ligands **E** (purple), **F** (light blue) and **G** (green) shown as sticks. Protein is represented as cartoon, with residues shown as lines. Carbon atoms are colored per ligand, with oxygen (red), nitrogen (dark blue), chlorine (yellow) and hydrogen (white). Any non-polar hydrogen atoms are removed for clarity. Hydrogen bonding (yellow dash) interactions are shown, determined using pymol.[161] GCMC hydration sites have been labelled $a - d$, with water occupancies labelled for waters that are present <95% of the simulation. Water locations have been calculated by clustering,[95] and a representative snapshot of the simulation is shown that represents the cluster centres.

water molecules provide to the relative affinities of ligands. As surface-GCAP is performed at a range of B values, it is possible to calculate additional free energy contributions, of the relative ligand binding free energy of the dry pocket, and the free energies of the water networks. From the surface-GCAP simulations, the binding free energy of the water network with each of the ligands can be calculated using the GCI Equation where λ is 0 or 1. This is equivalent to performing a GCMC titration simulation, with the addition of RE in λ as well as B. This has been calculated for each of the surfaces, and is shown as vertical legs in Figure 4.13. This shows that ligand **E** has the most tightly bound water network, followed by ligand **G**, while ligand **F** has the least tightly bound water network, despite being the highest affinity ligand. From surface-GCAP simulations, it is also possible to calculate the relative free energy of the ligands in a dry pocket by performing one-dimensional (1D) MBAR along the lowest B value, where the GCMC region has an average water occupancy of zero. This dry free energy cycle is shown in Figure 4.13, and while it is not intended to reproduce the experimental results, it can be useful — along with the water network binding free energies — for understanding from where the various energetic contributions arise.

4.4.3 Ligands F — G

The ligands **F** and **G** have the largest difference in affinity. As the relative hydration free energy of the two ligands is effectively zero, Table A.3, the difference in affinities arises from active site interactions. The GCAP simulations are able to show that the perturbation from ligand **F** to ligand **G** results in the loss of low-occupied water site b and the introduction of water site d as the hydroxyl group is removed. The water network with ligand **G** is 1.5 kcal·mol^{-1} more stable than with ligand F. This insight, provided by surface-GCAP, suggests that the high affinity of ligand **F** is predominantly due to the protein-ligand complementarity, rather than water stabilisation. This is illustrated by the dry leg affording a relative binding free energy of 3.0 kcal·mol^{-1}. The free energy surfaces for the vdW leg of this perturbation are shown in Figure 4.14. As before with SD, the surface shows that the minima in the free energy coincides with the B_{eq} value (-7.65).

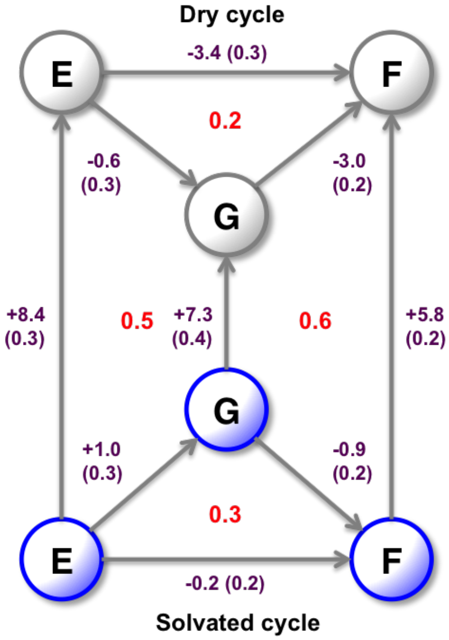


Figure 4.13: Relative binding free energies of the three A_{2A} ligands in kcal·mol^{-1}. All results are calculated from the surface-GCAP simulations. The dry cycle is calculated from using MBAR at B = -21.65, where the GCMC region is free of water molecules. The solvated cycle is calculated using MBAR on the whole surface, where the ligands will be correctly hydrated, Figure 4.12. The vertical legs are the free energy of the GCMC water networks, calculated using GCI at the λ end points of the surface. Standard errors are shown, calculated from three repeats for ligand perturbations and six repeats for water network calculations. Thermodynamic cycle closure is shown in red.

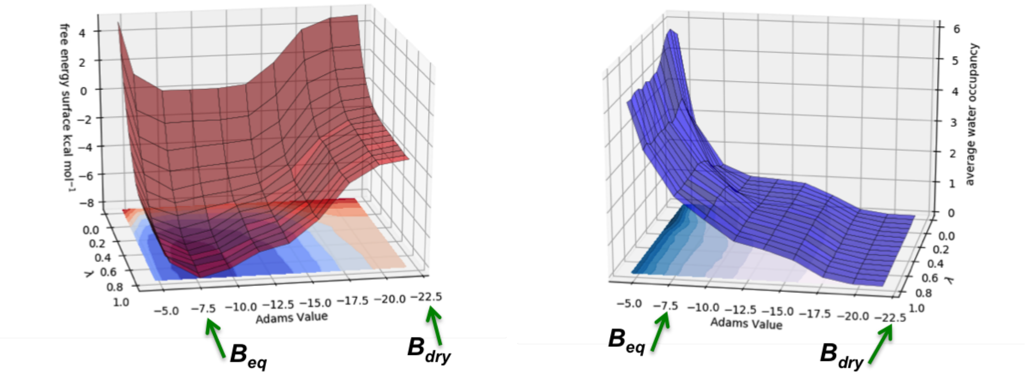


Figure 4.14: Free energy surface, and corresponding water occupancy from the vdW leg of the ligand **F** (λ=0) to ligand **G** (λ=1) perturbation. Note that this is not fully ligand **F** as the electrostatics have been perturbed to those of ligand **G**. The lowest free energy region of the surface is at B_{eq} (-7.65), where both ligands and any intermediate states will be dynamically hydrated. B_{eq} is indicated on the surface, along with B_{dry} the B value at which the dry cycles are calculated, where the average water occupancy is zero.

Where the GCMC region is under or over hydrated at lower or higher B values, the free energy of the system increases. Looking also at the water occupancy at B_{eq}, the water occupancy increases as λ changes from 0 to 1. This corresponds to the replacement of partially occupied site b with fully occupied d. The partially occupied nature of site b is easily simulated with the grand canonical ensemble, and would be challenging with a fixed N ensemble. At λ=0, where the system has perturbed electrostatics, but the vdW interactions are still of ligand **F**, the minimum in the free energy surface is broader than for λ=1 (ligand G). This suggests that ligand **F** (with its perturbed electrostatics) is stable with either **1** or **2** water molecules, whereas ligand **G** has favourable affinity only with two water molecules present. Little change is seen between ligand **F** and ligand **F** with the electrostatics perturbed, Figures A.4, A.5 and A.6.

4.4.4 Ligands E — F

The difference between ligands **E** and **F**, is the substitutions at the *meta* position. The alchemical perturbation that removes the methyl group close to water site *b* results in the stabilisation of the water site, and its occupancy increases from 40% to 100% across the alchemical pathway. This stabilises an additional water site, water *c*, which is in turn 30% occupied when ligand **E** is bound. The changes to sites *b* and *c* correspond to a 2.6 kcal·mol^{-1} favourable stabilization of the water network. The relative free energy of the ligands when the pocket is dry is +3.4 (0.3) kcal·mol^{-1} less favourable for ligand **E** than ligand **F**, which shows that the strong interactions of ligand **F** to the pocket directly, are mostly compensated by the increased stability of the water network with ligand **E**. While the relative free energy of the perturbation can be determined from the single-GCAP simulation, the surface-GCAP simulation in addition allows the binding free energies of the water network and the dry simulation to be calculated, which provides deeper understanding of the energetics and stability of the different systems. The **E** to **F** perturbation is also most improved when comparing to the experimental relative binding free energies by surface-GCAP, relative to single-GCAP. This may be because the difference in stability of the two ligands' water networks is the largest in the set.

4.4.5 Ligands E — G

As the perturbation from ligand **E** to ligand **G** involves both the addition and removal of functional groups, it has been performed in two steps, via the intermediate, where the C-OH group of ligand **E** has been perturbed to a N atom, but the meta groups are unchanged, shown in Figure 4.5. This perturbation from ligand **E** to **G** results in the loss of water sites *b* and *c*, and the introduction of water site *d*, corresponding to a loss in water network binding affinity of 1.1 kcal·mol^{-1}. The relative affinity of the dry leg, finds that ligand **G** is more tightly bound than **E** by 0.6 kcal·mol^{-1}; however, as the water network is able to better stabilise ligand **E**, ligand **E** is 1.0 kcal·mol^{-1} more tightly bound than ligand **G** when solvated.

Both GCAP methods perform well for A_{2A}, with the more computationally expensive surface-GCAP simulations outperforming the cheaper single-GCAP results. For the three ligands considered for A_{2A}, the GCAP methodologies are able to correctly reproduce the experimental relative binding free energies to within 1 kcal·mol^{-1} accuracy, while also determining the locations of the water molecules proximal to ring A. Attempting to calculate these relative free energies by naïvely solvating the system results in poor experimental agreement, with the lowest affinity ligand, ligand **G**, calculated as having the highest affinity. The starting locations of the water locations of the naïve simulations are shown in Figure 4.10 and indicate a water close to water site *d*, that is observed with ligand **G**, but not with ligands **E** or **F**. This coincidental similarity in the position of water *d* could explain why ligand **G** is predicted to be the most stable ligand in the naïve set of simulations. With ligands **E** and **F**, a water is not predicted in this location with the GCAP methods and is kinetically prevented from diffusing out of the pocket. Using the GCMC methodology, whereby water molecules are located on the fly throughout the simulation, means that there is no assumption of the number or location of water molecules within the region of interest. This allows for ligands with different active site water networks to be calculated directly. Although single-GCAP is computationally cheaper than surface-GCAP, the surface simulations provide smaller errors, better thermodynamic closure and better experimental agreement. In addition, simulating the whole surface through using a range of B values not only allow the stability of the water networks to be determined, by using GCI at a set λ value, but also the relative free energy of the ligands at a given level of hydration to be calculated, by using 1D MBAR along λ for a set B value. This information allows the energetic contributions from the water network to be decomposed. However, this additional information comes at computational cost, proportional to the number of additional B values simulated.

4.4.6 His$_{278}$ protonation

The protonation of an A_{2A}active site histidine residue has a significant impact on the relative ligand affinities. The active site histidine (His278)

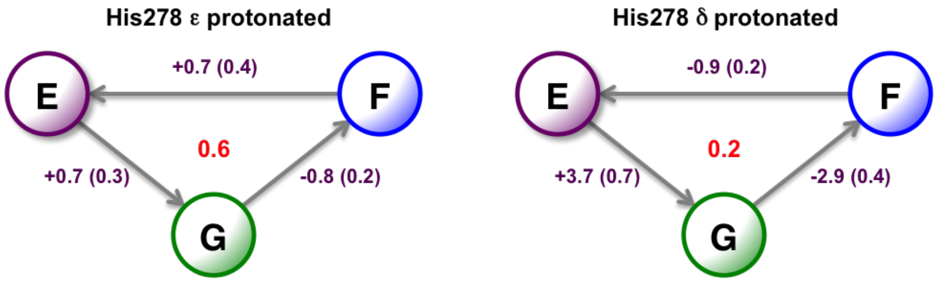


Figure 4.15: Relative ligand binding free energies for A_{2A} ligands with the two protonation states of active site residue, His278

was ϵ protonated by Maestro set up tools. As the residue is in close proximity to the ligand and the GCMC region, the simulations were repeated also with the δ protonation state. The results of this are shown in Figure 4.15. This changes the rank ordering of the ligands, with ligand **E** stabilised, and ligand **G** destabilised. The relative destabilisation of ligand **G** may be rationalised as it is the only complex that contains a His278-water hydrogen bond, Figure 4.12. As the ϵ protonated form was suggested in the set up, and has significantly better experimental agreement, this was shown in the main text. The δ protonated results show how sensitive results can be to choices made in the system set up — whether that be location of water molecules (demonstrated by the naïve results in the manuscript), and by the effect of protonation here. The alternate histidine rotamers have not been considered. Ideally, GCAP would be performed with a constant-pH protocol that would exchange the titratable active site residues during the simulation.

4.5 Conclusion

Issues arise in relative protein-ligand binding free energy calculations in cases where water molecules become trapped in the protein binding site. This can occur where the ligands considered have differing active site water networks. Conventional alchemical perturbation methods do not always cope with this situation, particularly in occluded pockets, where exchange with bulk water may be prevented within a feasible timescale due to kinetic barriers. GCMC has been developed to determine both active site water locations and water network free energies, all within a single series of simple to perform simulations.[54,157] In this paper, GCMC has been combined with MBAR to achieve dynamic adaptation of water networks with relative protein-ligand binding free energy calculations. Two protocols have been presented; low-cost single-GCAP that simulates only at B_{eq}, thereby ensuring equilibrium with bulk water, and high-precision surface-GCAP that simulates at a range of B values. Using surface-GCAP it is possible to calculate relative binding affinities between ligands at a chosen level of hydration, as well as isolate the contribution that the displacement, or rearrangement, of a water network has on the relative ligand binding affinity. Thus not only are robust, reproducible protein-ligand binding free energies produced, but the associated changes in water network in the binding site are observed. Moreover we have demonstrated the decomposition of the protein-ligand free energies into terms related directly to protein-ligand interactions and separately, to water stabilisation. We have shown with two protein ligand systems that this can produce experimentally consistent affinities, useful for drug design, and usefully rationalise Structure Activity Relationships. We anticipate that this methodology will prove a powerful tool in structure based drug design.

Chapter 5

Conclusions and future work

This thesis studies the role of water molecules in ligand binding, and illustrates how computational methods can be used in drug design. Grand canonical simulations of active site water molecules has been applied in various ways to enhance the sampling of occluded water molecules that can become trapped in conventional simulations due to limited achievable timescales and kinetic barriers. First, replica exchange has been introduced into GCMC simulations to improve sampling of titration simulations, with the motivation of improving the reliability of water network binding affinities. Second, GCMC has been applied to a large dataset to understand the success rate of the method for correctly identifying crystallographic water sites without *a priori* knowledge. Finally, GCMC has been introduced to conventional ligand free energy calculations, to allow for direct calculation of relative binding affinities for pairs of ligands with differing active site water networks.

Chapter 2 demonstrates that replica exchange between neighbouring B values is computationally cheap and improves the monotonicity of GCMC titration results. Improved monotonicity of results allows for better fitting of logistic functions to the data, allowing for the precise determination of the Gibbs binding free energy of networks of water molecules to an active site. Theoretical improvements to the GCI Equation affords results that are both as accurate and reliable as DD methods. The accuracy and reliability of GCMC has been demonstrated for two systems; BPTI and SD. DD requires the laborious set up of multiple simulations to decouple each water molecule of interest individually with the use of restraints of constraints. GCMC can calculate the binding free energy of multiple water molecules simultaneously, allowing for network contributions to be accounted for automatically. The theoretical improvements afforded the derivation of B_{eq}, whereby the system is in dynamic equilibrium with bulk water. Previously, a full GCMC titration plot was required to calculate the optimum water occupancy of a GCMC region, through minimising the Gibbs free energy; however, B_{eq} allows for this to be simulated using only a single B value. This does, however, lose the ability to calculate the water network binding free energy. Use of a single

B_{eq} allows for simulations to be computationally cheaper, as is demonstrated in Chapter 3. The results of Chapter 3 illustrate how binding free energies afforded by the new GCI Equation afford excellent agreement with values calculated using DD methods.

Chapter 3 illustrates the success of GCMC to reproduce crystallographic water locations for a dataset of 105 crystal structures. The dataset has been generated of complexes from FDA approved drug molecules, with care taken to prevent over representation of any particular drug or target. The focus on FDA-approved drug molecules ensures the usefulness of the test-set for pharmaceutical applications. Simulations were performed where the system was sampling (ligand, bulk water and protein with 15 Å of the ligand in addition to GCMC sampling) and where the system was fixed (only GCMC sampling). Fixed system sampling was shown to have higher success rates than sampling simulations, as the system does not dynamically move away from the starting conformation. The main focus of this Chapter, beyond determining the success rate of GCMC, was to study the variability that is possible in success rates based on different protocols. The success rate of GCMC was shown to vary with; simulation protocol, GCMC occupancy, classification of water molecules, and filtering the experimental data based on various quality metrics. The variation in the success rate makes apparent the need for methods to be applied to one dataset with a consistent analysis protocol. The intention is to make the FDA dataset readily available such that it can be used for future testing. To the best of our knowledge, this is the largest test of a simulation-based water placement method performed. At a distance cutoff of 1.0 Å, for GCMC clusters with occupancies $> 50\%$, for crystallographic water molecules that form at form two polar contacts to either protein or ligand, 59% and 72% of water molecules are found by sampling and fixed simulations respectively.

Chapter 4 presents GCAP — where relative ligand free energy simulations are performed in the grand canonical (constant μVT) ensemble, allowing for the ligand to perturb while the local active site water molecules can adapt accordingly. Performing relative binding free energy calculations in the NVT ensemble, where the number and location of active site hydration sites are assigned *a priori* can cause errors, both with cycle closures as for SD, and reduce experimental agreement for both systems tested. Two protocols were developed, one involving GCMC simulations performed at B_{eq} (single-GCAP), and another, more computationally expensive method, where a full titration is performed at each λ window (surface-GCAP). Single-GCAP is able to calculate the relative binding free energy with the same computational resources as a typical NVT, with only slightly longer simulation times (approximately 50% longer for single-GCAP as there are twice as many MC steps, but the insertion and deletion moves are computationally cheaper than conventional MC moves), for improved experimental agreement. Surface-GCAP involves simulating the perturbation at a range of λ windows and B values, resulting in two-dimensional free energy surfaces. Surface-GCAP is more computationally expensive, but it is able to afford the relative binding free energy between the pair of ligands in any of the simulated hydration states. Determining the relative free energy of the ligands with a dry pocket allows the free energy contributions to be decomposed to inspect the contribution that water molecules provide to GCMC, which can be useful for decision making in drug design projects. The binding free energy of the water networks with both ligands can be determined by using GCI on the λ end points. The GCAP method allows for more accurate binding free energies to be determined computationally, while fewer decisions are required during simulation set up, allowing the method to be more readily automated, which can reduce errors.

There are several avenues to pursue in future work. In Chapter 2, nearest-neighbour replica exchange was introduced to GCMC titration simulations. Future work could include all-pairs replica exchange to allow for faster mixing of states at different chemical potentials. Chapter 2 also demonstrated the ability to simulate only at B_{eq}, reducing the computational cost of the simulation, when only the locations of water molecules are of interest. If SAMS were applied to sampling B values in the proximity of B_{eq} then it may be possible to gain the benefits of replica exchange, without additional computational cost.

In Chapter 3, the success rate of GCMC was evaluated for a dataset of 105 protein-drug dataset. The dataset is of great value for future work. It is still unclear if water molecules should be targeted for displacement during drug design, and expanding this dataset to cover congeneric ligands and comparing the hydration patterns between them may be able to correlate SAR with the locations and affinities of active site water molecules. Various point-charge water models exist, and while TIP4P has been used throughout, the success rate of the data-set could be evaluated using a range of water force-fields, which would indicate which model — if any — is the most appropriate for modelling bound water molecules. This force-field comparison could be expanded to study different protein and ligand force fields for the application of identifying hydration sites.

MC sampling in simulations of the lengths performed herein are unlikely to differ significantly from the starting conformation, particularly when sections of the protein are held rigid. A hybrid methodology whereby GCMC simulations are intermixed with MD sampling would allow for better sampling of the full conformational space of the system, and would make the results more reliable. A GCAP-MD protocol would be suitable to model absolute ligand free energy calculations, whereby the ligand could be decoupled from the active site, and GCMC would control the appropriate hydration of the system throughout. MD sampling would out-perform a MC only methodology to capture the conformational changes in the residues of the active site on ligand unbinding.

To conclude, grand canonical sampling methods have been applied to a range of protein-ligand systems and its use has been demonstrated for the computational determination of positions and binding affinities of active site water molecules, and their influence on ligand binding.

Appendix

A.1 Optimising B-spacing

Replica exchange can be used for two non-exclusive benefits; improving the titration profile of GCMC results and enhancing sampling of a replica. The most efficient B-spacing for enhancing sampling of a given replica is found to be dependent on the water occupancy of the GCMC region, which can be approximated during the simulation set up. While RE can improve the reliability of calculated binding free energies using GCMC titration results, it can also be used to improve the sampling of the number of bound water molecules, N, at a given chemical potential. In the case of determining the equilibrium location of the water molecules, it is possible to simulate only at B_{eq}. Simulating only the equilibrium condition heavily reduces the computational resources required, as each B value requires a single processor. For the case of BPTI, if only the location of water molecules was required, simulating only B_{eq} would reduce the computational cost (and required disk space) by a factor of 32. However, as shown in Figure 2.4, considering a single B value in isolation, a range of N can be observed when each B value is run in isolation (i.e. without RE). A compromise between reducing the computational load, while benefitting from replica exchange, is to sample a few B values, at and around the equilibrium value. Before, the aim of RE was to smooth the titration curve, as measured by the Kendall τ score, Chapter 2. Now, as the GCI equation is not being used in Chapter 3, it is the sampling of N that is of primary interest.

For the optimal sampling of water occupancies at B_{eq}, successful RE moves should be as frequent as possible, but if the replicas are too closely spaced, then the exchange may be between repeats with the same N. One consequence of Equation 2.1 is that for ΔN=0, any attempted swap will always be accepted. While this is beneficial in terms of exchanging configurations across the B space, it does not enhance the sampling of water occupancies at a given B value. These simulations to determine only the equilibrium location of water molecules are presented in Chapter 3. In these simulations, good rates of RE are required to improve the sampling of N, but exchanges between replicas of the same occupancy are not as useful. For this reason, the ratio of accepted GCMC swaps that are between replicas of different N will be referred to as *useful* GCMC swaps. The optimal spacing of B values will be tested for a subset of 10 structures from the FDA dataset. Initially, a B-spacing of 1 was chosen, and the percentage of useful GCMC replica swaps has been plotted against the average equilibrium water occupancy for the system calculated from B_{eq}, shown in Figure A.1. It demonstrates that the greater the number of GCMC water molecules, the fewer B replica exchange moves are attempted. The acceptance rate drops off very rapidly, with fewer than 10% of moves accepted for systems with more than 10 GCMC water molecules. For nearly all the systems with more than 27 water molecules, the B value RE acceptance rate is 0%, which means that the additional replicas are trapped, and provide no information for equilibrium solvation. This plot makes sense; a B value that is $B_{eq}\pm$ 1, is going to either include or exclude water molecules that are either very weakly bound, favourably or unfavourably. If the number of water molecules in the GCMC is larger, then the variance in the occupancy will be larger, likely increasing ΔN. The larger N is for a system, the larger the variance in N, the larger ΔN is likely to be, and therefore the much less likely an attempted RE swap will be to be accepted.

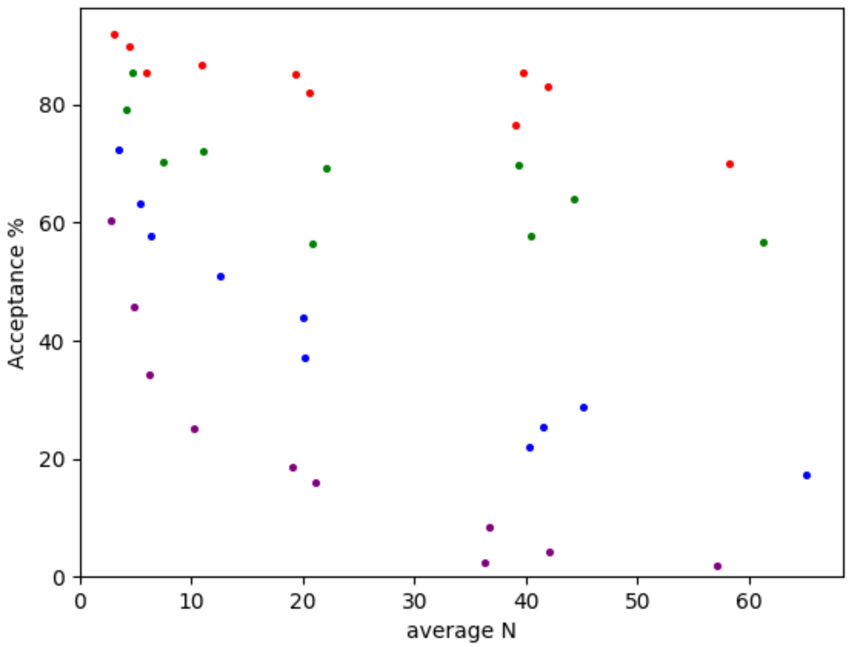


Figure A.1: Useful GCMC RE swaps for 10 systems of the FDA dataset, plotted against the average equilibrium water occupancy for different B-spacings. B-spacings are coloured as red=0.1, green=0.2, blue=0.5, purple=1.0

This suggests that the spacing between B values; when the enhanced sampling of N is of interest, should depend on the equilibrium average N. The equilibrium average N is a simulated observable, while the optimal B-spacing is something that we would like to set *a priori*. A naïve surrogate for average N would be the volume of the GCMC box, however Figure A.2 shows that there is little or weak correlation (R^2=0.48) between the GCMC volume and the average water occupancy. Two GCMC boxes of the same size may contain very different number of waters, depending if the ligand is surface binding, and most of the box is covering bulk water, or a very occluded, dry pocket, where very few water molecules will be located. An alternative metric was found to estimate N_{eq} of a GCMC region. Within the ProtoMS set up, ProtoMS will naïvely solvate the protein-ligand complex — retaining any crystallographic waters, and attempting to add additional waters at the density of bulk water, removing any that clash with protein or ligand atoms. When the GCMC box is generated, any waters within the GCMC box, crystallographic or otherwise are removed from the GCMC region. The number of water molecules removed from the system at this setup stage is a reasonable estimate of the equilibrium water occupancy of the system, shown in Figure A.2, with an R^2 of 0.92. This can be used in combination with Figure A.1 to indicate if a given B-spacing is appropriate. During set-up of a GCMC simulation, the number of water molecules cleared from the GCMC region can be used to estimate the expected value of N, using Figure A.2, right. This estimation for the expected average N can be used with Figure A.1 to approximate the expected RE acceptance rate for a range of B-spacings. An appropriate B-spacing can be chosen such that reasonable acceptance rates for RE is expected. For the research using the FDA dataset, Chapter 3, one B-spacing was preferred for the entire dataset. Based on Figure A.1, a B-spacing of 0.5 was chosen, as even the largest systems have a RE-acceptance rate of at least 15%. Using any closer spaced B values would both not represent simulating anything sufficiently dissimilar to B_{eq}, and would also be too finely spaced to reasonably adopt in a titration simulation.

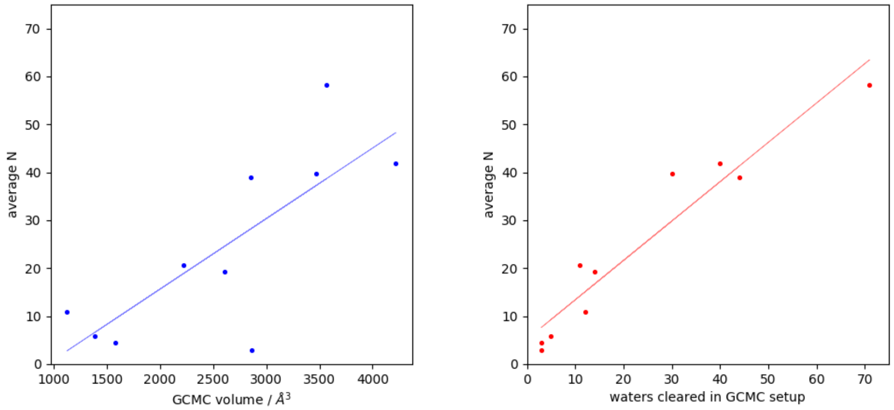


Figure A.2: Plots of the equilibrium average water occupancy of the region against the GCMC volume (blue,left) and the waters cleared from the GCMC region (red,right) for 10 systems. The number of waters cleared is a better predictor of the average water occupancy than the GCMC volume itself, with R^2 values of 0.92 and 0.48 respectively.

A.2 FDA Dataset

PDB	Organism	Protein	Ligand	Year	Res. /Å
1F9G	S. Ppneumoniae	Hyaluronate lyase	Vitamin C	2001	2.00
1FXV	E. coli	Penicillin acylase	Penicillin G	2001	2.25
1G5Y	H. sapiens	RXR alpha	Alitretinoin	2001	2.00
1GWR	H. sapiens	Estrogen receptor alpha	Estradiol	2002	2.40
1I1E	C. botulinum	Botulinum neurotoxin B	adriamycin	2001	2.50
1IE9	H. sapiens	Vitamin D receptor	Calcitriol	2001	1.40
1LHU	H. sapiens	Sex hormone-binding globulin	Estradiol	2002	1.80
1M2Z	H. sapiens	Glucocorticoid receptor	Dexamethasone	2003	2.50
1S19	H. sapiens	Vitamin D receptor	Calcipotriol	2004	2.10
1SQN	H. sapiens	Progesterone receptor	Norethisterone	2004	1.45
1SR7	H. sapiens	Progesterone receptor	Mometasone furoate	2004	1.46
1TUV	E. coli	YgiN	Menadione	2005	1.70
1UOU	H. sapiens	Thymidine phosphorylase	Tipiracil	2004	2.11
1X70	H. sapiens	Dipeptidyl peptidase IV	Sitagliptin	2005	2.10
1YI4	H. sapiens	PIM-1 kinase	Adenosine	2005	2.40
2A15	M. tuberculosis	RV0760	Nicotinamide	2005	1.68
2AA6	H. sapiens	Mineralocorticoid receptor (S810L)	Progesterone	2005	1.95
2AM9	H. sapiens	Androgen receptor	Testosterone	2006	1.64
2E5D	H. sapiens	Nicotinamide phosphoribosyltransferase	Nicotinamide	2007	2.00
2F9W	P. aeruginosa	Type III CoaA	Pantothenic acid	2006	1.90
2GQG	H. sapiens	ABL1	Dasatinib	2006	2.40
2HYY	H. sapiens	ABL	Imatinib	2007	2.40
2P16	H. sapiens	Coagulation Factor Xa	Apixaban	2007	2.30
2QK8	B. anthracis	DHFR	Methotrexate	2007	2.40
2RIN	S. meliloti	ABC-transporter choline binding protein	Acetylcholine	2008	1.80
2W26	H. sapiens	Factor Xa	Rivaroxaban	2008	2.08
2W9H	S. aureus	DHFR	Trimethoprim	2009	1.48
2WGJ	H. sapiens	c-Met	Crizotinib	2009	2.00
2XN3	H. sapiens	Thyroxine-binding globulin	Mefenamic acid	2011	2.09
2XRH	H. pylori	HP0721	Niacin	2011	1.50
2Y7J	H. sapiens	Phopsphorylase kinase, gamma 2	Sunitinib	2011	2.50
3APV	H. sapiens	Alpha-1-acid glycoprotein	Amitriptyline	2011	2.15
3APX	H. sapiens	Alpha1-acid glycoprotein	Chlorpromazine	2011	2.20
3AZZ	T. maritima	Laminarinase	Gluconolactone	2011	1.81
3B7E	I. A virus	Neuraminidase	Zanamivir	2008	1.45
3C7Q	H. sapiens	VEGFR2	Nintedanib	2008	2.10
3CSJ	H. sapiens	Glutathione S-transferase	Chlorambucil	2008	1.90
3D90	H. sapiens	Progesterone receptor	Levonorgestrel	2009	2.26
3EW2	R. etli	Rhizavidin	Biotin	2008	2.30
3EYG	H. sapiens	JAK1	Tofacitinib	2009	1.90
3F8F	L. lactis	LmrR	Daunorubicin	2008	2.20
3FL9	B. anthracis	DHFR	Trimethoprim	2009	2.40
3FUP	H. sapiens	JAK2	Tofacitinib	2009	2.40
3FUU	T. thermophilus	Methyltransferase	Adenosine	2009	1.53
3G0B	H. sapiens	Dipeptidyl peptidase IV	Alogliptin	2010	2.25
3G0E	H. sapiens	KIT kinase	Sunitinib	2009	1.60
3GN8	H. sapiens	AncGR2	Dexamethasone	2009	2.50
3I45	R. rubrum	Twin arginine translocation pathway signal protein	Niacin	2009	1.36
3L4W	H. sapiens	Maltase-glucoamylase	Miglitol	2010	2.00
3LXK	H. sapiens	JAK3	Tofacitinib	2010	2.00
3LXN	H. sapiens	TYK2	Tofacitinib	2010	2.50
3MYU	M. genitalium	MG289	Thiamine	2010	1.95
3OLL	H. sapiens	Estrogen receptor beta	Estradiol	2010	1.50
3QPS	C. jejuni	CmeR	Cholic acid	2011	2.35
3QT0	H. sapiens	PPARgamma	Mifepristone	2012	2.50
3RY2	S. avidinii	Streptavidin	Biotin	2011	0.95
3SG8	E. casseliflavus	Aminoglycoside-2”-phosphotransferase type 4a	Tobramycin	2011	1.80
3SG9	E. casseliflavus	Aminoglycoside-2”-phosphotransferase type 4a	Kanamycin	2011	2.15
3SZJ	S. denitrificans	Shwanavidin	Biotin	2012	1.45

PDB	**Organism**	**Protein**	**Ligand**	**Year**	**Res. /Å**
3TEG	H. sapiens	Phenylalanyl-tRNA synthetase	Levodopa	2011	2.20
3TI1	H. sapiens	CDK2	Sunitinib	2012	1.99
3U5J	H. sapiens	BRD4	Alprazolam	2011	1.60
3U5K	H. sapiens	BRD4	Midazolam	2011	1.80
3UE4	H. sapiens	ABL	Bosutinib	2012	2.42
3VHU	H. sapiens	Mineralocorticoid receptor	Spironolactone	2011	2.11
3VRI	H. sapiens	HLA class I histocompatibility antigen	Abacavir	2012	1.60
3VW1	S. enterica	RamR	Gentian Violet	2013	2.21
3WAR	H. sapiens	CK2a	Niacin	2013	1.04
4ASD	H. sapiens	VEGFR2	Sorafenib	2012	2.03
4BB2	H. sapiens	Corticosteroid-binding globulin	Progesterone	2012	2.48
4BBO	B. japonicum	Bradavidin	Biotin	2013	1.60
4DT8	E. casseliflavus	Aminoglycoside-2”-phosphotransferase type 4a	Adenosine	2012	2.15
4DVE	L. lactis	ECF-type ABC transporter	Biotin	2012	2.09
4E2J	H. sapiens	Glucocorticoid receptor 2	Mometasone furoate	2012	2.50
4EY6	H. sapiens	Acetylcholinesterase	Galantamine	2012	2.40
4G1Q	Immunodeficiency virus 1	Reverse transcriptase	Rilpivirine	2013	1.51
4GCP	E. coli	OmpF porin	Ampicillin	2012	1.98
4KS8	H. sapiens	PAK6	Sunitinib	2013	1.95
4LZR	H. sapiens	BRD4	Colchicine	2014	1.85
4MKC	H. sapiens	Anaplastic lymphoma kinase	Ceritinib	2014	2.01
4NMY	C. difficile	ABC transporter	Thiamine	2013	1.90
4O0S	H. sapiens	Aurora A	Adenosine	2014	2.50
4OAR	H. sapiens	Progesterone receptor	Ulipristal	2014	2.41
4P6W	H. sapiens	Glucocorticoid receptor	Mometasone furoate	2014	1.95
4P6X	H. sapiens	Glucocorticoid receptor	Hydrocortisone	2014	2.50
4QE6	H. sapiens	FXR	Chenodeoxycholic acid	2015	1.65
4QMN	H. sapiens	MST3	Bosutinib	2015	2.09
4QMS	H. sapiens	MST3	Dasatinib	2015	1.88
4QMZ	H. sapiens	MST3	Sunitinib	2015	1.88
4QRC	H. sapiens	FGFR4	Ponatinib	2014	1.90
4R38	E. litoralis	LOV protein	Riboflavin	2014	1.60
4RP9	E. coli	UlaA/SgaT	Vitamin C	2015	1.65
4RYA	A. vitis	ABC transporter	Mannitol	2014	1.50
4S0V	H. sapiens	OX2 orexin receptor	Suvorexant	2015	2.50
4TVJ	H. sapiens	PARP2	Olaparib	2015	2.10
4U0I	H. sapiens	KIT kinase	Ponatinib	2014	2.00
4U95	E. coli	AcrB	Minocycline	2014	2.00
4UDA	H. sapiens	Mineralocorticoid receptor	Dexamethasone	2015	2.03
4ZN7	H. sapiens	Estrogen receptor alpha	Diethylstilbestrol	2016	1.93
4ZOW	E. coli	MdfA	Riboflavin	2015	2.45
5EDL	B. subtilis	ECF transporter	Thiamine	2016	1.95
5G48	H. pylori	RORg	Diflunisal	2017	2.28
5I9X	H. sapiens	Ephrin A2	Bosutinib	2016	1.43
5P9I	H. sapiens	BTK	Ibrutinib	2017	1.11
5TE0	H. sapiens	AAK1	Nintedanib	2016	1.90
5UFS	H. sapiens	Glucocorticoid receptor 2	Triamcinolone acetonide	2017	2.12

Table A.1: Details of 105 complexes used in the FDA dataset for the validation of GCMC. All structures are of human, bacterial or viral origin, with an FDA approved drug molecule. The structures have been published since the year 2000, with a resolution of 2.5 Å or better.

A.3 Surface-GCAP results

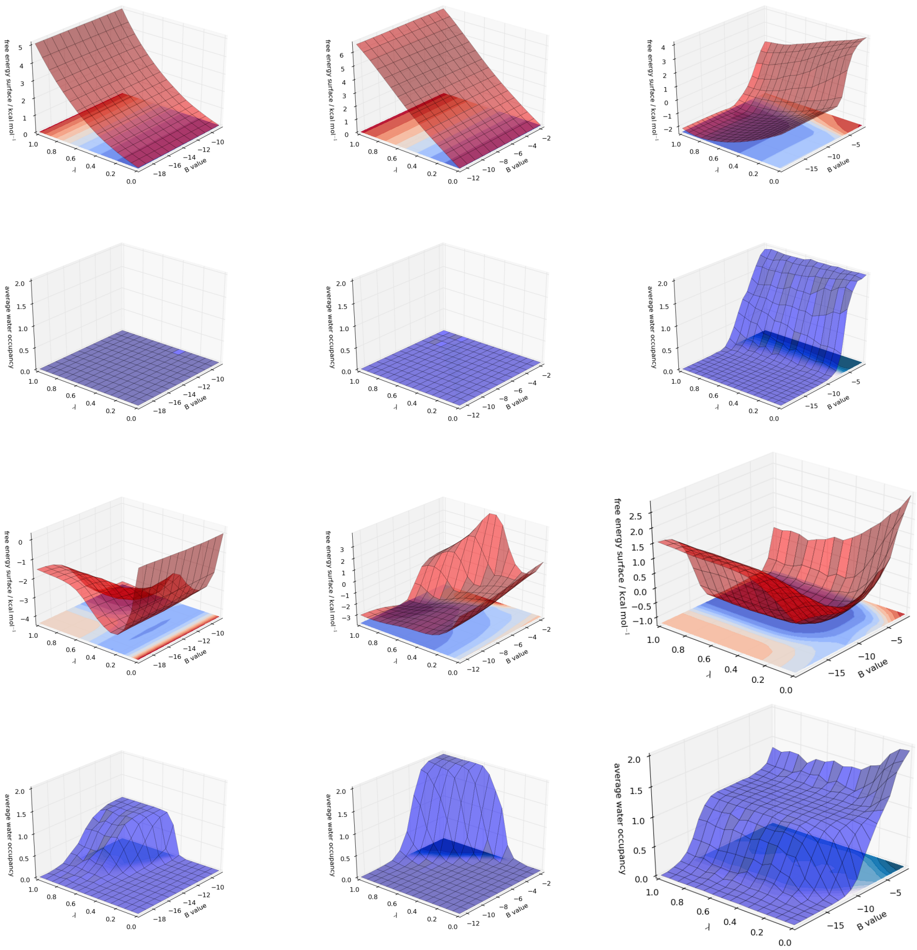


Figure A.3: surface-GCAP results for SD. Columns left to right: Ligands 1-2, 2-3, 3-1. Rows top to bottom: electrostatic surface, electrostatic solvation, van der Waals surface, van der Waals solvation.

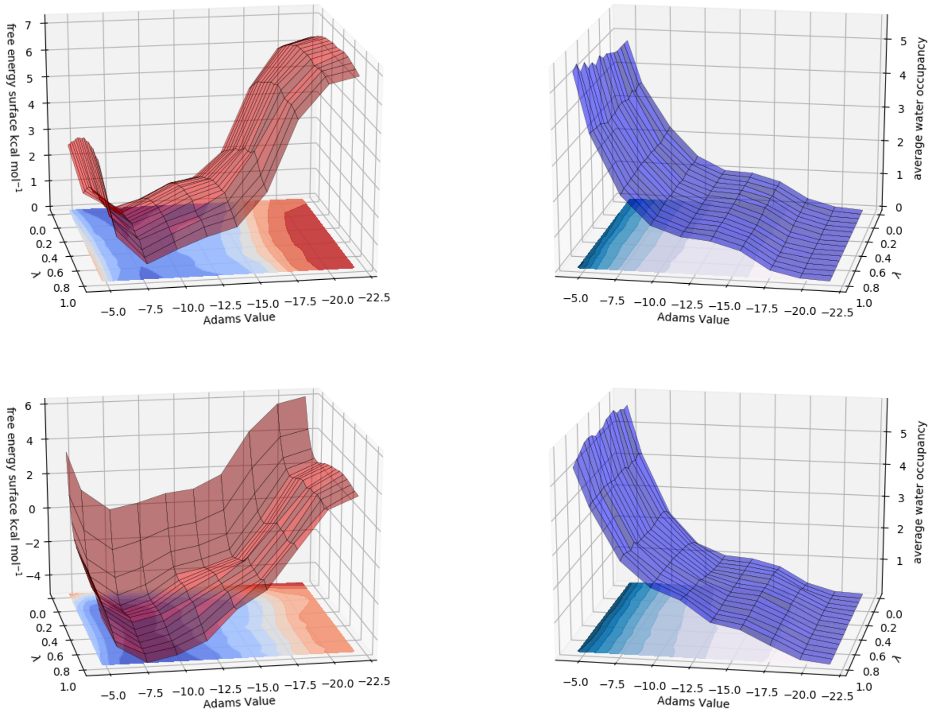


Figure A.4: surface-GCAP results for A_{2A}, ligands **E-F**. Surfaces shown are the free energy surface (left, red) and the average water occupancy (right, blue). Top row shows the electrostatics leg of the perturbation, and the bottom for shows the vdW perturbation.

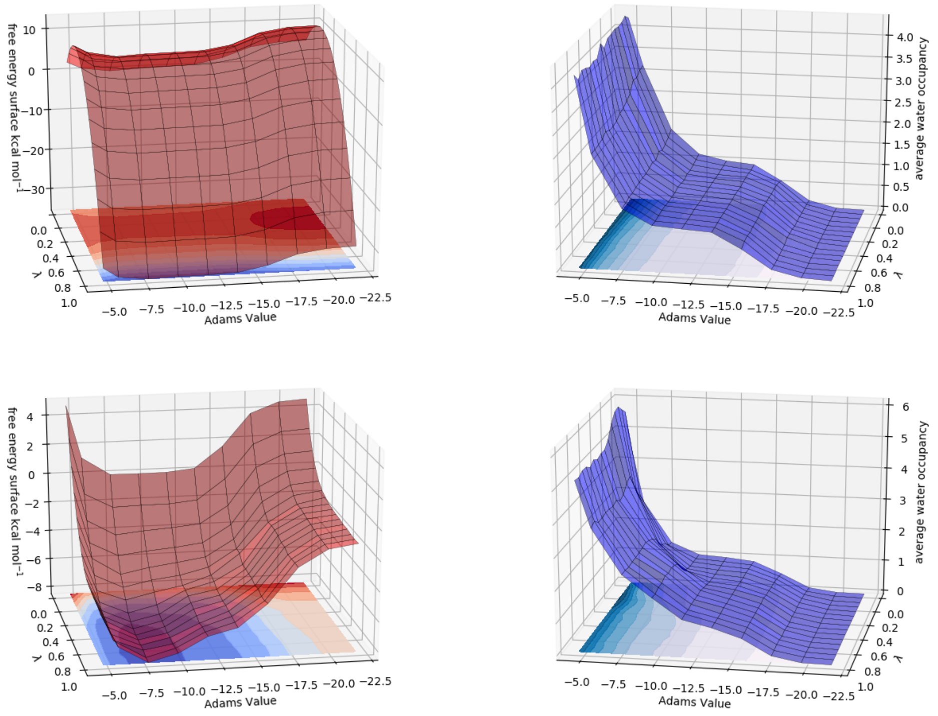


Figure A.5: surface-GCAP results for A_{2A}, ligands **F-G**. Surfaces shown are the free energy surface (left, red) and the average water occupancy (right, blue). Top row shows the electrostatics leg of the perturbation, and the bottom for shows the vdW perturbation.

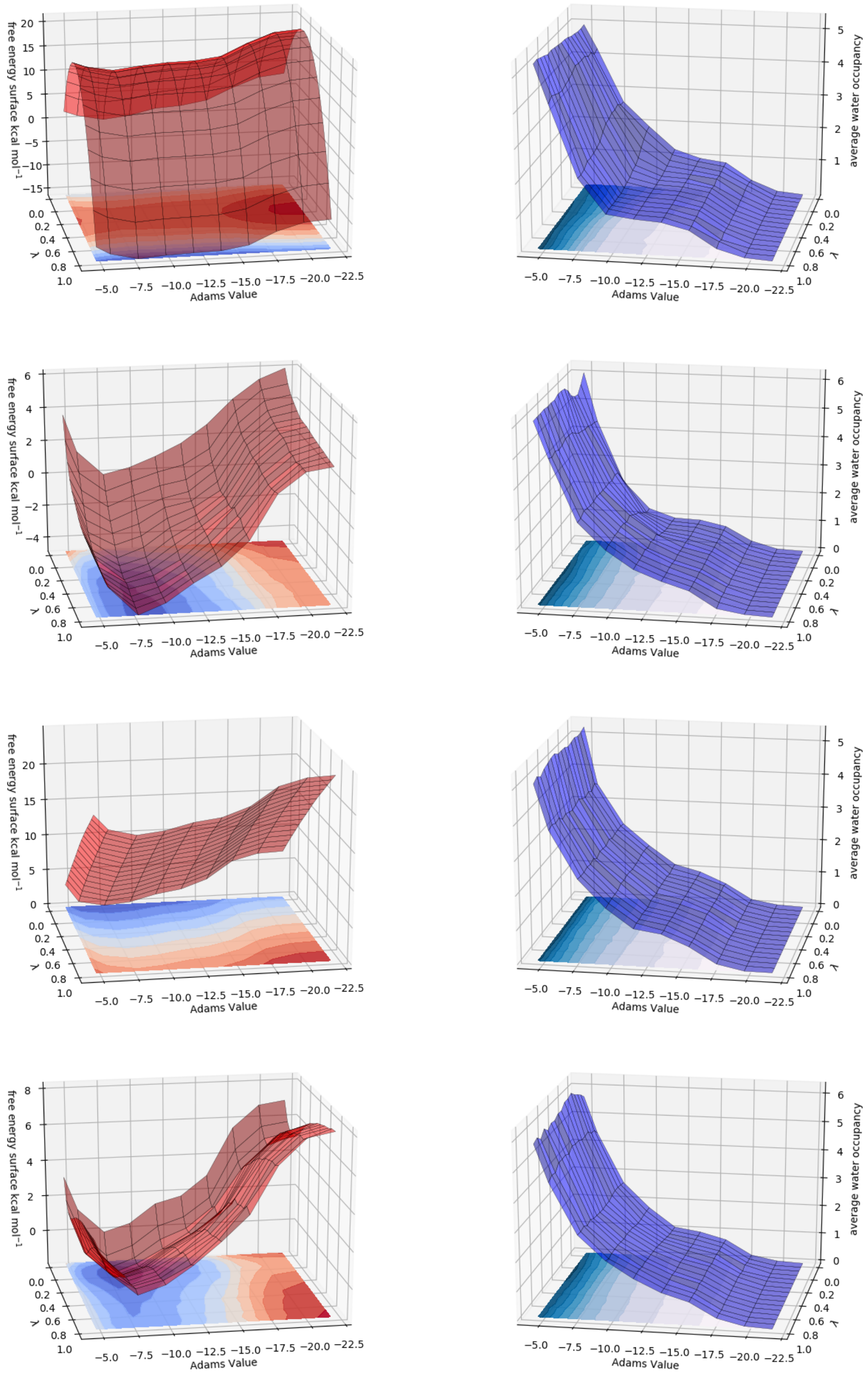


Figure A.6: surface-GCAP results for A_{2A}, ligands **E-G**. Surfaces shown are the free energy surface (left, red) and the average water occupancy (right, blue). The rows show; the electrostatics leg of the **E** - **M** perturbation, the vdW leg of the **E** - **M** perturbation, the electrostatics leg of the **G** - **M** perturbation, and the vdW leg of the **G** - **M** perturbation.

A.4 A_{2A} experimental binding affinities

The original publication of the A_{2A} ligand set considered herein[4] provides both K_i and K_D results for the set of ligands, measured using inhibition binding assays and SPR, respectively. As the free energy accuracy aimed for is typically 1 kcal·mol^{-1}in binding free energy calculations, we wanted to select a set of ligands where the relative free energies were within 1 kcal·mol^{-1}for demonstrating the GCAP methodology. The relative experimental free energies were considered as this reduces any possible systematic differences between the two measurements.

$$\Delta G_{K_D} = k_B T ln(K_D) \tag{5.1a}$$

$$\Delta G_{K_i} = -k_B T ln(K_i) \tag{5.1b}$$

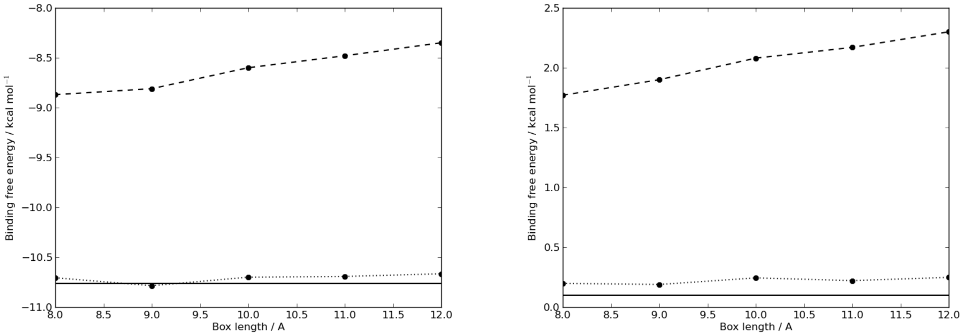
So the difference in the relative free energy for a pair of ligands (x and y), between the two methods, can be calculated from:

$$\Delta\Delta G(x-y)_{K_D} - \Delta\Delta G(x-y)_{K_i} = k_B T ln\left(\frac{K_D(x)}{K_D(y)}\right) + k_B T ln\left(\frac{K_i(x)}{K_i(y)}\right) \tag{5.2}$$

If the absolute value of Equation 5.2 is less than 1 kcal·mol^{-1}then the perturbation was considered for GCAP simulations. As crystal structures are only available for ligand G and **E**, any ligands where the binding mode was unclear, i.e. where either ring A or ring B was asymmetrically substituted, were excluded, as the ligand may bind in either orientation. This excludes ligands **B, J, K** and **L**. Ligand **E** is asymmetrically substituted, but the binding mode is available from the crystal structure. Of the 8 remaining ligands, only 7 have published data for both K_i and K_D. This results in 42 possible pairs of ligands. Of the 42 pairs, only 8 pairs satisfied the requirement that Equation 5.2 was less than ± 1 kcal·mol^{-1};EF, EG, EH, EI, FG, GH, GI and HI. Of these, the ligands **E**, **F** and **G** were chosen as both ligands **E** and **G** have crystal structures available, and the differences in the ligand seem significant enough to displace or disrupt active site water molecules.

system	GCI	DD
one *a*	-5.50 (0.02)	-5.52 (0.03)
one *b*	-5.17 (0.01)	-5.24 (0.02)
three *a*	+5.53 (0.02)	+5.47 (0.12)
three *b*	-5.30 (0.01)	-5.37 (0.07)

Table A.2: Binding free energy for each individual water for SD, calculated by both GCI and DD, with the correction applied. Errors shown are standard deviation over four repeats for the decoupling, and over five different sized GCMC regions for the GCMC results.



(a) waters *a* and *b*, ligand one

(b) waters *a* and *b*, ligand three

Figure A.7: Binding free energy of two-water networks with SD-ligand complex. Dashed line - GCMC results (Equation 1.61), dotted line - GCMC result including volume correction (Equation 2.7), Solid line - decoupling result.

A.5 A_{2A}hydration free energies

Table A.3: Relative free energy perturbations for ligands in the gas phase, and bulk solvent phase. ΔG_{hyd} is the relative free energy of hydration of the two ligands, calculated from ΔG_{sol} - ΔG_{gas}. ΔG_{sol} is used to calculate ΔG_{bind}. All energies are in kcal·mol^{-1}. Energies and standard errors for SD are calculated using MBAR from four repeats, and A$_{2A}$ from three.

Perturbation	ΔG_{gas}	ΔG_{sol}	ΔG_{hyd}
2 to **1**	-100.9 (0.0)	-101.1 (0.1)	-0.1 (0.1)
2 to **3**	-12.5 (0.0)	-11.5 (0.1)	1.0 (0.1)
3 to **1**	-90.1 (0.0)	-91.3 (0.1)	-1.2 (0.1)
F to **E**	-5.7 (0.0)	-4.9 (0.2)	0.8 (0.2)
F to **G**	-43.9 (0.0)	-43.7 (0.1)	0.2 (0.1)
E to **G**	-38.7 (0.1)	-39.0 (0.2)	-0.4 (0.2)

Bibliography

[1] S. H. Sleigh, P. R. Seavers, A. J. Wilkinson, J. E. Ladbury and J. R. Tame, *Journal of Molecular Biology*, 1999, **291**, 393–415.

[2] T. Lundqvist, J. Rice, C. N. Hodge, G. S. Basarab, J. Pierce and Y. Lindqvist, *Structure*, 1994, **2**, 937–944.

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

[4] M. Congreve, S. P. Andrews, A. S. Doré, K. Hollenstein, E. Hurrell, C. J. Langmead, J. S. Mason, I. W. Ng, B. Tehan, A. Zhukov, M. Weir and F. H. Marshall, *Journal of Medicinal Chemistry*, 2012, **55**, 1898–1903.

[5] J. Michel, J. Tirado-Rives and W. L. Jorgensen, *Journal of the American Chemical Society*, 2009, **131**, 15403–15411.

[6] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *Journal of Chemical Theory and Computation*, 2015, **11**, 3696–3713.

[7] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157–1174.

[8] G. Kaminski and W. L. Jorgensen, *The Journal of Physical Chemistry*, 1996, **100**, 18010–18013.

[9] R. A. Buckingham, *Proceedings of the Royal Society*, 1938, **168**, 264–283.

[10] S. E. Hill and W. Osterhout, *The Journal of General Physiology*, 1938, **21**, 541–556.

[11] A. R. Leach, *Molecular Modelling: Principles and Applications*, Pearson education, 2001.

[12] R. Zwanzig, *The Journal of Chemical Physics*, 1954, **22**, 1420.

[13] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, 1996.

[14] L. Verlet, *Physical Review*, 1967, **159**, 98–103.

[15] R. W. Hockney, *Methods of Computational Physics*, 1970, **9**, 136.

[16] D. Beeman, *Journal of Computational Physics*, 1976, **20**, 130–139.

[17] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford university press, 2017.

[18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *The Journal of Chemical Physics*, 1953, **21**, 1087–1092.

[19] L. Martino and J. Read, *Computational Statistics*, 2013, **28**, 2797–2823.

[20] J. G. Kirkwood, *The Journal of Chemical Physics*, 1935, **3**, 300–313.

[21] C. H. Bennett, *Journal of Computational Physics*, 1976, **22**, 245–268.

[22] M. R. Shirts and J. D. Chodera, *The Journal of Chemical Physics*, 2008, **129**, 124105+.

[23] C. J. Woods, M. A. King and J. W. Essex, *Replica-Exchange-Based Free-Energy Methods*, Springer Berlin Heidelberg, 2006, vol. 49, pp. 251–259.

[24] Y. Sugita, A. Kitao and Y. Okamoto, *The Journal of Chemical Physics*, 2000, **113**, 6042–6051.

[25] H. Fukunishi, O. Watanabe and S. Takada, *The Journal of Chemical Physics*, 2002, **116**, 9058–9067.

[26] Y. Sugita and Y. Okamoto, *Chemical Physics Letters*, 1999, **314**, 141–151.

[27] A. M. Baptista, P. J. Martel and S. B. Petersen, *Proteins*, 1997, **27**, 523–544.

[28] J. D. Chodera, W. C. Swope, J. W. Pitera, C. Seok and K. A. Dill, *Journal of Chemical Theory and Computation*, 2007, **3**, 26–41.

[29] P. Kollman, *Chemical Reviews*, 1993, **93**, 2395–2417.

[30] Y. Sun, D. Spellmeyer, D. A. Pearlman and P. Kollman, *Journal of the American Chemical Society*, 1992, **114**, 6798–6801.

[31] P. A. Bash, U. C. Singh, R. Langridge and P. A. Kollman, *Science*, 1987, **236**, 564–568.

[32] K. M. Merz, M. A. Murcko and P. A. Kollman, *Journal of the American Chemical Society*, 1991, **113**, 4484–4490.

[33] T. Steinbrecher, I. Joung and D. A. Case, *Journal of Computational Chemistry*, 2011, **32**, 3253–3263.

[34] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber and W. F. van Gunsteren, *Chemical Physics Letters*, 1994, **222**, 529–539.

[35] T. Steinbrecher, D. L. Mobley and D. A. Case, *The Journal of Chemical Physics*, 2007, **127**, 214108.

[36] B. O. Brandsdal, F. Österberg, M. Almlöf, I. Feierberg, V. B. Luzhkov and J. Åqvist, in *Free Energy Calculations and Ligand Binding*, Elsevier, 2003, vol. 66, pp. 123–158.

[37] T. A. Özal, C. Peter, B. Hess and N. F. A. van der Vegt, *Macromolecules*, 2008, **41**, 5055–5061.

[38] J. W. Pitera and W. F. van Gunsteren, *Journal of Physical Chemistry B*, 2001, **105**, 11264–11274.

[39] A. Rizzi, P. B. Grinaway, D. L. Parton, M. R. Shirts, K. Wang, P. Eastman, M. Friedrichs, V. S. Pande, K. Branson, D. L. Mobley and J. D. Chodera, *yank*, getyank.org, 2018.

[40] C. J. Woods, M. Malaisree, S. Hannongbua and A. J. Mulholland, *The Journal of Chemical Physics*, 2011, **134**, 054114+.

[41] C. J. Woods, M. Malaisree, J. Michel, B. Long, S. McIntosh-Smith and A. J. Mulholland, *Faraday Discussions*, 2014, **169**, 477–499.

[42] T. L. Hill, *The Journal of Chemical Physics*, 1955, **23**, 623–636.

[43] J. Michel, M. L. Verdonk and J. W. Essex, *Journal of Chemical Theory and Computation*, 2007, **3**, 1645–1655.

[44] J. Michel and J. W. Essex, *Journal of Medicinal Chemistry*, 2008, **51**, 6654–6664.

[45] S. Riniker, C. D. Christ, N. Hansen, A. E. Mark, P. C. Nair and W. F. van Gunsteren, *The Journal of Chemical Physics*, 2011, **135**, 07B604.

[46] S. Banba and C. L. Brooks III, *The Journal of Chemical Physics*, 2000, **113**, 3423–3433.

[47] S. Banba, Z. Guo and C. L. Brooks, *The Journal of Physical Chemistry B*, 2000, **104**, 6903–6910.

[48] M. K. Gilson, J. A. Given, B. L. Bush and J. A. McCammon, *Biophysical Journal*, 1997, **72**, 1047–1069.

[49] J. Hermans and S. Shankar, *Israel Journal of Chemistry*, 1986, **27**, 225–227.

[50] B. Roux, M. Nina, R. Pomès and J. C. Smith, *Biophysical Journal*, 1996, **71**, 670–681.

[51] M. K. Gilson and K. K. Irikura, *Journal of Physical Chemistry B*, 2010, **114**, 16304–16317.

[52] J. E. Mayer and M. G. Mayer, *Statistical Mechanics*, John Wiley & Sons, 1940.

[53] D. L. Mobley, J. D. Chodera and K. A. Dill, *The Journal of Chemical Physics*, 2006, **125**, 084902+.

[54] G. A. Ross, M. S. Bodnarchuk and J. W. Essex, *Journal of the American Chemical Society*, 2015, **137**, 14930–14943.

[55] R. H. Swendsen and J.-S. Wang, *Physical Review Letters*, 1986, **57**, 2607–2609.

[56] C. J. Woods, J. W. Essex and M. A. King, *The Journal of Physical Chemistry B*, 2003, **107**, 13703–13710.

[57] Y. Okamoto, *Journal of Molecular Graphics and Modelling*, 2004, **22**, 425–439.

[58] S. Park and V. S. Pande, *Physical Review E*, 2007, **76**, 016703.

[59] Y. Sindhikara, D. Meng and A. E. Roitberg, *The Journal of Chemical Physics*, 2008, **128**, 024103+.

[60] P. Brenner, C. R. Sweet, D. VonHandorf and J. A. Izaguirre, *The Journal of Chemical Physics*, 2007, **126**, 074103+.

[61] J. D. Chodera and M. R. Shirts, *The Journal of Chemical Physics*, 2011, **135**, 194110+.

[62] M. Levitt and B. H. Park, *Structure*, 1993, **1**, 223–226.

[63] R. Horst, G. Wider, J. Fiaux, E. B. Bertelsen, A. L. Horwich and K. Wüthrich, *Proceedings of the National Academy of Sciences*, 2006, **103**, 15445–15450.

[64] J. A. Ernst, R. T. Clubb, H. X. Zhou, A. M. Gronenborn and G. M. Clore, *Science*, 1995, **267**, 1813–1817.

[65] E. Persson and B. Halle, *Proceedings of the National Academy of Sciences*, 2008, **105**, 6266–6271.

[66] C. Barillari, J. Taylor, R. Viner and J. W. Essex, *Journal of the American Chemical Society*, 2007, **129**, 2577–2587.

[67] T. Young, R. Abel, B. Kim, B. J. Berne and R. A. Friesner, *Proceedings of the National Academy of Sciences*, 2007, **104**, 808–813.

[68] R. Abel, T. Young, R. Farid, B. J. Berne and R. A. Friesner, *Journal of the American Chemical Society*, 2008, **130**, 2817–2831.

[69] T. Lazaridis, *Journal of Physical Chemistry B*, 1998, **102**, 3531–3541.

[70] M. S. Bodnarchuk, *Drug Discovery Today*, 2016, **21**, 1139–1146.

[71] C. N. Nguyen, K. Young, T. Kurtzman and M. K. Gilson, *The Journal of Chemical Physics*, 2012, **137**, 044101+.

[72] T. Beuming, R. Farid and W. Sherman, *Protein Science*, 2009, **18**, 1609–1619.

[73] C. Higgs, T. Beuming and W. Sherman, *ACS Medicinal Chemistry Letters*, 2010, **1**, 160–164.

[74] C. N. Nguyen, A. Cruz, M. K. Gilson and T. Kurtzman, *Journal of Chemical Theory and Computation*, 2014, **10**, 2769–2780.

[75] J. Michel, J. Tirado-Rives and W. L. Jorgensen, *Journal of Physical Chemistry B*, 2009, **113**, 13337–13346.

[76] E. D. López, J. P. Arcon, D. F. Gauto, A. A. Petruk, C. P. Modenutti, V. G. Dumas, M. A. Marti and A. G. Turjanski, *Bioinformatics*, 2015, **31**, 3697–3699.

[77] K. Haider, A. Cruz, S. Ramsey, M. K. Gilson and T. Kurtzman, *Journal of Chemical Theory and Computation*, 2018, **14**, 418–425.

[78] W. R. Pitt and J. M. Goodfellow, *Protein Engineering, Design and Selection*, 1991, **4**, 531–537.

[79] W. R. Pitt, J. Murray-Rust and J. M. Goodfellow, *Journal of Computational Chemistry*, 1993, **14**, 1007–1018.

[80] J. Konc and D. Janežič, *Bioinformatics*, 2010, **26**, 1160–1168.

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

[82] E. Nittinger, F. Flachsenberg, S. Bietz, G. Lange, R. Klein and M. Rarey, *Journal of Chemical Information and Modeling*, 2018, **58**, 1625–1637.

[83] B. Widom, *The Journal of Chemical Physics*, 1963, **39**, 2808–2812.

[84] D. J. Adams, *Molecular Physics*, 1974, **28**, 1241–1252.

[85] D. J. Adams, *Molecular Physics*, 1975, **29**, 307–311.

[86] F. Guarnieri and M. Mezei, *Journal of the American Chemical Society*, 1996, **118**, 8493–8494.

[87] M. Clark, F. Guarnieri, I. Shkurko and J. Wiseman, *Journal of Chemical Information and Modeling*, 2006, **46**, 231–242.

[88] S. Vaitheeswaran, J. C. Rasaiah and G. Hummer, *The Journal of Chemical Physics*, 2004, **121**, 7955–7965.

[89] S. Vaitheeswaran, H. Yin, J. C. Rasaiah and G. Hummer, *Proceedings of the National Academy of Sciences*, 2004, **101**, 17002–17005.

[90] M. Mezei, *Molecular Physics*, 1980, **40**, 901–906.

[91] M. Mezei, *Molecular Physics*, 1987, **61**, 565–582.

[92] H.-J. Woo, A. R. Dinner and B. Roux, *The Journal of Chemical Physics*, 2004, **121**, 6392–6400.

[93] M. Clark, S. Meshkat and J. S. Wiseman, *Journal of Chemical Information and Modeling*, 2009, **49**, 934–943.

[94] A. Ben-Naim and Y. Marcus, *The Journal of Chemical Physics*, 1984, **81**, 2016–2027.

[95] J. E. Group, *ProtoMS*, www.protoms.org, 2018.

[96] J. A. Barker and D. Henderson, *The Journal of Chemical Physics*, 1967, **47**, 4714–4721.

[97] G. E. Norman and V. S. Filinov, *High Temperature*, 1969, **7**, 216–222.

[98] A. McPherson and J. A. Gavira, *Acta Crystallographica Section F*, 2014, **70**, 2–20.

[99] A. Krogh, B. Larsson, G. von Heijne and E. L. L. Sonnhammer, *Journal of Molecular Biology*, 2001, **305**, 567–580.

[100] E. P. Carpenter, K. Beis, A. D. Cameron and S. Iwata, *Current Opinion in Structural Biology*, 2008, **18**, 581–586.

[101] A. Wlodawer, W. Minor, Z. Dauter and M. Jaskolski, *The FEBS Journal*, 2008, **275**, 1–21.

[102] M. Cymborowski, M. Klimecka, M. Chruszcz, M. Zimmerman, I. Shumilin, D. Borek, K. Lazarski, A. Joachimiak, Z. Otwinowski, W. Anderson and W. Minor, *Journal of Structural and Functional Genomics*, 2010, **11**, 211–221.

[103] L. Zhou and Q. Liu, *The Journal of Physical Chemistry B*, 2014, **118**, 4069–4079.

[104] E. A. Merritt, *Acta Crystallographica Section D: Biological Crystallography*, 2012, **68**, 468–477.

[105] N. V. Nucci, M. S. Pometun and A. J. Wand, *Nature Structural and Molecular Biology*, 2011, **18**, 245–249.

[106] R. P. Joosten, F. Long, G. N. Murshudov and A. Perrakis, *International Union of Crystallography Journal*, 2014, **1**, 213–220.

[107] N. Research, *PDB_REDO*, https://pdb-redo.eu/, 2018.

[108] T. A. Jones, J. Y. Zou, S. W. Cowan and M. Kjeldgaard, *Acta Crystallographica Section A*, 1991, **47**, 110–119.

[109] T. A. Jones and M. Kjeldgaard, in *Electron-density Map Interpretation*, Elsevier, 1997, vol. 277, pp. 173–208.

[110] I. J. Tickle, *Acta Crystallographica Section D*, 2012, **68**, 454–467.

[111] E. Nittinger, N. Schneider, G. Lange and M. Rarey, *Journal of Chemical Information and Modeling*, 2015, **55**, 771–783.

[112] A. Meyder, E. Nittinger, G. Lange, R. Klein and M. Rarey, *Journal of Chemical Information and Modeling*, 2017, **57**, 2437–2447.

[113] T. D. Pollard, *Molecular Biology of the Cell*, 2010, **21**, 4061–4067.

[114] A. B. Ghisaidoobe and S. J. Chung, *International Journal of Molecular Sciences*, 2014, **15**, 22518–22538.

[115] L. Tian, S. Liu, S. Wang and L. Wang, *Scientific Reports*, 2016, **6**, 1–11.

[116] P. A. Van Der Merwe, *Protein-ligand Interactions: Hydrodynamics and Calorimetry*, 2001, **1**, 137–170.

[117] M. M. Pierce, C. S. Raman and B. T. Nall, *Methods*, 1999, **19**, 213–221.

[118] W. H. Ward and G. A. Holdgate, in *Progress in Medicinal Chemistry*, Elsevier, 2001, vol. 38, pp. 309–376.

[119] G. E. P. Box, in *Robustness in the Strategy of Scientific Model Building*, Elsevier, 1979, pp. 201–236.

[120] A. G. Smart, *The war over super-cooled water*, https://physicstoday.scitation.org/do/10.1063/PT.6.1.20180822a/full/, 2018.

[121] S. Hykes, *docker*, www.docker.com, 2010.

[122] D. Horinek, S. I. Mamatkulov and R. R. Netz, *The Journal of Chemical Physics*, 2009, **130**, 124507+.

[123] D. Mobley and C. Caitlin, *bioRxiv*, 2018.

[124] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon, *The Journal of Physical Chemistry B*, 2010, **114**, 2549–2564.

[125] A. Laio and M. Parrinello, *Proceedings of the National Academy of Sciences*, 2002, **99**, 12562–12566.

[126] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chemical Reviews*, 2016, **116**, 7898–7936.

[127] D. J. Earl and M. W. Deem, *Physical Chemistry Chemical Physics*, 2005, **7**, 3910–3916.

[128] S. G. Itoh, A. Damjanović and B. R. Brooks, *Proteins*, 2011, **79**, 3420–3436.

[129] G. Moore, *Electronics*, 1965, **38**, 114–117.

[130] S. Loukatou, L. Papageorgiou, P. Fakourelis, A. Filntisi, E. Polychronidou, I. Bassis, V. Megalooikonomou, W. Makałowski, D. Vlachakis and S. Kossida, *Journal of Molecular Biochemistry*, 2014, **3**, 64–71.

[131] Y. Duan and P. A. Kollman, *Science*, 1998, **282**, 740–744.

[132] D. E. Shaw, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Lerardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey,

M. M. Deneroff, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S. C. Wang, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson and K. J. Bowers, *Communications of the ACM*, 2008, **51**, 91–97.

[133] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror and D. E. Shaw, *Current Opinion in Structural Biology*, 2009, **19**, 120–127.

[134] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger and D. E. Shaw, *Journal of the American Chemical Society*, 2011, **133**, 9181–9183.

[135] W. Chen, B. H. Morrow, C. Shi and J. K. Shen, *Molecular Simulation*, 2014, **40**, 830–838.

[136] G. A. Ross, A. S. Rustenburg, P. B. Grinaway, J. Fass and J. D. Chodera, *The Journal of Physical Chemistry B*, 2018, **122**, 5466–5486.

[137] M. P. Blakeley, P. Langan, N. Niimura and A. Podjarny, *Current Opinion in Structural Biology*, 2008, **18**, 593–600.

[138] Y. Liu, S. Gonen, T. Gonen and T. O. Yeates, *Proceedings of the National Academy of Sciences*, 2018, **13**, 201718825+.

[139] C. J. Woods, J. W. Essex and M. A. King, *The Journal of Physical Chemistry B*, 2003, **107**, 13703–13710.

[140] V. B. Chen, W. B. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, *Acta Crystallographica Section D*, 2010, **66**, 12–21.

[141] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926–935.

[142] G. L. Warren, T. D. Do, B. P. Kelley, A. Nicholls and S. D. Warren, *Drug Discovery Today*, 2012, **17**, 1270–1281.

[143] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson and C. W. Murray, *Journal of Medicinal Chemistry*, 2007, **50**, 726–741.

[144] G. Warren, *Modeling Water: Real or an illusion?*, https://www.eyesopen.com/news/webinar-2018-modeling-water-real-or-illusion, 2018.

[145] Schrödinger, *Release 2018-1: Maestro, Schrödinger*, 2018.

[146] U. of Hamburg, *proteins plus*, https://proteins.plus, 2018.

[147] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin and K. S. Wilson, *Acta Crystallographica Section D*, 2011, **67**, 235–242.

[148] P. Setny, *Journal of Chemical Theory and Computation*, 2015, **11**, 5961–5972.

[149] X.-Y. Meng, H.-X. Zhang, M. Mezei and M. Cui, *Current Computer Aided-Drug Design*, 2011, **7**, 146–157.

[150] M. P. Baumgartner and C. J. Camacho, *Journal of Chemical Information and Modeling*, 2015, **56**, 1004–1012.

[151] R. S. Armen, J. Chen and C. L. Brooks III, *Journal of Chemical Theory and Computation*, 2009, **5**, 2909–2923.

[152] G. A. Ross, G. M. Morris and P. C. Biggin, *PLOS ONE*, 2012, **7**, e32036+.

[153] M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *Journal of Molecular Biology*, 1996, **261**, 470–489.

[154] X.-Y. Pan and H.-B. Shen, *Protein and Peptide Letters*, 2009, **16**, 1447–1454.

[155] Z. Yuan, T. L. Bailey and R. D. Teasdale, *Proteins: Structure, Function, and Bioinformatics*, 2005, **58**, 905–912.

[156] Y. Deng and B. Roux, *The Journal of Chemical Physics*, 2008, **128**, 115103.

[157] G. A. Ross, H. E. Bruce Macdonald, C. Cave-Ayland, A. I. Cabedo Martinez and J. W. Essex, *Journal of Chemical Theory and Computation*, 2017, **13**, 6373–6381.

[158] G. M. Saunders, H. E. Bruce Macdonald, J. W. Essex and S. Khalid, *bioRxiv*, 2018.

[159] A. Bortolato, B. G. Tehan, M. S. Bodnarchuk, J. W. Essex and J. S. Mason, *Journal of Chemical Information and Modeling*, 2013, **53**, 1700–1713.

[160] S. R. Zia, R. Gaspari, S. Decherchi and W. Rocchia, *Journal of Chemical Theory and Computation*, 2016, **12**, 6049–6061.

[161] Schrödinger, *The PyMOL Molecular Graphics System, Version 2.0*, 2018.