UNIVERSITY OF SOUTHAMPTON

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

Predicting the Location and Binding Affinity of Small Molecules in Protein Binding Sites

Michael Steven Bodnarchuk

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Southampton

September 2012

Supervisor: Prof. Jonathan Essex

Industrial Supervisor: Dr. Russell Viner

Advisor: Dr. Syma Khalid

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES
SCHOOL OF CHEMISTRY

Doctor of Philosophy

PREDICTING THE LOCATION AND BINDING AFFINITY OF SMALL MOLECULES IN PROTEIN BINDING SITES

by Michael Bodnarchuk

In this thesis, various methods for locating and scoring the binding affinity of water molecules and molecular fragments in protein binding sites are described. The primary aim of this work is to understand how different methodologies compare to one another and how, by carefully choosing the correct method, they can be used to extract information on how small molecules interact with proteins. Three different methods are used to predict the location and affinity of water molecules; Just Add Water Molecules (JAWS), Grand Canonical Monte Carlo (GCMC) and double-decoupling. By applying the methods to the N9-Neuraminidase system, it can be shown that all of the methods predict the same binding free energy of the water molecules to within error. The JAWS method was shown to be advantageous for the rapid prediction of the binding free energy of water molecules, whilst GCMC was preferred for the prediction of hydration sites. The combination of the methods were used on a variety of novel test cases, including hydrophobic cavities and protein kinases. These test cases highlight how the methods can be used to accurately predict hydration patterns as a function of the binding free energy in GCMC simulations, and how these patterns can be used to dictate ligand design in a drug discovery context. The approaches described are likely to be of interest to the pharmaceutical industry. A JAWS based fragment based drug discovery methodology is also described, which takes into account key features commonly neglected by existing computational approaches such as fragment-solvent competition and fragment desolvation. This method is used upon the Kinesin Spindle Protein and factor Xa, and demonstrates that the method is able to accurately locate the position of molecular fragments and water molecules compared to crystallographic ligands.

DECLARATION OF AUTHORSHIP

I, MICHAEL STEVEN BODNARCHUK, declare that the thesis

"PREDICTING THE LOCATION AND BINDING AFFINITY OF SMALL MOLECULES IN PROTEIN BINDING SITES"

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission

Signed : .....

Date : .....

Acknowledgments

Firstly I would to thank my supervisor, Jon, for his support and guidance throughout my PhD. In addition, for helping to arrange an extra years funding, without which many of the ideas presented in this thesis would not have come to fruition.

I would also like to thank my industrial supervisor at Syngenta, Dr. Russell Viner for his advice throughout my PhD, and to Syngenta and the BBSRC for funding the first three years of my research.

Many of the ideas and implementations in this thesis have come from discussions with various people, some of which I would like to thank here. Thanks to Dr. Julien Michel for his help and advice in implementing the JAWS biasing scheme, and for his support in ensuring the results obtained for test systems were accurate. I would also like to thank Marcel Verdonk and Richard Hall at Astex Pharmaceuticals for their help in developing a script for the visualisation of JAWS and GCMC density plots with AstexViewer, and in particular to Marcel for many useful fragment-based discussions.

Many thanks to Vernalis, and in particular Nicolas Foloppe, for providing the Chk-1 system as a test of the methodologies. In addition, for helpful discussions relating to the analysis of the results. I would also like to thank some of the other collaborators who I have worked with throughout my PhD, in particular Andrea Bortolato at Heptares.

I would like to express my gratitude to the various members of the Essex research group who I have worked with throughout my PhD. In particular, I would like to thank Sophia Wheeler and Chris Cave-Ayland for helping to proof-read my thesis, and Patrick Schöpf for helping me when I first joined the group and for many useful discussions over the years. I would like to reserve special thanks to Michael Carter, without whom my time in the group would have been significantly less enjoyable and productive, and has been a true friend throughout my time in the group.

I would like to thank all of my housemates who I have lived with throughout my time in the group, but in particular Andrew Guy and Dan Mason who have been amazing friends and housemates. Also I would like to thank Cally Haynes and David Robson, who have been so generous in allowing me to stay in their flat during the week in the last few months of my PhD. I would also like to thank all of my other friends who I have known during my PhD, but in particular Stephen Moore and Sam "Sammy_Kings" Keltie for many fun evenings in the last 4 years.

I would like to thank my amazing and beautiful girlfriend, Rebekah, who has made me so, so crazily happy ever since I met her, and has always been there for me during the final stages of my PhD and writing of my thesis.

Finally I would like to thank my parents, who have always supported me in everything I do. Your love and encouragement throughout the years has lead to this thesis, and I dedicate this thesis to you both.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

In the past 20 years, computational methods have been widely adopted by both academic and industrial researchers as a vital part of the drug development cycle. Whether they are used to predict the correct pose of a drug in a binding pocket [1], simulate the dynamics of a protein [2] or calculate the affinity between two protein-ligand complexes [3], it is now commonplace to use computational methods alongside the more traditional wet chemistry approaches to aid the discovery process.

For a computational method to be applicable in drug discovery it must fulfil several criteria:

- It must be efficient, and capable of producing results rapidly
- The results must be consistent with experimental evidence
- Alternatively, the method should deliver results which experimental methods cannot easily achieve

In recent years, there has been a trend towards bottom-up approaches in drug discovery. Rather than trying to fit an entire molecule into a target binding site, medicinal chemists attempt to piece together potential drug molecules from smaller molecules.[4] The size of these small molecules vary significantly. At the bottom end of the spectrum lie water molecules. It has long been recognised that water molecules play a crucial role in protein-ligand complexes, with classical thought proposing that the incorporation of water molecules into a protein-ligand complex can provide a boost to the binding free energy.[5, 6, 7, 8] Alternatively weakly bound water molecules can be displaced upon binding, providing an entropic boost to the binding free energy.

At the other end of the spectrum lie approaches associated with Fragment-Based Drug Discovery.[4, 9, 10] Compounds with molecular weights up to 300 Da are screened against a protein target to find both the location and affinity of the fragment. With knowledge of the bound fragment, locations drug molecules can be constructed around the fragment scaffolds, delivering compounds which should be more potent and display improved Lipinski properties.[11]

Experimentally, locating water molecules and fragments is typically performed using X-ray crystallography. Although the method gives direct evidence for binding, it is often limited by the resolution of the equipment and the ability to obtain a high quality crystal.[12] Such drawbacks make the prediction of small molecules in protein binding sites an appealing prospect for computational chemistry.

1.2 Thesis aims and outline

Various computational methods have been applied to predicting the location and affinity of water molecules, although no study has ever critically appraised them

in a comparative context. This provides one of the key aims of the project; to understand how different methodologies perform when applied to the same system. This knowledge is important, since it allows the correct method to be used for specific problems.

A second aim is to understand the role of apo hydration in the protein-ligand binding process. For a ligand to bind to a protein, the ligand must either displace or incorporate the apo waters. Understanding how the role of these unbound apo waters influences ligand binding is of great interest, since it opens the possibility for rational drug design through exploiting strongly or weakly bound water molecules in the protein binding site.

A final aim is to develop a computational approach to simulate fragments in protein binding sites. For such a method to be of use, it should take into account competition with water molecules - something which existing computational approaches typically neglect and plays a crucial role in the fragment binding process. In addition, the predictions made by the approach should be able to be validated against experimental data.

Chapter two of this thesis describes some of the key principles behind this project, and how the Boltzmann distribution underlines the whole of computational chemistry. Chapters three and four provide an overview of the literature concerning water molecules and fragment-based drug discovery respectively, and highlight some of the current work in these fields. Chapter five describes the development of the Just Add Water Molecules (JAWS) methodology [13], and how this method has been used alongside the double-decoupling [14] and Grand Canonical Monte Carlo (GCMC) [15, 16] to compare and contrast the methods on the N9-Neuraminidase system. Chapter six takes the three methods and applies to them to a number of different case studies, to help understand the role of active site

hydration in protein-ligand complexes. Chapter seven describes the development of a JAWS-based Fragment-Based Drug Discovery method, and demonstrates its use on two different test systems. Finally, chapter eight concludes this thesis and evaluates the work performed herein.

Chapter 2

Computational Methods

2.1 Introduction

With ever increasing computational power, it has now become possible to begin to simulate complex processes such as protein folding [17] and membrane permeation [18], simulations which 20 years ago would never have been possible. Such processes are typically modelled using *Monte Carlo* or *Molecular Dynamics* techniques. The following chapter attempts to briefly explain some of the basic principles behind molecular modelling. Considering the huge number of studies and methods available in the literature, it is impossible to fully, or even partially, explain some of the methods. As such, there are excellent introductionary sources which go into some of the finer elements of the methods, such as *Molecular Modelling: Principles and applications* by Leach [19] and *Computer Simulation of Liquids*, by Allen and Tildesley.[20]

2.2 Statistical Mechanics

Statistical mechanics is a method which allows the thermodynamic properties of a system to be expressed using information from the microscopic level. The theory is based upon the idea of entropy, and more specifically the second law of thermodynamics. Statistical mechanics, using the laws set down by Boltzmann, allows for a precise definition of entropy based upon the number and population of microstates in the system.

2.2.1 The Boltzmann Distribution

The following derivation is taken from pages 510-511 of *Physical Chemistry* by Atkins.[21]

The Boltzmann distribution can be derived by considering a system of N non-interacting molecules. Each molecule, n_i, can exist in a state of energy ε_0,ε_1...where ε_0 is the state with lowest energy. The instantaneous configuration of the system fluctuates with time, yet some configurations are intrinsically more likely others. Crucially, the total energy of each configuration must be the same. For example, consider Figure 2.1. In this figure, each configuration has a total energy of 7 units and 6 particles. The likelihood of any one of these states can be expressed numerically as the configurational weight of the configuration:

$$W = \frac{N!}{n_0!n_1!n_2!...} \tag{2.1}$$

Using equation 2.1, it can be seen that there are 180 ways of achieving the first configuration, 30 ways of the second but only 6 of achieving the third.

One question which might be posed is whether there is a dominant configuration which is indicative of the system properties which are observed. To go about

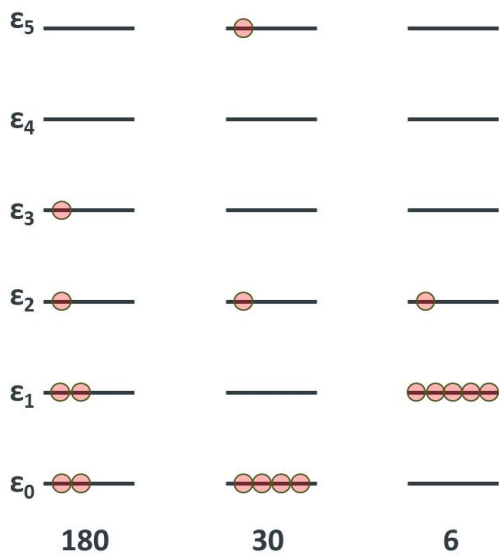


Figure 2.1: 3 sample configurations satisfying the criteria of the same energy and number of particles. The energy of each level, ϵ_n is denoted by n

answering this question, two constraints need to be applied:

$$\sum_i n_i \varepsilon_i = E \tag{2.2}$$

$$\sum_i n_i = N \tag{2.3}$$

Equations 2.2 and 2.3 ensure that any configuration in the same system will have the same total energy, E, as well as the same number of particles.

In order to find the most probable configuration, we need to differentiate W with respect to all of the populations in the system subject to the constraints. This is not an easy transformation however, and Lagrangian multipliers are required. This gives the condition for the most likely configuration as:

$$\frac{\partial \ln W}{\partial n_i} + \alpha - \beta \varepsilon_i = 0 \tag{2.4}$$

In equation 2.4, α and β are constants. The solution to this equation can be estimated using Sterling's approximation:

$$\ln x! \approx x \ln x - x \tag{2.5}$$

Combining this equation with equation 2.1 gives:

$$\ln W = \ln \frac{N!}{n_0! n_1! n_2! ...} \tag{2.6}$$

$$\ln W = \ln N! - \sum_i \ln n_i! \tag{2.7}$$

$$\ln W \approx \ln N! - \sum_i (n_i \ln n_i - n_i) \tag{2.8}$$

We can now estimate a solution to equation 2.4:

$$\frac{\partial \ln W}{\partial n_i} = -\frac{\partial}{\partial n_i} (n_i \ln n_i - n_i) \tag{2.9}$$

$$\frac{\partial \ln W}{\partial n_i} = -\ln n_i \tag{2.10}$$

Putting this result into equation 2.4, we get:

$$-\ln n_i + \alpha - \beta \varepsilon_i = 0 \tag{2.11}$$

Hence, the most probable populaton of the state of energy ε_i is:

$$n_i = e^{\alpha - \beta \varepsilon_i} \tag{2.12}$$

Using constraint 2.3, we get:

$$N = \sum_i n_i = e^{\alpha} \sum_i e^{-\beta\varepsilon_i} \tag{2.13}$$

Rearranging equations 2.12 and 2.13, we get the Boltzmann distribution:

$$n_i = e^{-\beta\varepsilon_i} e^{\alpha} = \frac{N e^{-\beta\varepsilon_i}}{\sum_i e^{-\beta\varepsilon_i}} \tag{2.14}$$

2.2.2 The Molecular Partition Function

The probability distribution, π, of the canonical ensemble is inexplicitly related to equation 2.14 and the Boltzmann equation:

$$\pi_{NVT}(i) = \frac{1}{Q_{NVT}} exp(-\beta\epsilon_i) \tag{2.15}$$

The denominator of equation 2.14 is known as the molecular partition function, Q, and plays the role of a normalisation constant in equation 2.15.

$$Q = \sum_i e^{-\beta\varepsilon_i} \qquad \beta = 1/k_b T \tag{2.16}$$

This equation contains all the information about the thermodynamics of a system of non-interacting particles. Knowledge of the partition function allows the calculation of all of the thermodynamics of the system. As a result, the partition function is of critical importance. Such properties which can be calculated are the Helmholtz energy, A and the pressure, p:

$$A = -k_b T \ln Q \tag{2.17}$$

$$p = k_b T \left(\frac{\partial \ln Q}{\partial V} \right)_T \tag{2.18}$$

In equations 2.17 and 2.18 k_b is the Boltzmann constant, equal to 1.38 x 10^{-23} J/K. T is the temperature of the system and V the volume of the system.

In the limit of a finite number of states, equation 2.16 can be used to define the partition function. However, when dealing with a large number of states, equation 2.16 can be replaced by an integral which considers the 6 dimensional phase space.

$$Q_{NVT} = \frac{1}{N!} \frac{1}{h^{3N}} \int \int d\boldsymbol{p}^N d\boldsymbol{r}^N exp(-\beta E(\boldsymbol{p}^N \boldsymbol{r}^N)) \tag{2.19}$$

In equation 2.19, $\mathbf{p}^N$ and $\mathbf{r}^N$ are the positions and momenta of the each of the N particles. h is Planck's constant, equal to 6.63 x 10^{-34} Js, and 1/N! is a term which deals with indistinguishable particles.

The partition function can be generalised to introduce the idea of interacting particles through the idea of ensembles. An ensemble can be thought of as a large collection of system replicas, whereby each replica in the ensemble could be representative of the true state of the system. One example of an ensemble is the *canonical ensemble*. This ensemble can be thought of as a collection of identical systems which are in thermal contact with each other, allowing the exchange of energy between systems but keeping the temperature and number of molecules in each system constant. Crucially however, the total energy across all of the systems remains constant.

The reason for using ensembles is that it is extremely difficult to know the exact state of a system at any moment, since the system is constantly exchanging energy with its surroundings. Hence, it becomes preferable to collect a time average of the properties of the system which we want to observe. Recognising the

need for time averages, but also that they were unobtainable in the 19^{th} century, Gibbs invoked the *ergodic hypothesis*, whereby it is stated that the time average of a system is equal to the average of an infinite number of replicas at a single instant. The calculation of time averages can now be performed using Molecular Dynamics simulations, discussed in section 2.4.

2.3 Empirical Force Fields

From looking at the partition function, it can be seen that the energy of the system needs to be calculated. The way which this is commonly achieved is by splitting the energy contributions into two parts; the kinetic and potential energy. The kinetic energy of the system can be found analytically from the masses and velocities of the particles, using equation 2.20.

$$E_k = \frac{1}{2}mv^2 \tag{2.20}$$

In equation 2.20, m is the mass of the particle and v the velocity of the particle.

The potential energy cannot be found in such a way. The way the potential energy of the system is commonly found is via the use of molecular mechanics, which typically finds the potential energy of the system as a function of the coordinates of the system.

The total potential energy of a system can be thought of as a sum of all of the intra- and inter-molecular contributions within the system:

$$E_{total} = E_{bond} + E_{angle} + E_{dihedral} + E_{coulombic} + E_{dispersive} \tag{2.21}$$

E_{bond} and E_{angle} are described via a harmonic potential:

$$E_{bond} = \sum_{bonds} \frac{k}{2}(l - l_{eq})^2 \tag{2.22}$$

$$E_{angle} = \sum_{angles} \frac{k}{2}(\theta - \theta_{eq})^2 \tag{2.23}$$

where k is the force constant, l is the bond length, θ is the value of the angle and l_{eq} and θ_{eq} are the equilibrium values of the bond length and angle respectively.

The dihedral energy is modelled with a cosine function:

$$E_{dihedral} = \sum_{dihedral} k_n(1 + \cos(n\phi - \delta)) \tag{2.24}$$

Here, n is the multiplicity of the function (the number of minima in the function as the bond is rotated through 360^o). δ is the phase factor, which gives the point where the dihedral energy is at its lowest value. ϕ is the rotation angle, whilst k_n is the amplitude of the cosine function and represents the force constant.

The inter-molecular contributions are made up of two parts; electrostatic (coulombic potential) and dispersive and repulsive terms (Lennard-Jones potential):

$$E_{inter} = \sum_{i} \sum_{i>j} \left\{ \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \tag{2.25}$$

Here, *i* and *j* represent intermolecular atom pairs, with q_i and q_j the atomic partial charges on atoms *i* and *j*. ε_{ij} and σ_{ij} are the Lennard-Jones well depth and collision diameter for atoms *i* and *j*, with r_{ij} the inter-atomic distance.

The parameters used within force-fields can be derived differently, meaning that the combination of different force-fields is rarely performed. The Generalised Amber Force Field (GAFF) obtains its parameters through a combination of em-

pirical and quantum mechanical means.[22] Partial charges are typically obtained using a high level quantum theory such as Hartree-Fock calculations, whilst bond lengths are typically obtained by a combination of experimental methods (such as X-ray crystallography) and high level *ab initio* calculations. A similar procedure is performed to obtain the parameters for the angle bending and torsional terms. In comparison the GROMOS force-field is purely empirical, with non-bonded terms parameterised to fit experimental properties such as the free enthalpy of hydration. Bonded terms are parameterised purely against crystallographic and spectroscopic data.[23]

2.4 Sampling the phase space

Empirical force fields allow us to calculate the relationship between atoms in a system, but it does not allow us to sample the phase space of the system. There are two major methods which are used to sample the phase space of systems; Monte Carlo (MC) and Molecular Dynamics (MD). These methods are now explained.

2.4.1 Monte Carlo simulations

Monte Carlo simulations are an example of a stochastic technique which can be used to sample properties of a system. Attempting to calculate the properties of a system by averaging over every configuration is impossible, so a method for only sampling states which make the biggest contribution to the partition function is required. This can be achieved by generating a Markov chain of configurations, whereby each new configuration is generated by a random change in the preceding configuration. Such a change is typically made through making a change in the Cartesian coordinates of one or more particles in the system. One problem with

this method is that a huge fraction of the phase space does not make an important contribution to the partition function. Hence, a method is required to sample the states which contribute most to the partition function.

A method for sampling the most relevant states to the partition function was developed by Metropolis.[24] The Metropolis algorithm is detailed below:

1. Start in state *i* and attempt a move to state *j* with probability p_{ij}
2. Accept this move with probability α_{ij} = min(1,χ), where χ is ratio of the probability density, π, of the states *j* and *i*
3. If the move is accepted, then state *i* becomes state *j*. Else, i=i
4. Measure the property of interest, and return to 1

It is important in the Metropolis Monte Carlo algorithm that detailed balanced is preserved. That is, the probability of moving from *i* to *j*, before weighting by π_i and π_j is the same as the probability of moving from *j* to *i*.[25] When this is obeyed, the acceptance test for the move from state i to j is:

$$\frac{\pi_j}{\pi_i} = \frac{exp(-\beta U_j)/Z_N}{exp(-\beta U_i)/Z_N} \tag{2.26}$$

In equation 2.26, Z is the configurational integral of the system, and is proportional to the potential energy part of the partition function. Fortunately it can be seen that the two configurational integrals in equation 2.26 cancel, since this parameter cannot be determined for large systems since it is a 6 dimensional integral. This leaves the acceptance test as equation 2.27:

$$\frac{\pi_j}{\pi_i} = \frac{exp(-\beta U_j)}{exp(-\beta U_i)} = exp(-\beta(U_j - U_i)) \tag{2.27}$$

In a simulation, when a move is performed the energy of the new configuration is calculated and the move accepted or rejected according to the Metropolis acceptance criterion, detailed below:

1. If the energy of the new configuration is lower than that of the preceding configuration then the new state is automatically accepted.

2. If the energy of the new configuration is higher than that of the preceding configuration, then a random number between 0 and 1 is chosen. The Boltzmann factor of the two configurations is calculated and compared to the random number. If the random number is less than the Boltzmann factor the new configuration is accepted, else the preceding configuration is retained and counted again in the overall average.

Such a procedure ensures that only the configurations which make the largest contribution to the partition function are included in the running average. Once the simulation has finished, the properties of interest are found by averaging over all accepted configurations using equation 2.28 below:

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A\left(\mathbf{r}^N\right) \tag{2.28}$$

In this equation, M is the number of configurations and $\mathbf{r}^N$ the Cartesian coordinates of that particular configuration. Monte Carlo simulations can be performed in at least three different ensembles, each with their own acceptance tests:

- Canonical ensemble (NVT): constant temperature, number of particles and volume. Equation 2.27 shows the acceptance test for this move.

- Isothermic-Isobaric ensemble (NPT): constant temperature, pressure and number of particles. Equation 2.29 shows the acceptance test for this move.

$$acc\,(A \rightarrow B) = min\left[1, exp\left(\frac{-\Delta E + P(V^n - V^o)}{k_B T} + N \ln \frac{V^n}{V^o}\right)\right] \tag{2.29}$$

In equation 2.29, P is the pressure in the system, with V^n and V^o denoting the new and original volumes of the system respectively. N is the number of molecules within the system.

- Grand-Canonical ensemble (μVT): constant chemical potential, volume and temperature. The acceptance tests for the Grand Canonical ensemble will be discussed in a later section.

2.4.2 Molecular Dynamics

Whereas MC is a stochastic technique, MD is deterministic, meaning that the preceding configurations can be found from the current configuration. In MD, the n+1th configuration is found by integrating Newton's laws of motion. A so-called "trajectory" is found during a MD simulation, which tracks the positions and velocities of the particles as a function of time. This trajectory is based upon Newton's second law of motion, **F** = m**a** and the resultant differential equations:

$$\frac{d^2\mathbf{x}_i}{dt^2} = \frac{\mathbf{F}_{xi}}{m_i} \tag{2.30}$$

Given an initial starting configuration and velocities, all details relating to the trajectory can be found at any point in space. Since the movement of one atom in the system can affect the velocity and position of other atoms in the system, integrators are used in MD to help to calculate the new positions and velocities. Although these integrators are often extremely efficient they require a large number

of computations to be performed, meaning that running a MD simulation requires significantly more computional power than a MC simulation due to the number of forces which need to be calculated. Although this is one potential drawback to MD, it also allows the efficient sampling of extremely large systems ($>$ 100000 atoms) to be performed, something which can be difficult for Monte Carlo simulations to achieve, especially on polymeric systems.

2.5 Free Energy Simulations

For the canonical ensemble, the free energy is expressed as the Helmholtz function, A, as in equation 2.17. The importance of free energy cannot be understated, since it is the driving force behind chemical processes. Thus it can be seen as a highly desired quantity, but it is difficult to calculate for systems with a large number of particles. As shown in equation 2.19 the partition function Q is a 6-dimensional integral, and evaluating this integral for large systems becomes an intractable task. Methods such as Monte Carlo or Molecular Dynamics only sample the low energy regions of the space, and attempting to sample the large number of degrees of freedom which contribute to Q becomes impossible. As such, methods have been devised which allow the calculation of relative free energies between two different systems; a calculation which is significantly easier to perform since it can be calculated as the ensemble average of the Boltzmann exponent between the two systems. Such approaches can be put into two categories; rigorous methods and approximate methods.

2.5.1 Rigorous methods

Free Energy Perturbation

According to the Zwanzig equation,[26] the free energy difference between two states, A and B, can be expressed as:

$$\Delta G_{A\rightarrow B} = -k_B T \ln \left\langle exp\left(-\frac{\Delta E}{k_B T}\right)\right\rangle_A \tag{2.31}$$

In equation 2.31, $<>_A$ represents the ensemble average over system A and Δ E represents the change in energy between states B and A. A is defined as the reference state of the simulation, whilst B is the perturbed state. One potential problem with this approach lies when states A and B do not overlap in phase space. Such a scenario means that the simulation run with potential U_A will not sample sufficient configurations of U_B and results in the values of ΔE becoming large. Equation 2.31 demonstrates that this results in the exponent becoming extremely small - resulting in small numbers being contributed to the overall average. Whenever the phase space of states A and B overlap the value of ΔE is much smaller, causing a large contribution to the overall average. As such the overall average converges very slowly, and generally results in the free energy being overestimated. A solution to this is to link states A and B in configurational space by the use of a coupling parameter, λ, which introduces intermediate states. For the above example, the reference state A would be defined to be the λ=0 state, whilst the final state, B, would be defined as the λ=1 state.

In the free energy perturbation approach (FEP), the simulation is broken into multiple λ windows between the two end states. Each window is defined a specific λ value, and the free energy of that reference state is calculated. The energy between each λ window is found, and then used in equation 2.32.

$$\Delta G = \sum_{\lambda=0}^{1} -k_B T \ln \left\langle -\frac{\Delta E'}{k_B T} \right\rangle_{\lambda} \tag{2.32}$$

In equation 2.32, $\Delta E'$ is the energy difference between the states λ+$\Delta\lambda$ and λ, where $\Delta\lambda$ is the interval between two successive λ windows.

Thermodynamic Integration

In the thermodynamic integration (TI) approach to calculating the free energy, the rate of change in free energy with respect to λ is calculated across the λ trajectory. The gradients are then integrated to give the relative free energy as in equation 2.33.

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\partial G}{\partial \lambda} \right\rangle_{\lambda} d\lambda \tag{2.33}$$

The resultant integral is typically calculated numerically.

Finite Difference Thermodynamic Integration

In the finite difference thermodynamic integration approach (FDTI),[27] a combination of FEP and TI is used. Instead of calculating the partial derivative of the free energy gradient, the finite difference approximation is used to calculate the free energy gradient.

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{\Delta G}{\Delta \lambda} \right\rangle_{\lambda} d\lambda \tag{2.34}$$

The simulation is broken up into multiple λ windows and, for each value of λ, the free energy is computed over a small interval, typically λ+0.001, using FEP. The total free energy change is then calculated numerically by integrating

over the computed values. Conceptually, FDTI is very similar to FEP. In FEP, the perturbed states are the neighbouring windows, whilst in FDTI the perturbed states are $\Delta\lambda$ above and below each window. The calculation of a numerical free energy gradient rather than an analytical gradient saves the need for a differentation step, although this is not achievable for all potentials.

Replica Exchange Thermodynamic Integration

The replica exchange thermodynamic integration method (RETI) can be considered to be a combination of FDTI and the Hamiltonian replica exchange method.[28, 29] The λ coordinate scales the force field terms linearly, leading to a system which has a different Hamiltonian at each λ value. In RETI, the coordinates between neighbouring λ values are periodically swapped according to the following Metropolis test:

$$rand(0,1) \leq exp\left[\frac{1}{k_B T}\left(E_B(j) - E_B(i) - E_A(j) + E_A(i)\right)\right] \tag{2.35}$$

In equation 2.35, A and B are two different λ values and i and j are replicas of the system at those λ values. RETI has been applied to the calculation of relative binding free energies of ligands to proteins, and has been shown to enhance the sampling of the phase space. Since the configurations of different λ values are passed across the λ coordinate, it has been observed that better free energy convergence is obtained.[28, 29]

2.5.2 Approximate free energy methods

Whereas rigorous free energy methods take into account the intermediate states between λ = 0 and λ = 1, approximate methods typically take into account the two end points of the simulation. As a result such simulations are typically significantly faster than their rigorous counterparts, although the accuracy of the methods are typically significantly poorer.[25] One of the most widely used approximate methods is the Molecular Mechanics / Generalised Born (Poisson Boltzmann) Solvation Area (MM/GB(PB)SA) method.[30] In MM/GB(PB)SA the two end points are simulated using either MD or MC, with the binding free energy calculated as equation 2.36:

$$\Delta G_{bind} =< \Delta E_{mm} > +\Delta G_{solv} - T\Delta S \tag{2.36}$$

In equation 2.36 $< \Delta E_{mm} >$ is the difference in the molecular mechanics energy between the complex and the isolated protein and ligand. ΔG_{solv} is found as the difference in the solvation free energy between the complex and the individual components, although it is often challenging to calculate the non-polar contribution to the solvation free energy.[31] The method also requires an entropic term to calculate the change in binding free energy, although this is often difficult to estimate. This term is commonly ignored when structurally similar ligands are compared, since it is assumed that the change in entropy between similar ligands upon binding to the same receptor is extremely similar.[32] The results obtained using the method are often significantly poorer than those found using rigorous methods, although the methodology has found use in the rescoring of docking poses.[32]

Another example of an end-point method is the Grand Canonical Monte Carlo

approach.[15, 16] This will be discussed in depth in section **??**.

2.6 Conclusions

In this chapter, a brief overview of statistical mechanics was presented. Since all of the key quantities which we would like to know from a molecular simulation, such as the free energy, can be calculated from the partition function, the target for all computational methods is to try and calculate this property. Owing to the complex nature of the partition function, however, direct calculation is impossible, so instead various methods have been devised to calculate the *relative* free energy between two systems. A sampling method such as MD will allow us to look at large systems since it is easily parallelised, although it is often limited in its ability to perform novel sampling schemes, unlike MC. Using TI to look at the relative binding free energy of two inhibitors is a rigorous and well-understood approach, although it is unlikely to be utilised by a pharmaceutical company which desires to screen a large number of compounds in a short space of time due to the computational expense of the method. The choice of the correct methodology to the correct system is something which is often difficult to decide, and is a matter which will be addressed in this study.

Chapter 3

Water Molecules in Drug Design

3.1 Introduction

The following chapter reviews the different roles of water molecules in drug design, focussing upon their relevance in protein binding sites. The importance of their interactions are assessed, before the different computational approaches available for predicting the location and binding free energy of water molecules are described. The known strengths and limitations of the methods are described, followed by the identification of the need for the various techniques to be compared on the same system. Such a comparison has never been described in the chemical literature to our knowledge, and will help to understand the true potential of the various simulation methods.

3.2 The importance of water molecules in drug design

3.2.1 Classical views and applications of water molecules in protein binding sites

The importance of water molecules in drug design and protein structures has become of considerable interest in the recent years.[33, 34, 35, 36, 37] Fuelled by the early work of Poornima and Dean in 1995 [5, 6, 7], it has now commonplace to take into account water molecules at all stages of the drug discovery process. Broadly speaking, there are two major roles which water molecules play in ligand binding. The first role is stabilising the complex via creating a hydrogen bonding network, seen in systems such as N9-neuraminidase. The second role is when a water molecule can be displaced upon ligand binding. This can lead to an increase in the binding affinity of a ligand, since the release of a weakly bound water molecule into the bulk carries an entropic boost to the binding free energy, coupled with a enthalpic gain of strong protein-ligand interactions. Equally however, the displacement of a water molecule can decrease the binding affinity of a ligand, showing that understanding the role of water molecules in drug design is of upmost importance.

A number of studies have shown that incorporation of explicit water molecules can considerably improve docking and virtual screening results.[38, 39] Since bound water molecules can act as essentially part of the protein, neglecting them in docking screens will inevitably lead to poor results. Also, as previously highlighted, water molecules can play a big role in the rational design of drug molecules, meaning that knowledge of whether a water molecule is present or absent from a

complex is highly desirable.

Work by Barillari [40] has helped to understand the nature of water molecules which can be displaced and those which cannot. Utilising double-decoupling simulations with replica exchange thermodynamic integration [28, 29], the paper focussed on understanding whether the binding affinity of a water molecule was related to the propensity of a water molecule to be displaced. The paper demonstrated that, on average, water molecules which are more tightly bound are less likely to be displaced than those which are more weakly bound. With this knowledge the medicinal chemist can decide whether to target a particular water molecule for displacement, or to try and design a ligand which is capable of utilising the hydrogen bonding opportunities affording by the conserved water.

An alternative method for predicting whether a water will be displaced or retained has been proposed by Ross *et al.*[41] Using a rapid docking method termed WaterDock, a method using the freely available AutoDock Vina toolset, they found that they could successfully predict whether a water molecule was displaced at least 75 % using the Astex test set. The method locates the possible hydration sites by attempting to dock water molecules into both apo and holo structures, and found an encouraging 97 % of the waters when compared to the native structures. Water molecule classifiers were identified which were used in a data mining, heuristic and machine learning algorithm, and allowed for the prediction of whether a water would be displaced or conserved when a ligand was overlayed on a possible site. This method shows promise, although it is not capable of predicting the binding affinity of water molecules in a simulation context.

3.2.2 Water : The active player in binding thermodynamics

More recent work has highlighted that the nature of water molecules in protein binding sites goes considerably beyond the two roles previously mentioned. In particular, numerous studies have explored the role which water performs in the hydrophobic effect. It has been traditionally held that the binding of two hydrophobic solutes is primarily entropically driven, due to the release of configurationally restricted water into the bulk. This is also coupled with a favourable increase in the enthalpy in the system due to the formation of new hydrogen bonds in the bulk, providing a thermodynamic force for the solutes to bind together.

An excellent 2007 study by Homans [42] looked at the binding of ligands to the Major Uninary Protein (MUP), a protein which is known to accept small, hydrophobic ligands. In order to understand the subtle differences in the thermodynamics of structurally similar ligands, Isothermal Thermal Calorimetry (ITC) was used. ITC is an experimental technique which is capable of measuring the binding affinity and enthalpy of interactions in the aqueous phase, allowing the direct contribution from entropy and enthalpy to the binding free energy to be determined. Two structurally similar ligands were examined in the study, differing in the addition of a methylene linker in the scaffold of the ligand.

The study examined the contribution of solvent to the binding of the ligands, and found some surprising results. First, it was found that the change in entropy upon protein desolvation was practically zero. This would appear to be in violation of the hydrophobic effect, since the release of conformationally bound water in a hydrophobic pocket should be favourable entropically. Secondly, it was found that, for both ligands, the enthalpy of binding was extremely favourable. The hydrophobic effect suggests that solute-solute dispersion terms should offer a minor contribution to the free energy of binding, since they should, in principle,

be offset by the loss of similar solvent-solute dispersion interactions prior to the binding event. As such, the favourable enthalpy to the binding free energy was unexpected.

The explanation for both phenomena was found in the solvation of the protein in the uncomplexed state. Molecular Dynamics simulations were performed upon the protein and found that the pocket was sub-optimally hydrated, with the pocket only filled with approximately 20 % density of bulk water. As such, the enthalpic binding signature can now be fully explained. The disordered and poorly hydrated binding pocket does not experience a large entropic loss upon solvent removal, since the pocket is already largely empty. Indeed, it was reported that the pocket is actually entropically *favourable* to be sub-optimally hydrated rather than have the solvent in the bulk. The lack of solvent in the pocket allows a large contribution to the binding free energy by solute-solute dispersion interactions, since these are not offset by the opposing solvent-solute interactions. As such, this offers an explanation for why the larger compound with an additional methylene linker has a more favourable binding free energy.

The discovery that not all proteins are fully hydrated has important consequences for drug design. Targetting sub-optimally hydrated proteins with ligands which optimise solute-solute dispersion interactions through shape complimentarity is likely to lead to potent inhibitors, since these interactions will not be offset by the solvent-solute interactions in the unbound state. In addition, targetting specific, poorly hydrated, regions of binding sites with strong solute-solute dispersion interactions is likely to be successful. The obvious barrier to such an approach lies in the identification of such regions within protein binding sites; something which is highlighted by Homans.[42]

Following on from Homans work, a number of other groups have started to

concern themselves with the idea of solvent-based thermodynamic signatures to binding events. A study by Setny looked at the binding thermodynamics of a model hydrophobic receptor-ligand system, using explicit MD simulations to derive a PMF for the receptor-ligand interaction.[43] The model hydrophobic receptor was made out of a hexagonal closed packed lattice of Lennard-Jones spheres, with the ligand treated as a neutral Lennard-Jones sphere. The TIP4P model was used to model the water molecules.[44] It was found that for this model system the overall free energy for complete hydration of the pocket was zero, with the surprising result that the removal of water from the pocket was entropically *unfavourable* - suggesting that water in this model system is highly mobile. It was recognised in the study that the solvation of the cavity was due to expansions and retractions of the bulk phase. The entropy was calculated by observing the temperature dependence of the free energy based on 5 MD simulations performed at different temperatures.

Interestingly this increase in entropy is offset by the gain of enthalpy in the bulk through solvent-solvent interactions, resulting in the net free energy change of zero for protein desolvation. The driving force for the binding of receptor and ligand in this study was found to be predominantly due to the strong gain in solvent-solvent enthalpy upon desolvation and receptor-ligand binding, with the receptor-ligand dispersion terms playing a much smaller role.

It can be seen that the conclusions drawn from this study appear to be in contradiction to the work performed by Homans. Whilst both studies agree that hydrophobic association can be enthalpy driven, Homans attributes the driving force to strong protein-ligand interactions which are stronger than the protein-water interactions in the hydrated protein. In comparison, the work by Setny concluded that the driving force is the formation of strong water-water interactions in the

bulk upon desolvation, rather than protein-ligand association. It should be noted, however, that the Setny system which was studied has several key differences to the MUP system. Firstly the receptor used was a model system, built up of a close-packed grid Lennard-Jones spheres, rather than a true protein-ligand system. Such an approach is likely to underestimate the interactions which would normally be made between protein, ligand and water, and as such is not necessarily indicative of reality. Secondly the receptor was solvent accessible, whilst the binding site of MUP is shielded from the bulk solvent. As such it is difficult to compare the studies on an equal footing.

The hydrophobic interaction study used by Setny was used as a basis for a second study looking at the role of water in cavity-ligand recognition.[45] A series of different model systems were used, looking at 7 different combinations of the charge states of the cavity and ligand. As with the previous study, Lennard-Jones spheres were used to describe the cavity and the ligand. For each combination, the different thermodynamic signatures to the binding event was examined. It was found that the thermodynamic signature of water was responsible for the binding (or non-binding) of the ligand to the cavity for each model system, with other effects playing a much more minor role. A possible explanation for this is the over-simplification of the system, since only the water molecules were modelled in explicit detail. Whilst also providing enthalpic and entropic contributions to the binding thermodynamics, it was also found that water plays an electrostatic screening role in the attraction of two oppositely charged species.

Whilst looking at individual water molecules is of great interest in lead discovery and development, wherein specific water molecules are often targetted and exploited, the role of water networks is beginning to be recognised as of equal importance. A recent communication by Hummer highlights the fact that seemingly

subtle changes in the hydrogen bonding network of water molecules can have profound effects upon the binding free energy of ligands.[46] As such, considering water networks alongside individual waters is an idea which is beginning to be incorporated. Indeed, a 2011 paper by Barillari looked at the apo hydration patterns of different kinases, finding qualitative differences in the water networks between different, yet sequentially similar, kinases.[37]

3.2.3 Locating water molecules experimentally

It is therefore apparent that the role of water in protein-ligand interactions is much more than just as solvent [47], and it plays an active role in the binding events. As such, knowledge of where water molecules like to reside in protein binding sites is of paramount importance. The traditional way of identifying water molecules is through X-ray crystallography, although this approach is often limited. A 1999 paper by Carugo [12] suggested that protein resolution is typically a limiting factor in determining the number of water molecules in a protein structure. Looking at 873 known crystal structures, the paper indicated that, on average, a protein structure with a resolution of 2 Å had one water molecule per residue resolved, whilst at a resolution of 1.0 Å around 1.6-1.7 waters are resolved. The same behaviour has been noted by Abel in a recent 2011 paper observing the hydration of the apo structure of thrombin.[48] Another issue with relying on crystallographic methods to predict waters lies in the role of the crystallographer. It has been demonstrated that two indepedently resolved structures of the transforming growth factor-β2 were found to have a different number of crystallographic waters with varying temperature factors [49], suggesting that the addition of water molecules into a crystal structure can be problematic.

Since crystallographic approaches will not necessarily give a true picture of

the hydration patterns within a protein binding site, computational methods have been used to try and locate the position of water molecules. The following subsections review some of these approaches, and highlight the advantages and disadvantages of the methods.

3.3 Grand Canonical Monte Carlo

Formulated by Adams in 1974 [15, 16], the Grand Canonical Monte Carlo (GCMC) technique is one which is capable of predicting the location of molecules in both biological and inorganic systems. Unlike traditional MC and MD simulations, GCMC utilises the μVT ensemble which allows the number of molecules in the system to fluctuate as a function of the applied chemical potential. As such, the methodology is ideally suited to looking at systems where the number of molecules in a system is unknown, such as an apo protein binding site.

In a GCMC simulation the moves associated with the canonical ensemble are permitted, alongside three unique moves associated with the μVT ensemble. The first type of move is a particle creation move, whereby the number of molecules in the system increases by one. The second type of move is a particle deletion, whereby the number of molecules in the system is decreased by one. The final type of move is a localised translational move, whereby inserted molecules are allowed to translate around the system. The acceptance tests for these moves are shown in equations **??**, **??** and **??**.

$$P_{in} = min\left[1, \frac{exp(B)}{N+1} exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{3.1}$$

$$P_{del} = min\left[1, N exp(-B) exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{3.2}$$

$$P_{dis} = min\left[1, exp\left(\frac{-\Delta E}{k_B T}\right)\right] \tag{3.3}$$

In the above equations, N is the number of particles in the simulation and B is the Adams parameter (B = μ'/k_BT + ln $\bar{n}$). $\bar{n}$ is the expected number of particles in the system given the volume of the simulation region and is equal to $\bar{p}$v, where $\bar{p}$ is the number density of the particle and v the simulation volume.[50] μ' is the excess chemical potential, k_B is the Boltzmann constant and ΔE the change in energy between the new and old states. Historically, B has been used in simulations instead of μ, for computational simplicity.[51] No explanation has been found for this parameter, but one possible explanation is that it allows the simulation results to be compared to the expected number of molecules in the bulk, $\bar{n}$. Since B and μ' differ by a constant, performing a simulation at constant B is equivalent to performing a simulation at constant chemical potential, μ'.

One problem which arises from using the Grand Canonical ensemble is the acceptance rate of insertions. If an insertion is attempted into a dense system then it is extremely likely that the move will be rejected due to repulsive van der Waals interactions. Acceptance rates of less than 1 % have been reported in the literature,[13, 50] meaning that adequate sampling of the chemical space can be a problem. In order to try and increase the efficiency of GCMC simulations, strategies have been devised to increase the acceptance rate of insertion moves.

3.3.1 Increasing the acceptance rates of GCMC simulations

The first method developed to attempt to increase the rate of GCMC insertions was to employ cavity bias in the simulation procedure.[51, 52] Rather than randomly inserting a molecule into the system, a preliminary test is made to look for regions where the test particle will fit into the system without overlapping with existing

particles. Typically the van der Waals radius of the test particle is used to look for suitable gaps in the system. Once the potential sites for insertion are identified, an insertion attempt is performed on a random chosen cavity site. If the move is accepted then the cavity positions are recalculated, whilst if the move is rejected the grid is not updated. Equally, if a deletion move is accepted then the cavity positions are recalculated. Since the cavity searching procedure biases insertions into particular points in space, a biasing term needs to be introduced into the Metropolis tests for insertions and deletions, shown in equations **??** and **??**.

$$P_{in} = min\left[1, \frac{exp(B)P_{cav}}{N+1} exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{3.4}$$

$$P_{del} = min\left[1, \frac{N exp(-B)}{P_{cav}} exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{3.5}$$

In equations **??** and **??**, P_{cav} is the probability of finding a cavity of radius R_c or larger. This is found by generating a number of uniformly distributed test points and calculating the fraction which have a cavity of size R_c or greater. If no potential insertion points are found then an insertion attempt is made randomly into the system utilising the standard GCMC acceptance tests.

Even with cavity bias schemes the acceptance rates for GCMC are still not optimal. In order to further increase the efficiency of GCMC simulations, Shelley and Patey utilised a configurational bias scheme to increase the acceptance rate.[53] Since calulating the entire system energy for insertions can be expensive, the authors utilised the fact that calculating the van der Waal's energy is considerably faster than calculating the electrostatic energy. A scheme can therefore be developed which is, in theory, more efficient than standard GCMC:

1. Select an insertion point which satisfies the cavity bias pre-screen.

2. Calculate the van der Waals energy of the new molecule in this position. If the energy of this position is less than zero then move onto the next stage. If the energy of the position is greater than zero then reject with the probability 1-$\alpha\frac{LJ}{ij}$, where:

$$\alpha\frac{LJ}{ij} = 1 \text{ if } \Delta E\frac{LJ}{ij} \leq 0, = e^{-\beta\Delta E\frac{LJ}{ij}} \text{ if } \Delta E\frac{LJ}{ij} > 0 \qquad (3.6)$$

In equation **??**, α is the rejection criterion for the position *j*, β is 1 / k_bT, and ΔE is the change in van der Waals energy upon an insertion attempt. If the value of ΔE is negative then the position is accepted for the next step, else the Boltzmann factor is calculated to determine the acceptance of the position.

3. Trial *n* orientations of the water molecule and weight towards the one with the lowest potential energy.

4. Accept or reject the move according to a modified Metropolis test.

Application of the configurational bias algorithm was found to remove 48 % of potential insertion attempts within a water box at minimal CPU cost, demonstrating the effectiveness of the method. Despite the apparent merits of the technique there has not been a widespread adoption of the method in the chemical literature.

3.3.2 Use of GCMC in biological systems

Despite the wealth of literature relating to the use of GCMC, there are relatively few examples where the method has been used to predict the location of waters in biological systems. The first paper describing such an application involved the hydration of the major and minor grooves of DNA.[54] In order to attempt to

understand the hydration patterns, the authors developed a technique known as simulated annealing of chemical potential.

The method works by starting the simulation at a high value of B. From looking at equation **??** it can be seen that a high B should increase the likelihood of insertion attempts being accepted, resulting in the simulation box being flooded with water molecules. As the chemical potential is gradually lowered the more weakly bound water molecules begin to leave the simulation box, whilst those which are strongly bound remain. This process continues until all of the water molecules have evacuated the system. Since the chemical potential of the system is known at all stages of the simulation a quantitative estimate of the water binding free energy at relevant hydration sites can be calculated.

The application of the method to DNA demonstrated proof of concept in the simulated annealing approach, in that as the chemical potential was lowered the number of molecules in the system decreased. Despite this, the study lacked several aspects of physical realism, primarily that the DNA in the system was held rigid throughout when it is widely held that this is not the case in biology. By not allowing flexibility the system will not sample the entire configurational space, meaning that the results are not necessarily realistic. The authors noted that the waters binding in the major groove were more weakly bound than those in the minor groove, yet it is possible that if such a scenario were to happen in nature the DNA would change conformation to accommodate the change in hydration state.

Another study by Mezei and Resat looked at applying the GCMC routine to the hydration of a cavity in the sodium salt of hyaluronic acid.[55] An isolated cavity within the crystal was chosen and GCMC insertions performed within this region. The molecules were restrained with a hardwall potential to prevent them from leaving the region of interest. Rather than employing simulated annealing, the

authors simply ran the simulation at a range of B values to record the number of waters in the subunit. The simulation corroborated experimental evidence relating to the optimum occupancy of the cavity, although as with the previous example the polymer was kept rigid throughout the simulation.

A study by Pan *et al.* looked at using GCMC to assist lead optimisation.[56] Looking at the HIV-1 Tat system, the authors tried to rationalise changes in ligand binding affinity by looking at the hydration sites near to the ligand. A lead compound was chosen and subjected to a GCMC simulation within the binding pocket, using a chemical potential which matched experimental conditions. 5 potential hydration sites were identified, with one site identified as displacable by ligand substitution. Derivatives of the lead compound were then simulated which attempted to displace this site. For each compound, the number of water deletions in the vicinity of the ligand was recorded. Those compounds which displaced the water most efficiently were claimed to experience the fewest number of deletions and were then tested in an experimental assay, indicating that these were more potent than the lead compound.

Despite the apparent success of the method in identifying potential sites for lead optimisation there are several issues arising from the study. Although techniques such as cavity-bias [51] were employed, the acceptance rate for insertions is still likely to be less than 1 %. As such, monitoring the number of deletions is not likely to be a statistically significant way of assessing the propensity of water displacement. Indeed it would be expected that the compound which displaced the water most effectively would experience the *most* number of successful deletions, since this would suggest the water is not present for the majority of the time. This assertion is backed up by looking at the data supplied in the paper, whereby the compound which was found to have the lowest EC_{50} did not have the fewest num-

ber of deletions. In addition the number of MC moves performed in the study was not mentioned, giving no indication of the efficiency of the method. A better measure would be to monitor the number of successful insertions and deletions into the region, which would allow a probability of the site occupancy to be obtained.

A 2004 paper by Woo *et al.* looked at utilising the grand canonical ensemble to help in using dual solvation approaches.[50] Rather than simulating an entire system using explicit solvation, the approach allows the region of interest to be simulating using explicit solvent whilst the remainder is simulated with implicit solvent. This approach, known as the generalised solvent boundary potential (GSBP),[57] requires waters to be able to cross over the boundary which is governed by allowing GCMC moves set at the chemical potential of water. Such a solvation set-up has two major advantages. First it allows the solvent to sample deeply buried cavities within the protein, something which often results in long timescales if simulated by MD. Second, the system is considerably quicker to simulate if not all of is treated explicitly. In order to achieve adequate sampling in the explicit phase, the authors utilised similar schemes to the cavity bias and configurational bias approaches described previously.

The methodology was tested on two different systems to determine the efficiency of the process. First a pure implicit-explicit water system was tested, with the methodology predicting the expected number of waters in the system given the correct chemical potential. Second the KcsA potassium channel was tested as an example of a system where water passage between the bulk and the interior cavity is typically problematic using standard MD. By utilising the GSBP approach equilibration within the pore was achieved, with an average GCMC acceptance rate of 0.81 %.

A later study by Deng and Roux looked at applying the GSBP/GCMC tech-

nique to the binding of camphor in cytochrome P450cam.[58] The binding of camphor occurs into a deeply buried pocket in the protein, resulting in the explusion of seven water molecules into the bulk. Standard FEP simulations with a fixed number of waters to calculate the absolute binding energy can suffer from energy convergence issues, making the system ideal for the GSBP/GCMC. To calculate the absolute binding energy the FEP simulation was performed in three stages. First, the solute repulsion terms in the pocket are gradually turned on. The solute attraction terms are then included, before the solute electrostatic terms are switched on. By allowing the number of water molecules to fluctuate within the pocket, the calculated binding free energy was found to be within 0.50 kcal/mol of the experimental result compared to an error of $>$ 6 kcal/mol when a fixed number of waters are used.

A more novel use of a Grand Canonical approach has been used by Collins *et al.*[59] Studying at the hydrophobic T4-lysozyme cavity, they looked at the thermodynamics of filling the cavity with a different number of water molecules. A MD simulation was performed, where the number of water molecules within the cavity was varied between 0 and 5. The canonical average of the Boltzmann factor of the potential-energy change ΔU = U_{N+1} - U_N was calculated for the process of inserting a water molecule into the cavity already occupied by N waters, with this related back to the chemical potential and number density of bulk water to arrive at an occupancy probability for each number of molecules within the cavity. The results from the study were verified by X-ray crystallography. This system will be further described in section **??**.

3.4 Just Add Water Molecules

The Just Add Water Molecules (JAWS) methodology was developed by Michel *et al.* as a tool to help predict the water content with protein binding sites.[13] As previously mentioned, this is typically a problem for solvent inaccessible binding pockets since the timescales required for waters to diffuse in and out of such pockets are typically too long to simulate. Although the GCMC method is capable of locating the waters within such pockets, it requires prior knowledge of the chemical potential which matches experimental conditions. This is typically non trivial, and requires several additional simulations if the desired conditions are not known *a priori*. In addition, the acceptance rates for GCMC insertions are typically < 1 %, meaning that adequate sampling of the cavity can be problematic.[50]

The JAWS methodology was designed as a method which is capable of both locating the position of waters within a protein binding and also providing an estimate for the binding affinity of these waters compared to the bulk. Based upon λ-dynamics [60], the approach works by simulating so called "θ-water" molecules which can appear and disappear across a grid located on the binding site. θ is an energy scaling parameter which controls the interaction energy between each θ_i water and the rest of the molecules in the system. If the value of θ_i is 0 then the molecule acts like a ghost particle and does not interact with the system. Equally, if θ_i = 1 then the molecule interacts fully with the system. The potential energy function, E(r), that describes N water molecules with scaling parameters θ_i interacting within a protein binding site is shown in equation **??**:

$$E(r, \sum_{i=1}^{N} \theta_i) = E_0(r) + \sum_{i=1}^{N} \theta_i E_{inter}(r, \text{water } i) \tag{3.7}$$

, where E_{inter} is the intermolecular energy of θ-water molecule *i* and E_0 con-

tains the other energy terms.

A MC simulation is performed whereby the θ-water molecules are allowed to sample the binding site grid whilst also sampling the value of θ. N θ water molecules are distributed across the binding site, so that 1 water molecule per 30 Å^3 is added randomly onto a 3D grid, corresponding to approximately the number density of bulk water. The θ-water molecules are allowed to freely sample the grid, and attempted MC moves consist of rigid body translations, rotations and, for 50% of the moves, a random variation in θ. If the value of θ_i is greater than a predefined threshold, typically 0.95, then a counter is increased by one on the nearest grid point.

The MC simulation is carried out for typically 5-15 M moves, and results in a probability density of water occupancies over the grid. These occupancies are then converted into an integer number of hydration sites using a clustering algorithm. There are two potential problems with performing such an approach:

- The number of hydration sites is not always known
- Clustering the information gives no indication of whether the potential sites are correlated.

Having identified the potential sites, an estimate of the binding free energy is sought. In order to achieve this, statistics need to be collected whereby the molecule experiences both the $\theta > 0.95$ and $\theta < 0.05$ states. Since the free energy barrier between the two states is typically large, a biasing potential is applied to each of the θ_i to enable such transitions. This biasing potential is based upon the hydration free energy of water and can be seen in equation **??**.

$$V(\theta_i) = (-\Delta G_{hyd} + \Delta G_{constr}(\text{ideal, site } i))\theta_i \tag{3.8}$$

In equation **??**, ΔG_{constr} is the free energy for constraining an ideal particle in a volume of V^{constr} instead of the bulk, V^0.

$$\Delta G_{constr} = -k_B T \ln \frac{V^{constr}}{V^o} \tag{3.9}$$

Each water molecule is constrained to occupy a volume of 27 $Å^3$, based on the water locations derived after clustering the first set of simulations, and a new MC simulation is performed with the biasing potential added onto the potential energy function for each θ_i water. Since the biasing potential is based upon the hydration free energy of water, the biasing term penalizes the high θ_i states by an amount that accounts for the desolvation of the water from bulk at the localised site *i*. As such, tracking the $\theta > 0.95$ and $\theta < 0.05$ states can allow an estimation of the binding free energy using equation **??**.

$$\Delta G_{bind}(\text{water, site } i) = -k_B T \ln \left(\frac{P(\theta_i \rightarrow 1)}{P(\theta_i \rightarrow 0)} \right) \tag{3.10}$$

At the end of the JAWS simulation, water molecules with negative binding free energies are retained. One drawback of the technique is that strongly bound water molecules often cannot have their binding free energy estimates found since the biasing potential is not sufficient to sample the $\theta < 0.05$ state.

3.4.1 Applications of JAWS

The original test system for JAWS was the zanamivir bound N9-neuraminidase complex. The binding pocket of N9-neuraminidase is isolated from the bulk solvent and crystallographic evidence resolves 6 water molecules within the site. The JAWS algorithm was trialled upon the system and located 7 potential hydration sites. 6 of these sites closely matched the crystallographic positions, whilst

another additional site was found towards the top of the pocket. These 7 sites were studied using JAWS stage 2, with the resultant binding free energies suggesting that the extra site had a positive binding free energy. Of the 6 crystallographic sites, only one did not experience a transition to the $\theta < 0.05$ state and hence this binding free energy was not converged. The results were validated using double-decoupling MC, with the results in good qualitative agreement.

Michel *et al.* have used the JAWS methodology to assist in free energy calculations.[61] Looking at test systems where water molecules are known to be displaced, the JAWS methodology was used to assess the energetic consequences of such displacements. The scytalone dehydratase system was chosen, whereby upon the substitution of a triazine moiety to a cyano diazine group there is a 30-fold increase in K_i. This transformation involves the displacement of a water molecule, bridging the triazine ring to two tyrosine residues. Another ligand in the series has just the diazine ring with no cyano group, but the occupancy of water for this ligand is ambiguous with a JAWS free energy of -1 kcal/mol. These ligands can be seen in Figure 3.1.

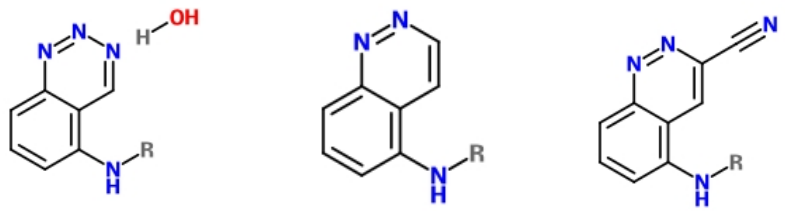


Figure 3.1: Ligands 1-3 in the scytalone dehydratase system

Since performing free energy simulations which involve both a ligand perturbation and water introduction/deletion typically require the use of GCMC, free energy cycles were employed in the study. Two different scenarios were imagined; one where the water was present for the routes 1-3, 1-2 and 2-3 and another

where the water was not present for the perturbations. It was found that for both setups the perturbation 1-2 was not in agreement with the experimental result, indicating that changes in hydration state need to be accounted for. In addition the free energy cycle which retained the water at all stages displayed poor closure ($>$ 4 kcal/mol), suggesting that the free energy for unphysical hydration states is unrealistic.

Utilising the hydration analysis from JAWS, new free energy pathways were devised whereby each ligand had its correct water occupancy. When this was applied, the relative free energy for the free energy perturbations was in much better agreement with experiment compared to standard 'fixed' hydration states. This highlights the importance of taking into account the potential changes in hydration during a perturbation calculation, something which can be rationalised using JAWS. The study also looked at analagous ligands in the EGFR and p38α MAP kinases and found similar behaviour.

A similar 2010 study by Luccarelli looked at the effect of JAWS water placement upon the binding affinity of p38α MAP kinase inhibitors.[62] 18 different ligands were examined, with the relative binding affinity of each calculated compared to a base ligand. When a standard solvent setup was utilised around the binding site the predictive index of the calculated $\Delta\Delta$G values was found to be 0.41.

The predictive index is a measure of ability of a method to correctly rank a series of inhibitors.[63] The method looks at the experimental and predicted binding free energies, and considers pairs of compounds one at a time. The form of the index means that large differences in binding free energies of compounds *i* and *j* will have a large weighting to the index and successfully predicting which of the two compounds is the more potent will provide a large positive contribution to

the final score. Equally, if *i* and j are similar in affinity then an incorrect prediction will have a minor impact. A method which perfectly predicts the ranking will have a value of +1, a model which always predicts incorrectly has a value if -1, whilst a value of 0 arises from predictions which are random.[3]

A JAWS simulation was then performed to obtain a more realistic solvent distribution and the relative binding affinities recalculated. This increased the predictive index to 0.62, indicating that the more realistic solvent distribution from JAWS gives better predictive power.

3.5 WaterMap

The WaterMap software is a commercial package developed by Schrödinger designed for locating and scoring waters in protein binding sites. The methodology utilises Inhomogeneous Fluid Solvation Theory (IFST), a method developed by Lazaridis which applies integral equations to assess the solvation properties of the system.[64, 65] IFST decomposes the solvation free energy into four terms; the solute-solvent energy, solute-solvent entropy, solvent reorganisation energy and the solvent reorganisation entropy.[66] The solute-solvent energy is found in the method as the the difference in energy between the system with the water molecule present and the system without the water molecule.[67] Each one of the entropy terms can be found at any position in the system by solving integrals relating to the solute-solvent and solvent-solvent correlation functions. The solute-solvent correlation function is zero over the region occupied by the solute, meaning that any contribution to the solvation energy and entropy arise from the regions occupied by the solvent.

Such an approach means the method can be used to look at specific regions of

a system and assess the various contributions to the solvation free energy specifically. Where there are several water molecules which bind to a system, IFST can break down the contribution of each to the entropy and enthalpy allowing an estimate of the binding affinity of each water to the system.[67] The entropy of each water is found as the first order correlation function between itself and the solute - crucially water-water entropy is neglected in the calculation. The binding affinity which is found by IFST is not the same as the binding free energy calculated via FEP/TI since the process does not simulate the removal of a solute water; rather it performs a decomposition of the free energy of the system and assigns a certain value to each water molecule.[66] It has not been clearly explained in the literature as to what this free energy calculated by IFST represents, since no reference state has been provided.

The first work utilising IFST within the WaterMap software was performed by Young *et al.* in 2007.[67] The work looked at understanding the differences in phase behaviour and thermodynamic properties of water in enclosed protein regions compared to the bulk. The major focus of the study looked at the binding site of streptavidin, a protein with a hydrophobic binding cavity which binds to biotin extremely strongly. A MD study was performed upon the system, with the positions of water molecules recorded across $>$ 10 ns of simulation time. These water positions were then clustered to arrive at an integer number of hydration sites, with these sites then subjected to IFST.

The simulation found that a 5 membered ring of water molecules is found within the binding cavity, something which is highly unusual in nature due to the high entropic penalty of ordering molecules in such a way. The IFST analysis indicated that these sites are indeed highly ordered and carry a high entropic cost, but this cost is outweighed by the enthalpic benefit of the molecules binding to

each other. This is facilitated by an ideal arrangement of hydrogen bonding motifs within the cavity, allowing the formation of a 5 membered ring. Release of the 5 waters into the bulk carries an entropic boost of 7 kcal/mol, suggesting one reason why the binding of streptavidin to biotin is so favourable.

Studies by Abel *et al.* have looked at using WaterMap to help derive a scheme for calculating relative binding free energies ,$\Delta\Delta$G values, between sets of inhibitors.[68] The studies make the assumption that in order for a ligand to bind to a protein, an equal sized cavity needs to be created in the active site of the protein. A further assumption states that any waters in the binding site are always displaced by a heavy atom from the ligand. Based upon these principles, a scoring scheme was developed which calculates an approximate binding free energy for a ligand based upon the system interaction energy and the entropy of the waters which are displaced. The scoring scheme does not take into account the strength of the ligand-protein interaction, nor does it take into account the entropic changes in the protein and ligand upon binding. As such, it was only used for ligands which are structurally similar.

The study by Abel used the scoring function to look at similar inhibitors of factor Xa. The calculated $\Delta\Delta$G values compared to the experimental activities were in good agreement (R^2 = 0.81) compared to the same calculations using MM-GBSA (R^2 = 0.29). The same functional was then used to look at more diverse ligands, whereby the R^2 value dropped to 0.48. This indicates that the use of such a scoring scheme is limited to similar ligands.

Recognising that one of the major limitations in the standard technique is the influence of ligand-protein interaction and reorganisation, Guimarães and Mathiowetz combined the method with MM-GBSA.[31] One typical problem with MM-GBSA is the poor estimation of protein desolvation, something which the

authors attempt to rectify with the WaterMap method. Rather than calculating the term using the GB model, the protein desolvation energy was instead calculated as the sum of the binding energies of the waters within the pocket. As such, combining the two should give a method which is more accurate that the two individual parts. Two different systems were examined; factor Xa and CDK2, with the results showing that the predicted R^2 values between the two individual methods and the combined method were extremely similar. The authors state that the reason for the lack of improvement is due to the fact that a model with 30 points can, at best, have an R^2 value of 0.80 considering the span of the experimental data.

A 2010 paper by Wang *et al.* introduced a new term into the WaterMap scoring function.[69] Whilst most protein binding sites are well hydrated, there are also those which have regions which are poorly hydrated or even not at all. For such systems the WaterMap scoring function will ignore the ligand heavy atom contribution in these areas, since there is no water to displace in these regions.As a result, the calculated energy will not be accurate. It is important to attempt to include these dry regions in the calculation, since the ligand will often gain affinity by binding in these regions since there is no desolvation penalty to pay. In order to correct for this the authors introduced a 'cavity correction' term, whereby ligand heavy atoms which bind into dry regions gain a binding affinity related to the size of the cavity and the solvation free energy of methane. Upon utilising the cavity correction term in conjunction with the standard Watermap scoring function, a qualitative improvement in the $\Delta\Delta G$ values of MUP inhibitors was observed. No improvement in the values of R^2 was given, whilst the results themselves are still an order of magnitude away from experimental activities.

In a 2010 paper by Robinson *et al.* the WaterMap method was used to attempt to rationalise the binding affinities of kinase inhibitors.[70] The binding

site of kinases are common across a large percentage of known structures, meaning that designing ligands which display selectivity towards a certain kinase can be challenging. The position of water molecules was determined across a MD run, followed by an energetic analysis using IFST. Structurally similar ligands were studied for 4 different kinase systems, and the resultant WaterMap method used to rationalise why certain ligands perform better than others for a particular system. From looking at the energies of the water molecules and the position of the ligands in the active site the method was able to suggest reasons why one ligand is more potent than another, something which was usually attributed to the displacement of a nearby water.

Whilst the results which have been obtained using the WaterMap method show promise, there are several key issues relating to the scoring schemes and methodology which have not been satisfactorily explained in the literature.

1. **Assuming that a ligand will displace any bound water in the system** - Whilst the form of the scoring function ensures that strongly bound waters do not contribute to the binding free energy of the ligand as much as a weakly bound molecule might, the assumption is made that all molecules in the system will be displaced. Work by Barillari has shown that this is not the case, since water molecules can and will form complexes between the protein and the ligand under the right circumstances.[40] Coupled with the semi-empirical nature of the scoring function this suggests that any values obtained from the function should be treated with great care, especially if there is the possibility for the formation of a ligand-water complex.

2. **Ignoring water-water entropy** - For systems where WaterMap predicts several water molecules bound to each other and the protein, only the water-

solute terms are used to derive the entropy term. The justification for ignoring the water-water entropy term lies in the assumption that the entropy which a water molecule will exhibit in a well hydrated binding pocket is similar to that which it will experience in the bulk. Although this assumption is fair in cases where the binding site is mobile and solvent accessible, for poorly hydrated binding sites this term is likely to be a potential source of error. Considering that recent studies have highlighted the critical role of water networks in protein binding sites, alongside the possibility of the water molecules being more entropically favourable within the pocket rather than the bulk [42, 43], it is perhaps surprising that this term is neglected in the calculation.

3. **Clustering water positions from a MD simulation loses orientational information** - Whilst observing the positions of water molecules during a MD simulation will give information on the most favourable positions within the system, any correlation between different molecules is lost once this information is clustered. As such, when molecules are placed upon the potential hydration sites there is no guarantee that this is a realistic representation of the solvent packing. If the chosen sites are incorrect then the estimated energy and entropy of the site will also not be correct.

4. **Are water free energies additive?** - In the WaterMap scoring function it is assumed that the free energy of liberation of a binding site will be equal to the sum of the individual energy and entropic terms across of waters in the presence of each other. This assumption will only hold true if all of the water molecules leave the binding site at the same time. This is unlikely to be the case, since the more weakly bound waters in the system are likely to

be displaced before the stronger binders. If these water molecules are removed from the system, then it is likely that the energetics of the remaining waters will change - something for which WaterMap is not able to account. This is particular important when WaterMap predicts a hydration site with a positive free energy, as in a Src kinase study.[70] This site was included in the analysis, although the presence of a highly unfavourable molecule in the system is unlikely. It is reasonable to assume that such a molecule would rather be in the bulk, and that including it in the analysis is likely to lead to erroneous results. However, in order to fully justify this criticism, a thorough understanding of the WaterMap free energies needs to be sought. The WaterMap method retains unfavourable water molecules in the analysis since it is claimed that the cost of creating of a vacuum at physiological conditions is extremely high in the condensed phase [69], although it can be argued that the driving force for a water molecule to reside in the protein is solely down to the binding free energy of the water - something which WaterMap does not calculate.

3.6 SZMAP

The SZMAP method is a tool developed by OpenEye to determine the position of water molecules within a protein binding site. Unlike other methods it uses a semi-continuum model, whereby an explicit water is used to probe the protein binding site whilst surrounded by an implicit continuum solvation model, modelled with a dielectric constant, ϵ. The water probe is moved around a grid, where at each grid point the energy of the water orientation and position is calculated. This is then used to determine favourable hydration sites. Two different types of probe can

be used; a traditional charged water probe and uncharged water probe, designed to represent hydrophobicity. The energetic difference at the same position using the two probes can then be used to determine whether this water is likely to be displaced by a polar or hydrophobic group on a ligand.

Although the SZMAP methodology has been presented at a number of conferences, there are no publications relating to the method. As such, it will not be considered for full comparison in this thesis.

3.7 Double decoupling method

The double decoupling method states that the absolute binding free energy of a substrate, S, to a receptor, R, can be found by performing two separate simulations.[14] The first of these involves decoupling the substrate from the bulk solvent, whilst the second of these decouples the substrate from the receptor. This can be visualised in a thermodynamic cycle, as in Figure **??**.

$$\begin{aligned} S_{sol} &\rightarrow S_{gas} & \Delta G_{hyd} \\ RS_{sol} &\rightarrow S_{gas} + R_{sol} & \Delta G_{dec} \\ R_{sol} + S_{sol} &\rightarrow RS_{sol} & \Delta G_{abs} = \Delta G_{hyd} - \Delta G_{dec} \end{aligned} \tag{3.11}$$

For the decoupling of water molecules from a protein binding site, the first term in Figure **??** refers to the simulation where a water molecule, S, is decoupled from the bulk, and is equal to the hydration free energy of water, ΔG_{hyd}. The second term requires the water molecule to be decoupled from the complex, R, and is equal to the decoupling free energy, ΔG_{dec}.

The process for decoupling a water molecule from the protein receptor is dependent upon the standard state of the system. For water, the standard state is known to be 55.56 M. As such, the value of ΔG_{dec} derived from the simulation needs to be corrected to arrive at the correct free energy. The overall expression for the decoupling of a water molecule from the protein can be expressed as equation **??**.[14]

$$\Delta G_{dec} = \Delta G_{comp} + \Delta G_{rest} - RTln\frac{\sigma_{RS}}{\sigma_R \sigma_S} + P^0(V_R - V_{RS}) \tag{3.12}$$

In equation **??**, ΔG_{comp} is the computed free energy for the decoupling the water molecule from the protein. This computed free energy is equal to the sum of decoupling the electrostatic terms of the molecule and decoupling the Lennard-Jones terms. ΔG_{rest} is the free energy of the restraint applied to the water molecule during the decoupling simulation. Such a restraint is required to guarantee reversibility since, if such a restraint is not applied, the molecule could leave its original position and drift away from the simulation region.

Several different types of restraints have been used in the literature to look at the binding free energies of water molecules. For example, studies by Zhang [71] and Olano [72] have applied a harmonic potential on the water molecule in question, whilst a study by Barillari [40] utilised a hardwall potential. It has been demonstrated that the calculated free energy is independent of the applied restraint/constraint.[8]

The third term in equation **??** is a symmetry related term. σ_{RS} is the symmetry number of the complex, σ_R is the symmetry number of the protein and σ_S is the symmetry number of water. Water has a symmetry number of 2 and, since the

other two terms have a symmetry of 1, the term can be found to be - 0.4 kcal/mol. The final term in equation **??** is taken to be negligible under standard pressures since the change in pressure can be taken to be miniscule.

3.7.1 Applications of the double-decoupling method

Being a well established and rigorous technique, there have been numerous studies where the authors have used the double-decoupling methodology. Early studies by Roux [73] and Helms [74] used a statistical thermodynamics approach similar to the double-decoupling method, but the first true application of the double-decoupling for the calculation of water-protein binding free energies was performed by Hamelberg.[75] In this study, the authors looked at the binding of a water molecule in the binding pocket of trypsin and HIV-1 protease. It was found that in both cases the binding was favourable, and it was suggested that the release of these water molecules into the bulk could be favourable. In addition, the effect of the harmonic potential upon the restraint term was examined. The study showed that the best results are obtained with a weaker restraining potential, providing that the correct region of configurational space of the water is sampled.

A study by Olano [72] used the double-decoupling method to look at the hydration of two different cavities; the bovine pancreatic trypsin inhibitor (BPTI) and barnase. These two systems are very different, in that one of them is polar, whilst the other cavity is more hydrophobic. The study found that the cavity in BPTI is likely to be hydrated (ΔG_{bind} = -4.7 kcal/mol), whilst the barnase cavity is likely to be empty (ΔG_{bind} = +4.7 kcal/mol). The observed results were in good agreement with earlier simulations performed by Hermans [71], which suggested that water molecules in non-polar cavities are likely to have positive binding free energies.

As previously mentioned, a key study by Barillari looked at a wide variety of water molecules in protein-ligand complexes to ascertain their propensity to be displaced.[40] 6 protein-ligand systems were examined, totalling 54 water molecules, with these waters having their hydration free energies found by double-decoupling. Based upon the results obtained, two different types of water molecules were identified; those conserved and not displaced by ligands, and those which are displaced. A Bayesian statistics approach was applied on the system, finding that, on average, those molecules which are more tightly bound are less likely to be displaced. A statistical model was also used to predict the probability of a water molecule being displaced by a ligand, given its binding free energy.

A 2011 paper by Fadda and Woods looked at using the double-decoupling method to the role of a conserved water molecule in Concanavalin A.[36] Using three different water models (TIP3P, TIP4P and TIP5P), they looked at the binding free energy of the water molecule in the apo protein and two different ligands. Surprisingly, they found a large dependence on the water model for predicting whether the molecule was conserved or displaced. Indeed variances in the binding free energy by 5 kcal/mol were observed, which were attributed to a shift in the Lennard-Jones decoupling stage. These variances are not however consistent; TIP5P sometimes appears to be the outlier in the data set, whilst in other simulations it is TIP4P which appears to the outlier. No firm energetic rationale is given for these changes. It is signficant to note that the wrong standard state and harmonic correction term was used in the study, which will result in a equal shift in the binding free energy for all the waters in this study.

3.8 Evaluation of currently available methods

In the previous sections, four major techniques for identifying and scoring water molecules have been described; GCMC, JAWS, WaterMap and double-decoupling. All of the methods have delivered results which are in good agreement with experiment, although they all have their drawbacks. Although theoretically rigorous, the double-decoupling method requires significantly more simulations to be performed than the other methods, since it requires a full examination of the thermodynamic pathway. The method also requires prior information on where the water molecule is, before the simulation is performed. As previously mentioned, the resolution of the crystal structure is often a limiting factor in locating waters in protein binding sites, meaning that the environment for the water may not be correctly defined. As such, the calculated free energy could be incorrect if the system is not fully and accurately defined.

The GCMC method, although having been around for 40 years, has not yet been fully exploited in terms of locating and scoring water molecules. Although the method has been used to account for fluctuating numbers of molecules across a boundary, such in the GSBP scheme, no studies have been performed where the binding free energy of a single water molecule is determined. In theory this should be possible, since the chemical potential of the system can be directly linked to the decoupling free energy and, therefore, to the binding free energy of a water molecule. One major drawback of the method is the poor acceptance rate, which is likely to introduce an error into the calculations.

Being a much more recent method than GCMC, the JAWS methodology has significantly fewer citations in the literature than the GCMC approach. As such, there has not yet been a full study assessing the limitations of the technique. Al-

though the method has shown great promise in both locating and scoring water molecules, it is incapable of calculating the binding free energy of strongly bound waters since the biasing potential is not sufficient to induce enough low θ sampling. In addition, the current method of using a clustering algorithm to locate waters is likely to be problematic in cases where water molecules can adopt a wide number of configurations.

Finally, the WaterMap methodology has been extensively covered in the literature as a strategy for locating water molecules in protein systems, as well as providing an estimate for their enthalpy and entropy. Although the results have been encouraging in qualitatively identifying ligand trends, the method has several drawbacks, being the use of clustering to locate waters, assuming that a ligand will displace bound waters and that the free energies of waters in a cluster are additive.

3.9 Conclusions

This chapter has looked at the importance of water molecules in protein binding sites, and the major computational methods available for locating and scoring them. Water molecules have been traditionally thought to play two major roles in the binding site, namely stablising the structure by forming a hydrogen bonding network, and being displaced by a ligand to provide an entropic boost to the binding free energy of the ligand. More recent studies have highlighted a plethora of roles which water plays in protein-ligand binding; whether it be sub-optimally hydrating the binding pocket to allow for enthalpic-based binding events, providing subtle electrostatic screening between protein and ligand, or forming part of a network which dictates ligand affinity. Regardless of the effect in which the medicinal chemist is interested in, the fundamental axiom is that water plays an

active role in biomolecular recognition and should be correctly treated to obtain high quality predictions.

The currently available methods have all shown promise in locating water molecules in protein binding sites, yet not all of them have been exploited fully. In particular, the GCMC and JAWS methods have not yet been fully tested and applied to novel systems. Another factor which has never been addressed is the consistency of the methods. For example, will the water locations predicted by JAWS be the same as the predictions given by GCMC? Equally, will the binding free energy calculations given by double-decoupling be the same as for GCMC?

From what has been previously discussed, it has emerged that there is currently no freely available computational method which has been proven to predict both the location of water molecules, and also providing a reliable free energy estimate, regardless of the binding free energy of the water molecule. In addition, no study has been performed which critically compares and contrasts the different methodologies to see if they give consistent results.

Based upon this, 4 initial aims were formulated regarding the placement and scoring of water molecules:

1. To adapt the JAWS methodology to calculate the binding free energy of strongly bound water molecules.

2. To determine whether JAWS, GCMC and double-decoupling all predict the same location and binding free energy of different water molecules.

3. To review the three methods to determine which is better for a particular problem.

4. To apply the methods to novel systems to further explore the strengths and limitations of each method.

Chapter **??** describes the work which has performed to achieve these aims.

Chapter 4

Fragment-Based Drug Discovery

4.1 Introduction

The following chapter describes the basic theory behind Fragment-Based Drug Discovery (FBDD), and how it is performed experimentally with examples drawn from the literature. An overview of the different computational methods available for performing FBDD is then presented, with the relative merits and drawbacks of each method assessed. The Grand Canonical Monte Carlo method for performing FBDD is then described, highlighting its differences with the other available methods. Finally, a critical review of all of the approaches is presented, highlighting the need for a new direction to be taken.

4.2 Why use FBDD?

A widely used approach in drug discovery is High-Throughput Screening (HTS), where a large range of molecules are screened against a particular drug target. One problem which has been identified with HTS is the poor hit rate. Since the

molecules which are typically screened in a HTS array are large ($>$ 300 Da), the chances of a good match between the chemical groups on the ligand and receptor is low, leading to hit rates of the order of around 1 %. Another consequence of screening such large compounds is that the optimisation of the hits typically involves an increase in both molecular weight and lipophilicity. This means that the hits typically will violate some of Lipinski's rules of 5 [11] and will provide a barrier to getting a drug to the market.[76]

One strategy to get around the HTS problem lies in FBDD. In FBDD, low molecular weight compounds (typically $<$ 250 Da) are screened against the protein target. Since the compounds are smaller than those screened in HTS, they tend to bind to the protein target more weakly. In order to detect the binding of the fragments, sensitive biophysical techniques such as NMR, surface plasmon resonance (SPR) and X-ray crystallography are used.[4] The resultant hits from FBDD can then be modified into suitable lead compounds.

Broadly speaking, there are three major advantages for utilising FBDD over standard HTS. The first advantage lies in the ability of FBDD to probe the chemical space more efficiently. It has been reported that the number of drug molecules with 30 or more heavy atoms is 10^{60}.[9] As such, attempting to sample this region of chemical space with HTS is both a challenge and inefficient. In comparison, the number of base fragments which make up the larger compounds is thought to be 10^7. It therefore becomes more efficient to screen the fragment chemical space than a standard HTS process. A recent review conducted by scientists from Astex demonstrated that, for a study on HSP90, 1600 fragments were used to probe the chemical space, with an average molecule weight of 170 Da.[77]

The second advantage is that, since fragments are usually significantly smaller and less complex, the probability of finding a successful hit is much larger in

FBDD compared to HTS.[78] For a binding event to occur the fragment must overcome a significant entropic barrier, with this barrier shown to be independent of the size of the fragment. As such, for a fragment to bind to the protein, the interactions formed between the fragment and the protein need to be of sufficiently high quality. The binding events which occur in FBDD are typically weaker than HTS, but the binding affinity per heavy atom is generally larger compared to HTS. For fragment hits to be of use in lead development, it is hoped that the interactions formed between the fragment and the protein are maintained when the lead compound is developed. Studies by Astex have shown that, for 39 fragment-lead campaigns, 80 % of the starting fragment interactions were retained in the lead compound.[77]

The ratio of binding affinity per heavy atom, expressed as the ligand efficiency, is the third major advantage of FBDD. Ligand efficiency allows fragments of different molecular weights to be easily compared, making it easier for the medicinal chemist to choose the optimum fragment. Careful development, and attempting to maximise the ligand efficiency, should give a lead which is more likely to obey Lipinski's rules and displays improved potency and, potentially, specificity.[4] Alongside ligand efficiency Astex have developed a ligand lipophilicity efficiency (LLE) measure, which allows the tracking of the lipophilicity through compound development. This measure is defined as the pIC_{50} - LogP of the compound. Starting from a fragment which has a high LLE gives the medicinal chemist control over the C*log*P during development, with evidence showing that fragment-derived compounds typically have a C*log*P of one unit less compared to HTS compounds. This is advantageous, since it has been shown that drugs which are more lipophilic are typically more promiscuous and more likely to display toxicity than those which are less lipophilic.[79] If the fragment hits are carefully optimised in the

drug discovery process, the resulting lead compounds will be both smaller and less lipophilic than drug molecules found through HTS.[77, 80]

4.3 Experimental FBDD

As previously mentioned, the three major biophysical techniques used to detect fragment binding are NMR, SPR and X-ray crystallography. A brief review of these techniques is now presented.

4.3.1 NMR spectroscopy

This technique works by one of two strategies.[81] The first involves looking at the chemical shifts of the target both prior to and during screening. If a binding event is present, then a shift in δ is observed. The second strategy involves looking at the rates of translation and rotation of the free ligand during binding. Upon binding, these rates will be reduced, since there is a decrease in the degrees of freedom of both the protein and ligand. NMR fragment screening has been used by Wang *et al.* [82] to look at fragment binding to BACE-1, a protein target which is believed to be associated with the onset of Alzheimer's disease. 2D ^{15}N-HSQC NMR was utilised, with positive hits identified through a large change in the chemical shift of near aspartic acid residues. Fragment NMR spectroscopy has also been used by Murray *et al.* [83], where the development of Hsp90 inhibitors was described. The hit to lead project created a lead compound with a 10^6-fold increase in potency from the initial hit, with the addition of only 6 heavy atoms.

4.3.2 X-ray crystallography

Fragment detection with X-ray crystallography works by soaking cocktails of fragments at high concentration, and then observing the position of the fragment in the target using X-ray crystallography.[84] One advantage of this method is that it is able to distinguish between different ligands, increasing throughput in the assay. In addition, it provides direct evidence for binding between fragment and the target.[85] One drawback in the method, however, lies in the fact that it is not capable of screening as many fragments as NMR.[10] X-ray crystallography is commonly seen as a gold standard in corroborating fragment hits, and as such is used in the vast majority of fragment studies. Recent studies such as the previously described work by Murray [83], the development of PDK1 inhibitors [86], and the discovery of allosteric inhibitors of farnesyl pyrophosphate synthase [87] all show the effectiveness of the method.

4.3.3 Surface Plasmon Resonance

SPR biosensors identifies fragment binding by looking at changes in optical refraction. Upon a fragment binding to a protein target, a change in the refractive index (RI) of the system is observed. A measure of the binding strength between the fragment and the protein is achieved by looking at the times to record the maximum RI and the subsequent relaxation back to the original RI as the fragment is removed from the protein by a flow system. Since the method does not give as much information as X-ray crystallography and NMR, there are significantly fewer studies where this method has been published. However, a 2010 study by de Kloe *et al.* used the approach to look at the binding of fragments to acetylcholine binding proteins, allowing the identification of hot spots in the protein.[88]

A study performed by Evotec demonstrated that all three methods are capable of identifying fragments in the same ligand efficiency range, suggesting that the rationale for choosing one method over another lies in the information desired or cost, rather than identifying more potent fragments.[80]

The wide number of examples in the literature relating to FBDD highlight how important it has become to drug discovery. Despite this, the timescales required for FBDD are much higher than standard HTS, since the fragment hits are often non discriminating and require significant effort to turn the fragment into both lead and final compounds. In addition, the experiments required to run the assays are typically expensive, meaning that it is important to ensure that the correct fragments are screened. Such a procedure could be performed computationally, performing the dual role of screening more fragments virtually and screening the optimal fragments experimentally.

4.4 Computational approaches to FBDD

Given the cost of performing FBDD experimentally, computational approaches are ideally suited to enhancing FBDD. Indeed researchers at Evotec have advocated using computational methods to study fragment binding to GPCRs, owing to the cost and complexity of obtaining high quality crystal structures.[89] A wide range of techniques have been reported in the literature, a selection of which are now reviewed.

4.4.1 Docking

Whilst traditionally performed for larger, lead-like molecules, docking has also been performed upon fragments.[90] In one study, the docking program Glide [91]

was used to dock fragments to two systems; prostaglandin D synthase (PGDS) and ligase. For the ligase study a 20K generic fragment library was used, with 794 actives present in the library with $IC_{50} \leq 1$ mM. Various different restraints were applied upon the hydrogen bonding to see if any improvement in docking accuracy was observed. The enrichment factor at 1% of the database was 3.3 when no restraints were applied, and application of various restraints resulted in an average enrichment factor of $\sim$3.1. This shows that there was no significant improvement in performance when restraints were applied in this case.

For both ligase and PGDS, use of the GlideSP protocol gave enrichment of actives over random sampling. Indeed, the enrichment rates of ligase were within the ranges which might be obtained for lead molecules. [90] This suggests that, in principle, fragment screening using docking methods should be possible. However, the authors note that such protocols have yet to be fully optimised, primarily due to limitations in the scoring functions used.[92] A similar conclusion was reached by Verdonk *et al.* in a docking study comparing the differences in docking performance between fragments and drug-like compounds.[93] It was found that the performance of the fragments and drug-like compounds was equally poor, but for different reasons. Whilst the performance of drug-like compounds can be attributed to poor protein sampling,[94] the inefficiency of scoring functions to describe fragments with low ligand efficiencies was identified as a reason for the poor fragment performance. It was noted that better performance was obtained when fragments with higher ligand efficiencies were studied, due to higher quality protein-ligand interactions.

4.4.2 The Multi-Copy Simultaneous Search Methodology

The Multi-Copy Simultaneous Search (MCSS) methodology [95] has proved to be a useful tool for probing targets with known structure.[96] The technique is primarily used to locate energetically favourable positions of fragments in the pre-defined binding site of the target. The methodology works by randomly distributing replicas of a fragment, typically around 5000-10000, in the binding site of the target. These fragments are not allowed to see each other, yet each one is allowed to fully interact with the target receptor. The fragments are then minimised simultaneously by conjugate gradient minimisation. Fragment replicas which begin to converge to the same position with a RMSD of less than 0.2 Å are removed to leave one replica in that site. The MCSS methodology typically picks up a large number of energy minima, which makes it ideal for finding favourable binding positions for that particular fragment. However, a consequence of this is that it is difficult to assess which minima are significant, a feature which is important in 'hot-spot' determination.[97]

Once minimisation has finished, the resulting fragment-protein minima are examined. The first stage of this post-processing is to remove minima with energy above a certain threshold. This threshold is usually based upon the solvation enthalpy of the functional group, which takes into account that the fragments within the binding site need to be desolvated.[96] In theory, this effect can be included by performing the minimisation process in implicit solvent. Original studies showed that errors could occur using this strategy, since the electrostatic interactions are often overstated.[98] However, a more recent study has demonstrated that rescoring the MCSS poses with MM-GBSA can lead to good agreement with X-ray crystallography.[99]

The MCSS method suffers from a few technical limitations. The first arises

when the fragment size is increased to around 20-30 atoms. If such a fragment has rotatable bonds, then finding the energy minima becomes complex for different conformations. Enough replicas must be included in the simulation to sample all of the different possible conformations, or undersampling can occur.[96]

The second limitation is the reliable incorporation of protein flexibility. Many proteins do not exist solely as a rigid structure, and hence to accurately describe the system it is desirable to take account of this flexibility. One method of attempting to include flexibility is by having multiple protein input structures, each one with a different conformation. Integration of the fragment maps for each protein conformation can then provide a method for accounting for the protein flexibility.[96] However this method is inherently limited by the number of protein structures available. In addition, for particularly flexible proteins, the question of how many structures are required to sample the entire structure ensemble arises. A final limitation in the methodology is that the methodology does not consider the role of water competing with the fragments.

4.4.3 FTMAP

FTMAP is a probe-based interaction energy technique which has been specific designed to look for 'hot spots' on the target.[97, 100] The method exploits Fourier transformations, which allows billions of probe positions to be positioned on rotational and translational grids. Simple energy expressions are then calculated to establish whether the site has a favourable interaction energy with the protein.

The algorithm can be broken down into 5 steps, detailed below:

1. **Fragment-docking**. 16 different fragments are chosen and are tested on the grids. Both the fragments and the protein are treated as rigid. The 2000 most

energetically favourable poses for each fragment are then taken forward to the next stage.

2. **Minimisation**. The interaction energy of each of the protein-fragment complexes are then minimised. In the FTMAP software, the CHARMM potential is used, with electrostatics treated using a Poisson-Boltzmann approach. Minimisation is performed, allowing the fragment to move whilst the protein is fixed. The fragments which have been docked in this stage and the previous stage can then be used as the building blocks for further fragment studies.

3. **Clustering**. For each fragment, the lowest energy conformer is chosen. Fragment conformers within 3 Å of this minimum are then clustered. The next lowest energy conformer is then chosen, and the clustering algorithm is again applied. This process continues until all of the fragments have been clustered. Clusters with fewer than 10 fragments are then removed, with the remaining clusters ranked on the average energy of the cluster.

4. **Determination of consensus sites**. A consensus site is a position in the protein where there is overlap of different fragment types. The cluster with the most fragments is chosen as the initial site, and all clusters within 5 Å of it are joined together to form the first consensus site. These are then removed from consideration. The cluster with the next most fragments in it is then chosen as the second consensus site, and clustering is then applied. This is repeated until all fragment clusters have been assigned.

5. **Binding site characterisation**. The first consensus site is generally the site which is the most important hot-spot in the protein.[100] All consensus

sites within 7 Å of the first consensus site are then used to describe the binding pocket of the protein. A second clustering is then applied, so that any consensus sites within 7 Å of a consensus site already in the binding pocket are included. This process continues until no further expansion is possible. This then forms the hot-spot of the protein binding site.

The FTMAP algorithm has been compared alongside the experimental Multiple Solvent Crystal Structure (MSCS) technique [84] to see if the same results are obtained. For the the glucocerebrosidase protein, a therapeutic target for Gaucher's disease, the two methods corroborated the same binding hot spot. In addition, other potential binding sites were identified using the FTMAP method. One drawback of the technique is that since the protein is kept rigid throughout the simulation, there is conformational dependence upon the results obtained. In a 2010 study by Ivetac and McCammon [101], the authors looked at mapping the allosteric space of GPCRs. Recognising that the system is extremely mobile, an ensemble of 15 input structures were generated via a MD simulation. These structures were then run through the FTMAP server, with the results analysed by looking at the interactions between a probe and a set residue and then building a probability map. One other drawback of the FTMAP method is that it does not take into account the role of water solvent in binding competition.

4.4.4 Site Identification by Ligand Competitive Saturation

Site Identification by Ligand Competitive Saturation (SILCS) [102] is an explicit solvent all-atom molecular dynamics method which has been designed to overcome some of the issues associated with computational fragment-based methods. Methods such as MCSS [96] and FTMAP [97, 100] are limited in their calculation

of fragment-protein affinities since protein flexibility and solvation is generally ignored or approximated. SILCS works by computationally immersing a protein in an aqueous solution and a collection of different fragments all at a concentration of $\sim$1 M. The protein, water and fragment system is then subjected to multiple MD simulations, which enables competitive binding to take place. The resulting snapshots are combined to form fragment probability maps which reveal where particular fragments prefer to bind.

Since the snapshots are generated from MD simulations, the SILCS output takes into account protein flexibility. The fragment maps which are generated give a Boltzmann distribution of conformations and have atomic-level solvation effects. As a result the fragment maps represent rigorous free energy distributions, something which is typically not observed in fragment-based approaches. These maps can either then be used as qualitative tools to assemble potential inhibitors or used for docking applications. The method has been used upon the oncoprotein BCL-6 and has been shown to reproduce the crystallographic binding modes of the SMRT and BCOR peptides.

A more recent application of the method looked at using the method on 7 different protein systems and attempted to create a more quantitative analysis.[103] MD simulations on the protein systems were performed, with benzene and propane used as probes alongside water. At the end of the simulation, a grid-based probability distribution of the fragment heavy atom positions is generated, termed a 'fragmap'. A grid free-energy term is then calculated, based upon the probability of finding the fragment at a particular site in the simulation and the probability of finding the fragment in a bulk system of water. A free energy comparsion is then made by summing up the grid free energies over the volume which a ligand occupies in the protein and comparing it to the experimental value of ΔG.

The observed results showed that the fragmaps correctly predicted the binding locations of functional groups in the examined ligands. However the grid free energy comparison was less successful, with the correlation between the experimental values and the grid terms modest at best. This is likely to be due to the fact that only simple fragments were used in the study.

There are two major limitations of the SILCS approach. Firstly, although the method allows solvent competition in the binding site, it does not take into account the desolvation of the fragments. As such, the method can only predict where fragments will bind in the system, and not predict whether they will actually be able to leave the bulk. Second, the method requires the fragments to physically pass by each other in the binding site. This is likely to be inefficient, especially in the presence of bulk water, and will inevitably lead to sampling issues and inefficiency.

4.4.5 3D-RISM

The 3D-RISM method is one which utilises integral equations to predict the location of molecules around a large macromolecule, such as a protein.[104] The method generates distribution functions of the species of interest by looking at the solute-solvent interaction potentials, based upon standard forcefields. By looking at these functions in three dimensions, the variance from the bulk solvent density can be found, giving an indication of whether or not that site is likely to be hydrated.[105]

The application of 3D-RISM to FBDD has recently been reported [104], whereby the fragments are examined simultaneously alongside water. By solving the integral equations for the species, the positions of fragments around the protein surface can be easily found, whilst including the effects of solvent competition.

The method was initially used upon the binding of isopropanol and acetone to thermolysin, with the results highlighting both the correct binding modes of the fragments and also cooperative water effects.

The 3D-RISM is theoretically rigorous, although it does have a few drawbacks. The method cannot take into account protein flexibility, so an ensemble of structures is required if the protein undergoes significant movement upon ligand binding. Second, no indication of the relative binding strength of each fragment is found in the method, meaning that it is predominantly a qualitative tool. Finally, as with many other computational fragment-based approaches, no estimate of the fragment desolvation cost is included in the method.

4.5 Calculation of binding free energies of fragments using GCMC

The use of GCMC in calculating the binding free energy of fragments is appealing, since in theory it should provide information on both the strength and pose of binding. Despite this, there have been relatively few publications detailing methods for observing protein-fragment GCMC binding. The first example [106] involved looking at the binding of fragments to T4-lysozome and thermolysin using the simulated annealing approach of Mezei.[54]

For the T4-lysozome case, the study looked at the binding of benzene derivatives to T4-lysozome. The authors adopted the cavity-bias methodology [51] to increase the probability of accepting insertions. Simulations were started at a value of B = -15, and the value of B was annealed stepwise in increments of 1B until all of the ligands left the system. 4 million GCMC steps were carried out for convergence, with a further 1 million steps carried out for data collection. A

similar set-up was used for thermolysin, except in this case acetone derivatives were explored.

In the T4-lysozome case, at high values of B, a high concentration of ligands were observed around the protein, resulting from attractive interactions between the ligands in the bulk. As the value of B was gradually reduced, the weaker binding fragments left the system. A point was reached where the average number of ligands in total was less than 0.1; at this point the simulation was terminated. The point immediately prior to the disappearance of the ligand was determined as the free energy of binding for the molecule. In the study, it was necessary to correct for the solvation energy of the ligand by calculating the GB/SA solvation energy. From this, a solvation corrected binding free energy was obtained. For benzene, the binding free energy was determined as -9.6 kcal/mol, against a value obtained using FEP [107] of -7 to -9 kcal/mol. The authors used a different force-field to the FEP methodology, which could account for the difference in estimated binding free energy.

In the thermolysin cases, the observed ligand poses from the GCMC method were compared to an analogous experiment which had been performed previously using the MCSS method. It was found that the poses which were observed using the GCMC method were consistent with those using the MCSS.

The second use of GCMC to calculate the binding free energies of fragments focuses solely upon the T4-lysozome test case using benzene [108]. Whereas previously the simulated annealing approach was used [106], here two different techniques are used, both of which are faster and are more rigorous.

4.5.1 Non-Interacting Particle Method

The first method involves flooding the system with ligand molecules which are not allowed to interact with each other, allowing them to 'ghost' past each other. This is carried out at a number of different chemical potentials. By calculating the concentration of ligand molecules, [$L_{simulation-cell}$], around the protein, the binding free energy is found using equation **??**, below:

$$\Delta G = -k_B T \ln \left(\frac{[L_{simulation-cell}]}{[L_{ideal}]} \right) \tag{4.1}$$

[L_{ideal}] is found is by noting that, for an ideal gas at equilibrium, the following relationship holds true:

$$\frac{N_i}{V} = \frac{N_0}{V_0} \exp(B) \tag{4.2}$$

Equation **??** shows that the concentration at a set value of B depends on the standard concentration, N_0/V_0, of species *i* and the perturbing B value. As such, the ideal reference can be easily found for any value of B, allowing the free energy to be found using equation **??**. Since the free energy of the system can be directly found from the concentration of the reservoir and the simulation, the free energy can now be found from a single simulation run whereas for the simulated annealing approach several simulations were required. In addition, the free energy for any smaller region in the entire simulation can be found by simply looking at the number of ligands in that region.

In that work, the free energy of benzene in the binding pocket was observed. This technique was performed at several different values of B, and the same binding free energy was found at each level. However it was also found that this technique works most efficiently at lower reference concentrations, correspond-

ing to a lower value of B, since equilibration times increase with the number of ligands in the system. The binding free energy was found to be -9.9 kcal/mol, consistent with the earlier study [106]. One drawback of this method is that it is only suitable for small and well-defined binding pockets. If the aim is to look over the entire protein for binding hot spots, or to find other binding poses in the active site, then the equilibration times are often not practical.

4.5.2 Interacting Particle Method

The second method involves allowing the ligand molecules to see each other during the simulation. Only ligand-ligand repulsion is allowed in this study; attraction terms are switched off. Unlike the other method, the number of ligands observed in the binding pocket is greatly reduced, since they exclude each others' volume. In this study, the maximum number of ligands observed was one. The value of B is scanned to identify simulations where the average population in the binding pocket is less than 1 but also statistically significant. One drawback of this method is that is if a high affinity pose is accepted, then subsequent removal attempts are likely to be rejected. This prevents potential poses of equal or lower energy from being sampled, meaning that the system will not reach true equilibrium. When a different value of B is simulated, all of the ligands are cleared from the system to minimise this error from the equilibration. Using equation **??** the authors found that the same result was found using the other method (-9.9 kcal/mol) once the average population in the binding pocket became sparse. In addition, the equilibration times were often far shorter due to far fewer ligands in the system.

As with the initial paper [106], the solvation energies of the fragments in the second study are found using a GB/SA model. The authors state that the calculation of these solvation energies are a potential source of error in the calculations;

for benzene the difference between the experimental and GB/SA solvation energies is 0.80 kcal/mol. Despite this, the method gives reasonable results compared to FEP methods.

4.5.3 Limitations of GCMC

Although the GCMC methodology has been shown to be effective for locating and calculating the affinity of fragment poses, there are a number of drawbacks to the method. The described method runs the simulation in the gas phase, meaning that the resultant gas phase free energies need to be post-processed.[109] Similarily, the fact that the simulations are run in the gas phase means that there is no solvent competition during the simulation. Protein flexibility is prohibited during the simulation, meaning that ensemble-style methods are required. Although this is achievable, questions such as the number of structures required to adequately sample the chemical space arise. As previously mentioned, one of the major limitations in GCMC lies in the poor acceptance rate. Since a typical fragment is significantly larger than a water molecule, this will serve only to exacerbate the problem.

4.6 Critical appraisal of computational methods

The previous sections have described the current computational methods available. Whilst all of the methods are capable of delivering results that are consistent with experiment, most of them suffer from at least one major limitation. Table **??** briefly summarises the advantages and disadvanatges of the methods.

Table **??** shows that most of the methods suffer from a lack of protein flexibility and a correct treatment of solvation. The lack of protein flexibility is par-

Method	Advantages	Disadvantages
Docking	High throughput	Fragments and protein are rigid. Scoring functions are not sufficiently accurate enough. No solvent competition. No fragment competition
MCSS	Fast. Well documented technique.	Fragments and protein are rigid. No solvent competition. No fragment competition.
FTMAP	Extremely fast. Can be run on online server.	Fragments and protein are rigid. No solvent competition. No fragment competition.
SILCS	Fragments can compete with each other and water. MD approach ensures that protein conformations are sampled.	No desolvation estimation. Method requires fragments to pass by each other in solvent.
3D-RISM	Includes atomic level solvation effects. Fast.	No protein flexibility. No estimate of fragment binding affinity. No desolvation estimation.
GCMC	Well documented. Can rank fragments based on binding free energy	No protein flexibility. No solvent competition. Acceptance rates are poor for large fragments.

Table 4.1: The advantages and disadvantages of the currently available computational FBDD methods

ticularly limiting, since it is well recognised that proteins can undergo conformational change upon ligand binding.[77] A 2012 study by Astex found that 50 % of fragments induced a 5 Å RMSD shift in the protein backbone in a sample of 25 targets, highlighting that it is vital such effects can be incorporated. Only a few of the methods are capable of incorporating fragment and water competition, something which occurs implicitly in most fragment assays. From looking at the

limitations of the current methods, for an approach to be truly effective in FBDD it must have the following characteristics:

1. The method should locate and rank fragments based on their binding free energy;
2. The method should use accurate energy functions, based on a simulation approach;
3. The method must allow competition between different fragments, and critically, water;
4. The method should include an estimate for the desolvation of fragments;
5. The method should allow protein and fragment flexibility;
6. The method should be reliable and efficient.

From looking at the above list, it is apparent that none of the methods fulfil more than four of the above criteria, clearly emphasising the need for new methodology. An approach which attempts to incorporate all of the criteria will be described in a later chapter.

4.7 Conclusions

In this chapter, the theory behind FBDD was discussed. Since FBDD uses signficantly smaller sized compounds than HTS, the chemical space explored in FBDD is larger, meaning that the method is much more efficient. In addition, since the compounds are much smaller, the probability of finding a good match between the fragment and the protein is greater, meaning that the hit rate is much improved

for FBDD.[78] As such, most pharmaceutical companies employ FBDD as part of their drug discovery programs.

Experimental methods such as NMR, X-ray crystallography and SPR are commonly used to detect fragment binding in assays. Although these methods can give high quality results they are often expensive to run, meaning that it would be desirable to have a method to pre-scan possible fragments. Such a method is ideally suited to computational approaches, since they are typically much faster and cheaper than their experimental counterparts.

All of the existing computational approaches suffer from at least two key drawbacks; commonly the lack of protein flexibility and a reliable incorporation of solvation. Since these factors directly influence the free energy of a fragment binding to a protein, it is imperative that they are included in a computational FBDD method. Six key criteria have been identified for a computational method to best mimic that of experiment and to provide high-quality, yet rapid, predictions. The application of the JAWS algorithm in this context will be discussed in a later chapter.

Chapter 5

Predicting the location and binding affinity of water molecules

5.1 Introduction

The following section details the work performed on comparing and contrasting three of the previously defined methods for locating water molecules and calculating their binding affinity; double-decoupling Monte Carlo, JAWS and GCMC. The development of the JAWS algorithm to predict the binding affinity of strongly bound water molecules is initially described, with this methodology applied to the N9-Neuraminidase test system. Drawbacks of each of the three methods are then highlighted via a series of examples based upon cavities in the bovine pancreatic trypsin inhibitor system. Finally conclusions are drawn between the three methods to identify the optimal approach for a particular problem.

5.2 N9-Neuraminidase

5.2.1 Biological Relevance

N9-Neuraminidase is an enzyme which is actively targetted in the treatment of influenza. Influenza viruses bind to the surface of healthy cells via hemagglutinin, a substance which has a high affinity for sialic acid which is present in mammalian cells.[110] The neuraminidase enzyme, present upon the surface of the viral influenza, is involved in the cleavage of sialic acid. Once the virus has infected the host cell, viral neuraminidase cleaves the sialic acid link between the virus and the host, allowing the virus to spread to other, unaffected cells through the release of progeny viruses.

Two major drugs have been developed in the treatment of influenza; oseltamivir (tradename Tamiflu) and zanamivir (tradename Relenza). Both target neuraminidase by displacing the sialic acid from the neuraminidase active site, and hence prevent the release of the virus from infected cells.[111] Whilst both drugs have proven to be successful in the treatment of influenza, new mutations, notably the H274Y mutation in N1-neuraminidase, have proven to be resistant to oseltamivir.[112] As a result there is a need to develop new influenza drugs, something which is of great interest to the pharmaceutical industry.

5.2.2 System Setup

The crystal structure chosen for the simulations was *1nnc* [113] (resolution = 1.80 Å). Polar hydrogens were added onto the structure using whatif [114], with non-polar hydrogens added using LEaP. The zanamivir ligand was parameterised using the antechamber module in AMBER, with the partial charges assigned using the

AM1-BCC [115] model. To reduce the computational cost, only protein residues that have a heavy atom within 15 Å of zanamivir were retained. Crystallographic waters within this region were retained, except for those in the region of interest. The complex was solvated by a sphere of TIP4P [44] water molecules of 23 Å radius centred upon zanamivir. The resulting complex was then equilibrated for 10 million moves in the NVT ensemble to remove bad contacts. For the forthcoming methods, the amber99 forcefield [116] was used, with a temperature of 25 ^{o}C and a non-bonded cutoff of 10 Å. Any ligands used in the studies were modelled using the GAFF forcefield.[22]

JAWS protocols

The JAWS stage 1 simulation was performed upon the entire binding site, encompassing a region of 1100 Å^3. 48 TIP4P [44] JAWS waters were added to to the simulation region, with these molecules allowed to move freely around the grid region for one million moves whilst turned off. Unless stated otherwise, the θ threshold applied for water molecules being classed as 'on' was 0.95. Statistics were then collected on the grid region for 40 million MC moves using a grid spacing of 1 Å, in line with the original JAWS study.[13] The resulting data was analysed using AstexViewer, and each grid point normalised according to the number density of the most frequently observed grid coordinate.[117] During the simulation, the JAWS waters were allowed to move and sample θ, with full sampling of the ligand angles and dihedral and bulk solvent performed. The bond angles and torsions for the side chains of residues within 10 Å of any heavy atom of zanamivir were also sampled, with the protein backbone restrained throughout the simulation. For the JAWS stage 1 simulations, solvent moves were attempted with a probability of 23 %, protein side-chain moves with a probability of 3.6 %

and solute moves with a probability of 0.4 %. Variations in θ_i were attempted with a probability of 50 %, in line with the original JAWS study [13], with translations and rotations of the JAWS waters attempted with a probability of 23 %.

The JAWS stage 1 simulation identified 7 hydration sites which were then used as starting points for the free energy methods.

JAWS stage 2 simulations were performed by placing a 3x3x3 Å^3 grid over the water molecule of interest. The biasing potential, as described in equations **??** and **??**, was turned on, and statistics on the value of θ collected for 40 million MC moves. The hydration free energy of water used in the biasing potential, ΔG_{hyd}, was taken to be +6.4 kcal/mol in line with previous studies.[13, 40] A binding free energy for the water molecule was found from the ratio of probabilities of observing a θ-water at high ($\theta > 0.95$) and low ($\theta < 0.05$) θ values, using equation **??**.

$$\Delta G_{bind}(\text{water, site } i) = -k_B T \ln \left(\frac{P(\theta_i \rightarrow 1)}{P(\theta_i \rightarrow 0)} \right) \tag{5.1}$$

In equation **??**, k_B is the Boltzmann constant and T is the temperature of the simulation. The θ thresholds for the high and low θ states are arbitrary, and are in line with the original JAWS study.[13]

For the JAWS stage 2 simulations, solvent moves were attempted with a probability of 23 %, protein side-chain moves with a probability of 3.6 % and solute moves with a probability of 0.4 %. Variations in θ_i were attempted with a probability of 50 %, in line with the original JAWS study [13], with translations and rotations of the isolated JAWS water attempted with a probability of 23 %.

Double-decoupling protocol

Double-decoupling [14] simulations were performed using RETI [28, 29] and the coordinates found from the JAWS stage 1 simulation. The binding free energy of a water molecule was found in two stages; firstly the electrostatic terms between the water molecule and its environment were perturbed to zero, followed by a gradual linear reduction in the Lennard-Jones terms on the oxygen atom to perturb it to zero. The water molecules were restrained by a hardwall potential of radius 1.8 Å to allow direct comparison with the JAWS hardwall. The hardwall was applied to only the water in question and forbids it from leaving this spherical region. Furthermore, other water molecules, solute atoms and protein atoms were not permitted to diffuse into this excluded region. As shown in equation **??** the volume of this spherical hardwall, V^{eff}, can be calculated to be 24.43 Å^3, which is of similar size to the cubic 27 Å^3 hardwall used in JAWS stage 2 simulations.

$$\Delta G_{rest} = RTln\frac{V^{eff}}{V^{o}} \tag{5.2}$$

In equation **??**, R is the gas constant, T is the temperature of the simulation, V^{eff} the volume occupied by the hardwall and V^0 the standard state volume of water, 29.89 Å^3 at 55.56 M. From this the correction term of the hardwall, ΔG_{rest} for double decoupling simulations can be found to be -0.12 kcal/mol.

For both the electrostatic and Lennard-Jones decoupling simulations, 16 equally spaced λ windows were used with a value of $\Delta\lambda$ of 0.001. The annihilation of both the electrostatic and Lennard-Jones interactions was performed in 40 million MC steps divided into 400 blocks of 100K steps each. Data was collected and averaged over the last 30 million steps for both sets of simulations. At the end of the simulation, the computed free energies for the decoupling of the electrostatic

terms of the molecule and decoupling the Lennard-Jones terms were summed, to give a value for ΔG_{comp}. The free energy of hydration used to calculate the absolute water binding free energy, ΔG_{hyd}, was taken to be +6.4 kcal/mol.

Having calculated the values of ΔG_{comp} and ΔG_{rest}, the binding free energy of a water molecule, ΔG_{abs}, was found using equations **??** and **??**.

$$\begin{aligned} S_{sol} &\rightarrow S_{gas} & \Delta G_{hyd} \\ RS_{sol} &\rightarrow S_{gas} + R_{sol} & \Delta G_{dec} \\ R_{sol} + S_{sol} &\rightarrow RS_{sol} & \Delta G_{abs} = \Delta G_{hyd} - \Delta G_{dec} \end{aligned} \tag{5.3}$$

$$\Delta G_{dec} = \Delta G_{comp} + \Delta G_{rest} - RTln\frac{\sigma_{RS}}{\sigma_R \sigma_S} + P^0(V_R - V_{RS}) \tag{5.4}$$

As previously described, the third term in equation **??** is a symmetry related term. R is the gas constant, T is the temperature, σ_{RS} is the symmetry number of the complex, σ_R is the symmetry number of the protein and σ_S is the symmetry number of water. Water has a symmetry number of 2 and, since the other two terms have a symmetry of 1, the term can be found to be - 0.4 kcal/mol. The final term in equation **??** is taken to be negligible under standard pressures since the change in pressure can be taken to be miniscule.

For the double-decoupling simulations, solvent moves were attempted with a probability of 85.7 %, protein side-chain moves with a probability of 12.9 % and solute moves with a probability of 1.4 %. As with the JAWS simulations, only the bond angles and torsions for the side chains of residues within 10 Å of any heavy atom of zanamivir were sampled.

Unless otherwise stated, error estimates from the double-decoupling simula-

tions were obtained as the standard error across at least three independent simulations.

GCMC protocol - Interacting Particle Method

The GCMC simulations for the individual water molecules were initially performed using the interacting particle method.[108] Unlike the original study, both attractive and repulsive terms were turned on throughout the simulations. Insertion and deletion attempts were accepted using the following Metropolis tests.

$$P_{in} = min\left[1, \frac{exp(B)}{N+1} exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{5.5}$$

$$P_{del} = min\left[1, N exp(-B) exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{5.6}$$

In the above equations, N is the number of particles in the simulation and B is the Adams parameter (B = μ'/k_BT + ln $\bar{n}$). $\bar{n}$ is the expected number of particles in the system given the volume of the simulation region and is equal to $\bar{p}$v, where $\bar{p}$ is the number density of the particle and v the simulation volume.[50] μ' is the excess chemical potential, k_B is the Boltzmann constant and ΔE the change in energy between the new and old states.

Unlike the double-decoupling and JAWS simulations, no formal hardwall region is applied in a GCMC simulation. Although other water molecules are prohibited from entering the defined region, solute and protein atoms are allowed to occupy the same region as the GCMC simulation. As a result a smaller 2x2x2 Å^3 grid was defined around each water molecule to obtain sufficient sampling of the localised water occupancy, since it was observed that in some cases the volume occupied by the water molecule was filled with a solute atom. Each B value was simulated for 40 million MC moves, divided into 800 blocks of 50K steps each.

At the end of each simulation the average population across the entire simulation was recorded. The decoupling free energy of the water was found using equation **??**.

$$\Delta G_{dec} = -k_B T \ln\left(\frac{[L_{sim}]}{[L_{ideal}]}\right) \tag{5.7}$$

In equation **??**, [L_{sim}] is found by initially recording the population at a set B value. This population is converted into a localised concentration by dividing by the simulation volume, and then converting this into a molar concentration using Avogadro's number. [L_{ideal}], as shown in equation **??**, is related to the B value of the simulation and is found using equation **??**.

$$[L_{ideal}] = 55.56M \times exp(B - \ln \bar{n}) \tag{5.8}$$

In equation **??**, $\bar{n}$ is the expected number of particles in the system given the volume of the simulation region and is equal to $\bar{p}$v, where $\bar{p}$ is the number density of the particle and v the simulation volume.[50]

Having calculated ΔG_{dec}, the binding free energy of the water was found using equation **??**.

$$\Delta G_{bind} = \Delta G_{dec} + \Delta G_{hyd} \tag{5.9}$$

For each water molecule, at least 5 B values were simulated to allow for a reliable estimate of the binding free energy, found as the average of the binding free energies across the range of B values. The free energy of hydration, ΔG_{hyd}, was taken to be +6.4 kcal/mol.

For the GCMC simulations, solvent moves were attempted with a probability of 44 %, protein side-chain moves with a probability of 5.8 % and solute moves

with a probability of 1.8 %. Insertion and deletions were attempted with an equal probability of 2.2 %, with translations and rotations of the isolated GCMC water attempted with a probability of 44 %.

Using the interacting particle approach, an estimate of the binding free energy of a water can be found as standard error of the binding free energy across a number of different B values.

GCMC protocol - Simulated Annealing Method

An alternative method for calculating the binding free energy of a water molecule through GCMC lies in the simulated annealing approach.[54] Rather than converting populations into localised concentrations, the populations obtained from simulating at a range of B values are instead used to make a free energy titration plot. The value of B can be related to the binding free energy using equation **??**.

$$\Delta G_{bind} = \Delta G_{hyd} + k_B T(B - \ln \bar{n}) \tag{5.10}$$

In equation **??**, the hydration free energy of water, ΔG_{hyd}, is taken to +6.4 kcal/mol, k_B is the Boltzmann constant, T the temperature of the simulation and $\bar{n}$ is the expected number of particles in the system given the volume of the simulation region.

By plotting the average population occupancy of the water molecule as a function of the binding free energy, found using B, the value of ΔG_{bind} can be found at the equivalence point of the graph i.e. extrapolating at a population of 0.50.

Using the simulated annealing approach, an estimate of the binding free energy of a water can be found as the difference between two consecutive B values; equal to $\pm$ 0.60 kcal/mol.

The simulation protocol for the simulated annealing approach is the same as for the interacting particle method.

Unless stated otherwise, the above protocols were used for all of the subsequent studies. The JAWS and GCMC protocols were coded into the in-house Monte Carlo software, ProtoMS [118], which was used to perform all of the simulations in this thesis unless otherwise stated. Approximately 2000 lines of Fortran77 code were written to implement the methods.

5.2.3 JAWS placement

A JAWS stage one simulation was performed upon N9-neuraminidase, incorporating a volume of approximately 1000 Å^3. The simulation identified 7 possible hydration sites, shown in Figure **??**, in good agreement with both the crystallographic data and the original simulations performed by Michel *et al.*[13] The native crystal structure contains six crystallographic waters [113], with the additional site, Wat7, found by the JAWS simulations.

Attempts were made to calculate the binding affinity of each of the waters using the JAWS stage 2 algorithm. It was found, however, that the majority of water molecules did not experience sufficient $\theta < 0.05$ transitions during the simulation timeframe. As a result, the binding free energies calculated by the method were either poorly converged or unobtainable. An example of this is shown in Figure **??**, where the standard biasing term does not induce any $\theta < 0.05$ transitions for Wat 5 in N9 neuraminidase.

It has been previously recognised that one the major drawbacks of the JAWS algorithm is that it cannot calculate the binding affinities of strongly bound waters.[41]

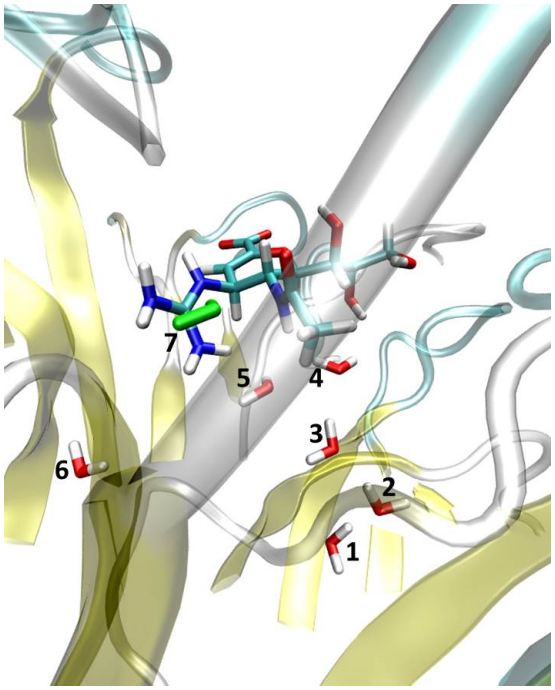


Figure 5.1: The 7 possible hydration sites identified by JAWS in N9 neuraminidase. The non-crystallographic site, Wat7, is highlighted in green

In order to calculate the binding free energies of strongly bound water molecules, modifications to the JAWS biasing term are required. These modifications are now discussed.

5.3 Development of the JAWS algorithm to calculate strongly bound waters

The calculation of the binding free energy of a water molecule is captured by equation **??**, where P($\theta_i \rightarrow 1$) and P($\theta_i \rightarrow 0$) is the probability of a water molecule being observed at a θ value of > 0.95 and < 0.05 respectively.

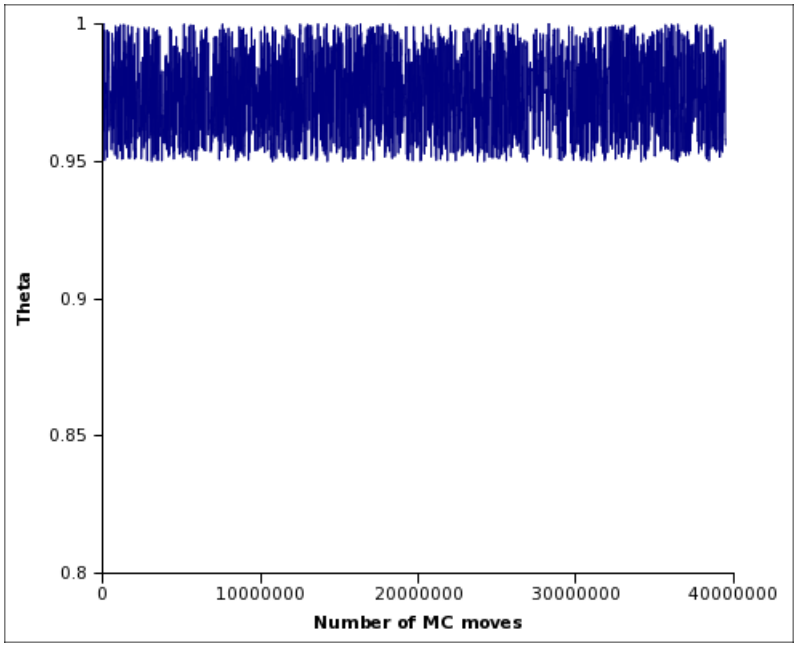


Figure 5.2: θ sampling for Wat5 in N9 neuraminidase using a biasing potential of 6.4 kcal/mol

$$\Delta G_{bind} = -k_B T \ln \left(\frac{P(\theta_i \rightarrow 1)}{P(\theta_i \rightarrow 0)} \right) \tag{5.11}$$

For weakly bound water molecules the biasing potential applied in the second stage of the JAWS algorithm is sufficient to ensure that the θ water molecule can sample both the on and off states, ensuring that enough statistical sampling is performed to obtain a reliable free energy estimate. However for strongly bound water molecules the standard bias potential of +6.4 kcal/mol is not sufficient to induce transitions to the off state, resulting in either poor or no sampling and an unreliable estimate of the binding free energy.

One way of ensuring that enough sampling is performed at both end states is by changing the biasing potential applied in the second stage of the algorithm. Rather than basing this upon the hydration free energy of water, the applied bias

can be changed to one which induces sufficient transitions between the two states. The resulting free energies obtained are indicative of the new biasing potential and hence must be corrected to take the standard hydration free energy of water into account, as shown in equation **??** where ΔG_{bias} is the value of the bias applied in the second stage of the algorithm. The free energies obtained by this method are broadly independent of the applied bias, with an example of this shown in Figure **??**.

$$\Delta G_{bind}(\theta_i) = -k_B T \ln\left(\frac{P(\theta_i \rightarrow 1)}{P(\theta_i \rightarrow 0)}\right) + 6.4kcal/mol - \Delta G_{bias} \tag{5.12}$$

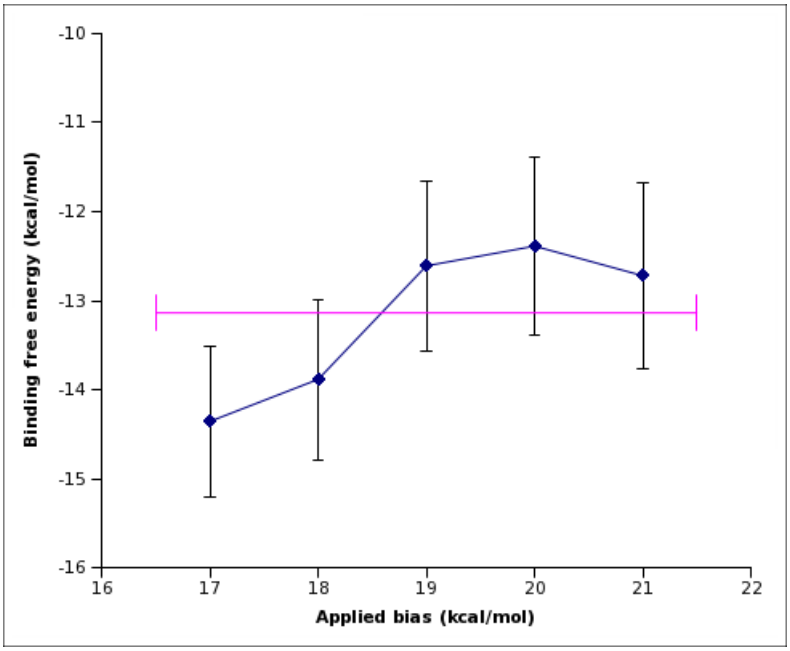


Figure 5.3: Effect of the applied biasing potential upon the JAWS stage 2 binding free energy of Wat5 in N9 neuraminidase

The error estimates observed in Figure **??** arrive from the fact that the biasing term is not completely assigned to both the on and off states, since the thresholds

for the end states are taken to be 0.95 and 0.05 respectively rather than 1 and 0. Assuming that the population distribution of θ across 0.95-1.00 and 0.05-0.00 is uniform, equation **??** can be modified to equations **??** and **??**. In these equations, the on and off states are approximated by the average of the threshold and the ideal end point:

$$\Delta G_{corr} = 0.975 * \Delta G_{bias} - 0.025 * \Delta G_{bias} \tag{5.13}$$

$$\Delta G_{bind}(\theta_i) = -k_B T \ln \left(\frac{P(\theta_i \rightarrow 1)}{P(\theta_i \rightarrow 0)} \right) + 6.4 - \Delta G_{corr} \tag{5.14}$$

There is also an error associated with the choice of θ threshold. For example, the threshold for an 'on' state could either be $\theta > 0.95$ or $\theta > 0.98$. To estimate this error, the binding free energy of a water molecule in N9-neuraminidase was calculated using different thresholds. The calculated results can be found in Figure **??**, with the error associated with the choice of θ estimated to be $\pm$ 0.30 kcal/mol.

The need for changing the applied bias can be seen in Figure **??**, whereby transitions between the on and off state can be induced by increasing the applied bias. At a biasing potential of 10 kcal/mol the water molecule experiences most of the simulation time in the on state, with no transitions to the off state being observed, meaning that a reliable free energy estimate cannot be obtained. However, upon a switch to 17 kcal/mol, the water molecule can sample both end states, allowing for a reliable free energy estimate.

5.3.1 Choice of biasing potential

Since the binding free energy is broadly independent of the applied bias, as demonstrated in Figure **??**, a JAWS stage 2 simulation needs to be run at only one value

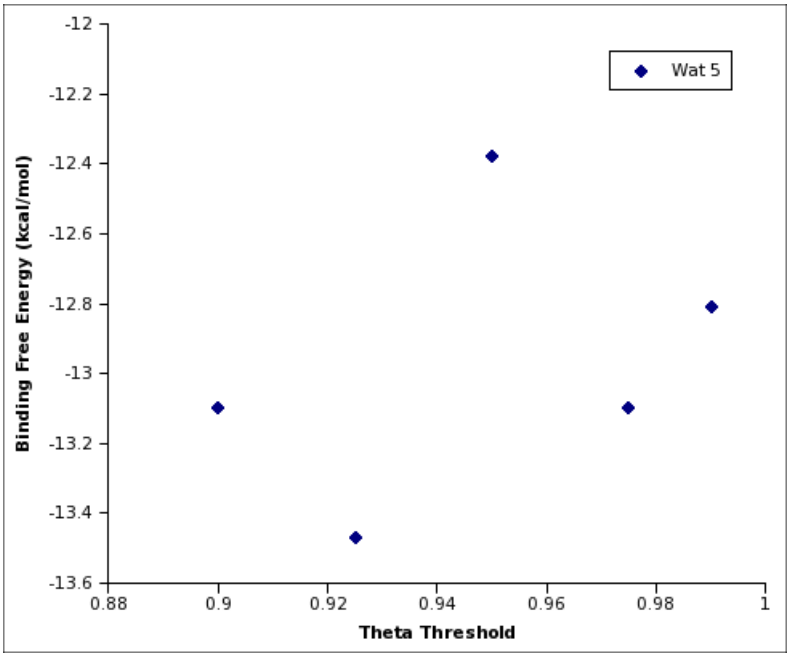


Figure 5.4: JAWS stage 2 binding free energy of Wat5 in N9 neuraminidase as a function of the θ threshold

of the bias to extract the binding free energy. The ideal bias should induce an equal number of on and off states, meaning that the binding free energy becomes the difference between the standard hydration free energy of water and the applied bias. An equal number of on and off states means that the water is, on average, present 50 % of the time and should give the most reliable estimate of the binding free energy. To achieve this, a simplex-style minimisation procedure is applied.

A short JAWS stage 2 simulation, typically one million MC moves, is performed and an estimation of the binding free energy found. The biasing potential is then iteratively changed to obtain a value of the bias which yields an equal number of on and off states. An example of this process is shown in Table **??**.

In table **??**, the iterative minimisation is trying to obtain a value of ΔG_{bias} which induces an equal number of on and states. As shown in equation **??**, an equal number of on and off states should result in a $\ln$ term approaching zero. At

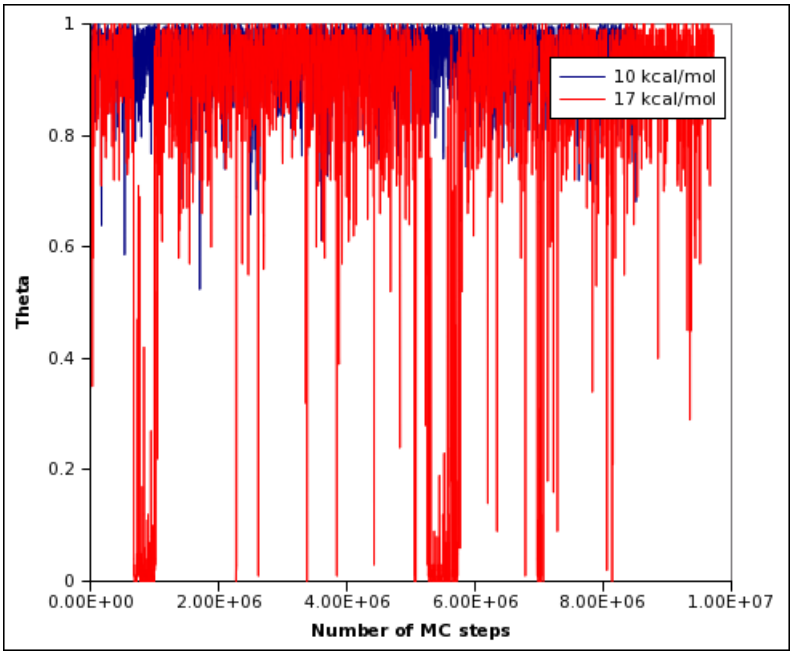


Figure 5.5: θ sampling as a function of the applied bias potential for Wat3 in N9 neuraminidase

Iteration	ΔG_{bias} (kcal/mol)	Ln(On/Off)	Iteration	ΔG_{bias} (kcal/mol)	Ln(On/Off)
1	15	12	6	17.5	-1
2	15.5	10	7	17	11
3	16	11	8	17.5	3
4	16.5	4	9	18	-2
5	17	3	10	17.5	0

Table 5.1: Simplex minimisation for Wat3 in N9 neuraminidase

the end of each simulation, the ln ratio of on and off states is calculated. If the value is positive, suggesting more on states than off, then the value of ΔG_{bias} is increased by 0.5 kcal/mol for the next iteration. If the value is negative, suggesting more off states than on, then the value of ΔG_{bias} is decreased by 0.5 kcal/mol. At the end of the process the value of ΔG_{bias} which gives a ln term of zero is chosen for the main simulation. In this example, the value of ΔG_{bias} was taken to be 17.5

kcal/mol.

5.3.2 Error calculation

For all of the subsequent studies, the error in the calculated JAWS stage 2 binding free energy is found as the sum of two possible sources. The first source of error is in the choice of θ threshold, estimated in Figure **??** to be $\pm$ 0.30 kcal/mol.The second source arrives from the fact that the biasing term is not completely assigned to both the on and off states, since the thresholds for the end states are taken to be 0.95 and 0.05 respectively rather than 1 and 0. This error is dependent upon the choice of ΔG_{bias}, and is found using equation **??**.

$$\Delta G_{error} = \Delta G_{bias} - (0.975 * \Delta G_{bias} - 0.025 * \Delta G_{bias}) \tag{5.15}$$

5.4 Binding Free Energy Calculations

5.4.1 JAWS vs RETI

Using the new modifications, a JAWS stage 2 simulation was performed upon each hydration site as identified in Figure **??**, with each site also studied using double-decoupling. The binding free energy for JAWS stage two simulations was found using equation **??**. The free energy comparison between the two methods can be seen in Figure **??**.

Figure **??** clearly demonstrates that the two methods give excellent agreement with each other. Both strongly and weakly bound water molecules are picked up by the two methods, showing that the modification to the JAWS algorithm has

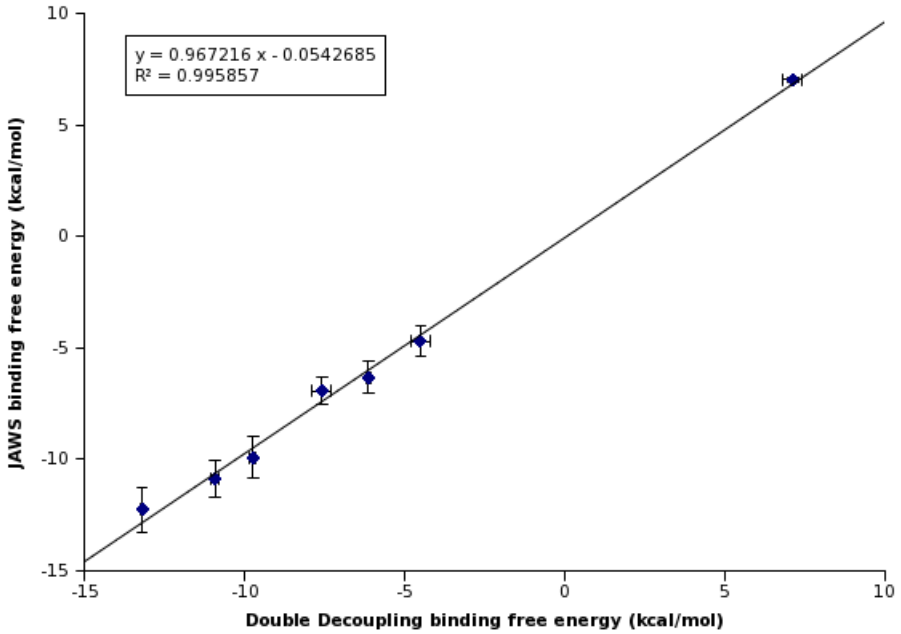


Figure 5.6: Binding free energies for the 7 hydration sites in N9 neuraminidase, found using JAWS stage 2 and RETI double-decoupling

been successful in predicting the binding free energy of water molecules which previously were incalculable. The biasing potentials used to calculate the JAWS free energies in Figure **??** can be seen in Table **??**.

Water Molecule	JAWS ΔG_{bind} (kcal/mol)	RETI ΔG_{bind} (kcal/mol)	JAWS Bias (kcal/mol)
1	-4.69 (0.70)	-4.51 (0.25)	14
2	-6.93 (0.60)	-7.60 (0.30)	12
3	-10.87 (0.85)	-10.86 (0.14)	17.5
4	-6.32 (0.70)	-6.14 (0.09)	14
5	-12.39 (1.00)	-13.21 (0.06)	20
6	-9.91 (0.95)	-9.76 (0.10)	19
7	7.04 (0.10)	7.13 (0.30)	2

Table 5.2: Binding free energies for the 7 water molecules in N9 neuraminidase, found using JAWS stage 2 and RETI double-decoupling. Errors are shown in parenthesis

The convergence for the JAWS stage 2 simulations can be seen in Figure **??**. It

can be seen that all of the simulations are well converged after ca. 10 million MC moves, demonstrating further that the biasing potentials used yield precise results.

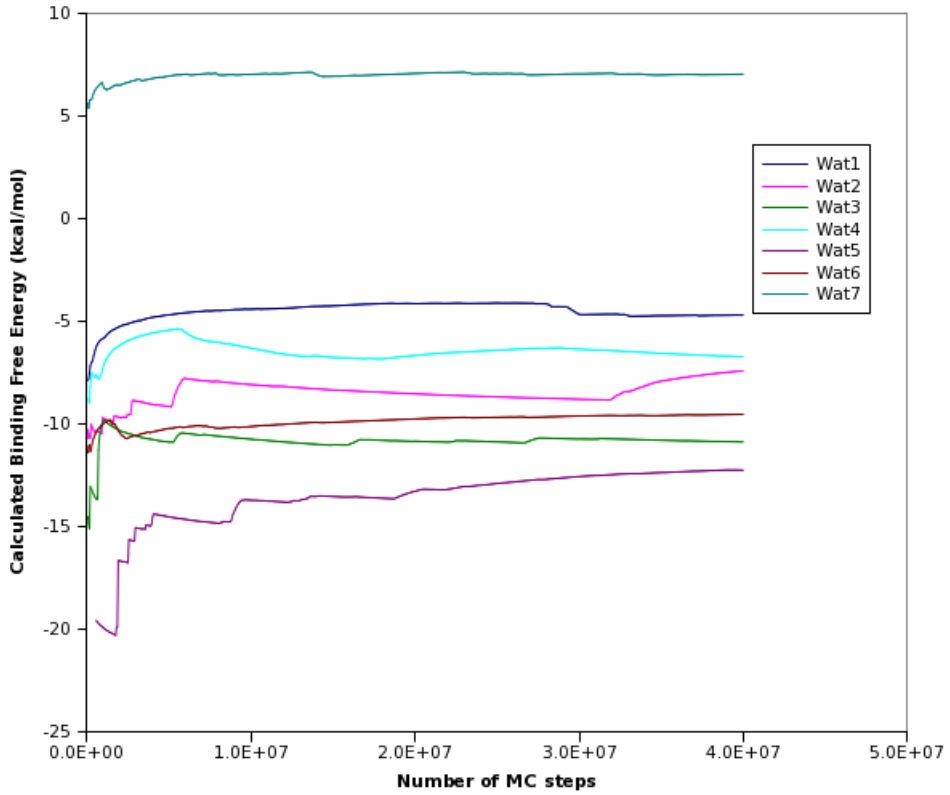


Figure 5.7: Convergence of the binding free energies in the JAWS stage 2 simulations

5.4.2 GCMC vs. RETI

Using the hydration sites identified by the JAWS stage 1 simulations, the free energy of binding of each site was calculated using the interacting particle method of Clark *et al.*[108], described previously in section **??**. Equation **??** yields the decoupling energy of the water molecule from the protein, and can be corrected with the hydration free energy of water to arrive at a binding free energy using

equation **??**. Figure **??** shows the predicted binding affinity of the 7 molecules using the two methods and shows an excellent correlation. Initial tests were made to apply the non-interacting particle method, but these proved to be unsuccessful. Since sidechain movement is allowed during the GCMC simulation, the high concentration of water molecules localised on one site allowed nearby sidechains to migrate towards the binding region. The end point of such a simulation is not the same as the other methods, and as such it cannot be reliably compared.

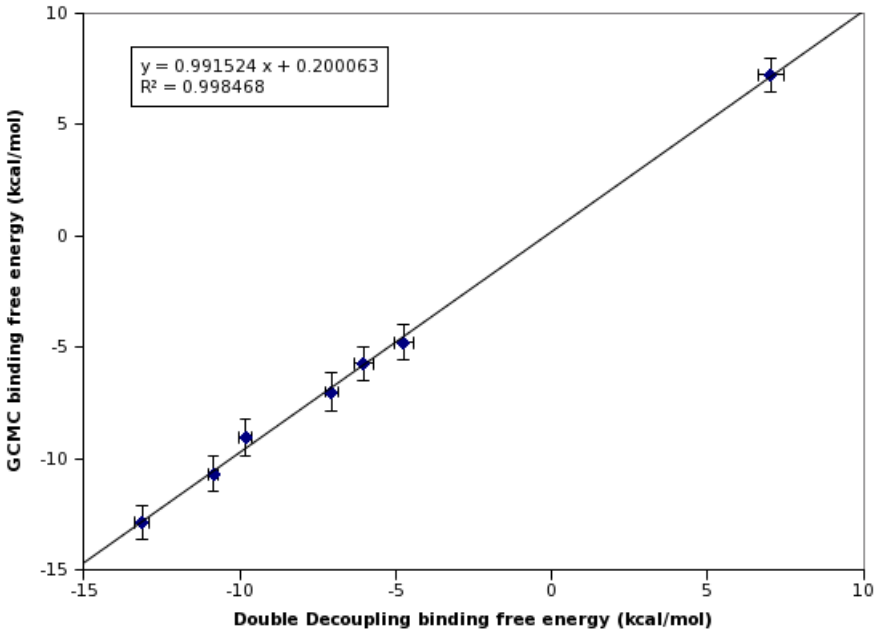


Figure 5.8: Binding free energies for the 7 hydration sites in N9 neuraminidase, found using GCMC and RETI double-decoupling

The reported GCMC binding free energies were calculated as the average of the binding free energy across a range of B values. An example of this for the Wat7 site, the weakest binder in the series, can be seen in Table **??**. The table shows that the binding free energy is consistent at around 7 kcal/mol once the average population drops below 0.60. This behaviour is similar to that demonstrated by Clark *et al.* in the calculation of benzene-T4 lysozyme binding free energies.

[106, 108] The maximum occupancy for the water site is one, meaning statistically significant occupancies of less than one need to be obtained to get a reliable estimate of the binding free energy.

B	Average population	[L_{sim}] (M)	[L_{ref}] (M)	ΔG_{dec} (kcal/mol)	ΔG_{bind} (kcal/mol)
4	0.9725	202	11400	2.39	8.79
3	0.9325	194	4120	1.82	8.22
2	0.7275	151	1540	1.37	7.77
1	0.570	118	565	0.926	7.33
0	0.368	76.3	208	0.593	6.99
-1	0.148	30.6	76.5	0.542	6.94
-2	0.04	7.78	28.1	0.761	7.16

Table 5.3: Calculated free energies for Wat7 in N9 neuraminidase, found at different B levels

Using the data for B values less than 2 in table **??**, the binding free energy can be found as 7.24 $\pm$ 0.17 kcal/mol.

One major drawback associated with the GCMC method lies in the acceptance rate of insertion and deletion moves. For an insertion to be accepted it is important that the orientation of the water molecule is correct, since otherwise it is likely that the intermolecular interactions between the water and its environment will be unfavourable.[53] As a result insertion rates as low as 0.1 % are seen in GCMC simulations, which in turn leads to poor sampling. It is this poor sampling which could potentially lead to an increase in the uncertainty in the free energies of GCMC simulations compared to both double-decoupling and JAWS, although no evidence of this has been seen in this study,

Since the interacting particle method can generate populations as a function of B, this information can also be used to derive free energies via the simulated annealing approach.[54] As previously described, the method can be considered to be analogous to a chemical titration whereby the decoupling free energy of the

water molecule is the equivalence point at which the average population is 0.50. The data from Table **??** has been used to generate such a titration profile, seen in Figure **??**.

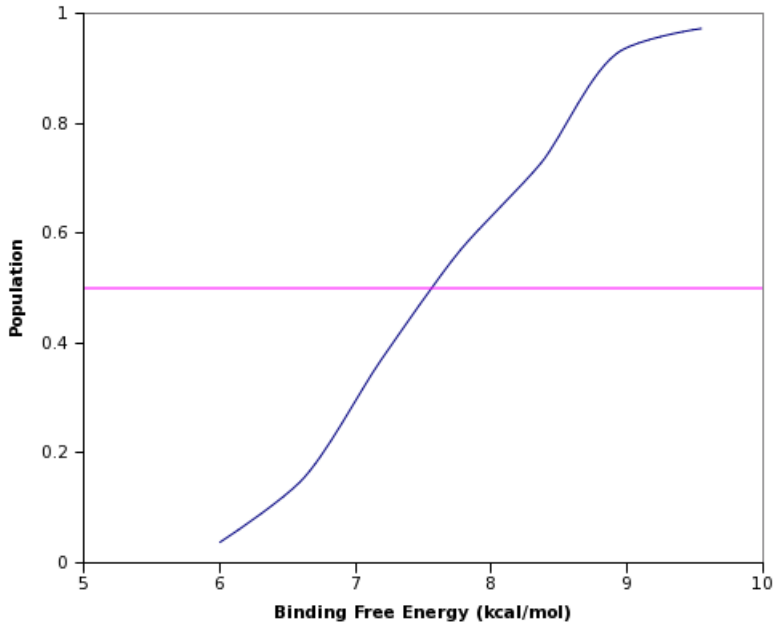


Figure 5.9: Free energy titration plot for Wat7 in N9 neuraminidase, found using the GCMC simulated annealing approach

Figure **??** shows that the estimated binding free energy of Wat7 is approximately 7.4 kcal/mol at the 0.50 equivalence point, in good agreement with the value calculated by the interacting particle method. Since either method can be used to derive the same result to within error, the question arises as to which of the GCMC methods is advantageous to calculate binding free energies. Whilst the simulated annealing approach gives information regarding the behaviour of the system as a function of the B, the interacting particle approach is significantly faster since it only requires the simulation to be performed at one value of B. However, the correct B to choose is not always known *a priori*, meaning that it can require several different simulations to arrive at statistically significant B val-

ues. It therefore appears to be more advantageous to use the simulated annealing approach. Utilisation of the simulated annealing approach to look at the influence of B upon the position and binding affinity of hydration sites will be discussed further in the subsequent sections.

5.4.3 Preliminary conclusions

The N9-neuraminidase system was chosen as a test for the different methodologies since there are a large number of different studies utilising the system in free energy calculations.[13, 40] The fact that all three free-energy methods give near identical results is clearly encouraging and lends itself to the question over which method is best suited to a particular problem. Whilst RETI double-decoupling is the most rigorous method of the three it is also the most computationally expensive. A typical simulation requires in excess of 300 CPU hours, whilst both GCMC and JAWS require an order of magnitude less time. As a result it is suggested that double decoupling is used in cases where precise free energies are required or in ambiguous cases.

One drawback of the double-decoupling approach is that it requires prior knowledge of the water binding positions; something which is found dynamically in both JAWS and GCMC. As such, if novel systems are studied then either JAWS, GCMC or both methods should be employed to identify potential hydration sites. Employing JAWS in free energy studies has already been utilised, whereby changes in hydration as a function of ligand perturbations are accounted for.[61, 62] Since both JAWS and GCMC take similar times to calculate the binding free energy, there is little reason to favour one method over the other. One possible advantage in the JAWS approach is that once the optimal biasing potential is found no further simulations need to be run, whilst the GCMC approach

requires several simulations run at different potentials to arrive at the binding free energy. This in itself, however, highlights one of the advantages in the GCMC approach; that it can give information on the binding of molecules as a function of the chemical potential.

One potential problem with the JAWS approach to calculate binding free energies is how to deal with the intermediate θ states. Since only the end points are considered when calculating the binding free energy, it is unclear whether or not the intermediate data should be considered. Such data incorporates a higher proportion of the recorded θ chemical space, highlighted in Figure **??**.

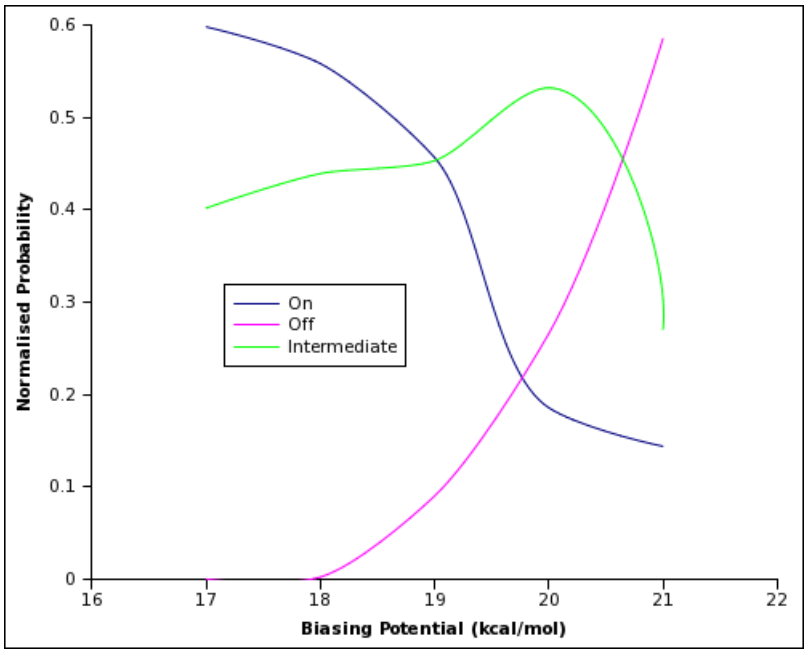


Figure 5.10: Normalised probabilities for the on, off and intermediate states as a function of the applied biasing potential for Wat 5 in N9 neuraminidase. The on state was defined as $\theta > 0.95$, the off state defined as $\theta < 0.05$ and the intermediate state was the remaining θ values

Based upon the excellent agreement between JAWS, GCMC and double-decoupling, it seems that ignoring the intermediate states is acceptable. This is, however,

a waste of potentially useful data, although it is unclear how the data could be analysed. In order to understand whether this data is significant, the profiles of strongly and weakly bound water molecules could be compared to examine whether there is a relationship between the binding affinity and the population of intermediate states.

5.5 Bovine Pancreatic Trypsin Inhibitor

5.5.1 Biological Relevance

The Bovine Pancreatic Trypsin Inhibitor (BPTI) is a small protein (58 residues) [119] which inhibits trypsin, a serine protease found in the digestive system. Inhibition of trypsin has been found to reduce bleeding, and as such the drug aprotinin was developed for use during surgery [120] before complications saw the drug removed from general usage. Within the structure of BPTI, two water cavities have been identified. The first, singly occupied by Wat122, finds the water stabilised by 4 hydrogen bonds within the protein cavity. The second cavity, occupied by Wat111, Wat112 and Wat113, finds the three molecules bound to both each other and the protein cavity.

5.5.2 System Preparation

The protein structure 5PTI (resolution 1 Å) [121] was used for the following simulations. The same protein preparation as for N9 neuraminidase, detailed in section **??**, was followed. For the JAWS stage one simulations, 6 JAWS θ molecules were used to simulate the Wat111 cavity, using the same simulation length and protocol as described previously.

5.5.3 Wat122

A schematic representation of the Wat122 cavity can be seen in Figure **??**. The water molecule was studied with JAWS stage 2, double-decoupling and the interacting particle GCMC method, with the results shown in table **??**. A biasing potential of 14 kcal/mol was used for the JAWS stage 2 calculations, using the modified protocol described in section **??**.

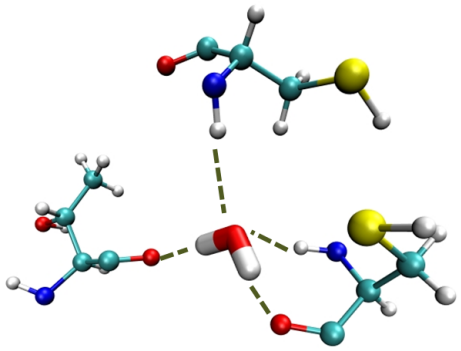


Figure 5.11: The binding pocket of Wat122 in 5PTI

Water Molecule	ΔG_{bind}-JAWS (kcal/mol)	ΔG_{bind}-GCMC (kcal/mol)	ΔG_{bind}-RETI (kcal/mol)
122	-6.75 $\pm$ 0.80	-6.81 $\pm$ 0.80	-6.84 $\pm$ 0.40

Table 5.4: The calculated binding free energy for Wat122 in 5PTI, found using JAWS, GCMC and RETI double-decoupling

The error estimates in table **??** from GCMC and RETI arrive from the calculated standard error across a number of simulations, whilst the JAWS error protocol is described in section **??**.

As with the water molecules in the N9-neuraminidase system, the binding free energy is extremely similar for all of the three methods used. The result

is also consistent with a previous study, with the energy calculated to be -7.08 kcal/mol.[8]

5.5.4 Wat111 cavity

Figure **??** shows the three molecules in the Wat111 cavity. The cavity was subjected to a JAWS stage 1 simulation, which identified the three molecules in close agreement with the crystallographic data. A volume of 245 Å^3 was simulated using 8 θ waters. Snapshots of the population density obtained from the JAWS stage 1 simulation can be seen in Figure **??**, and show that the three water sites are clearly identified.

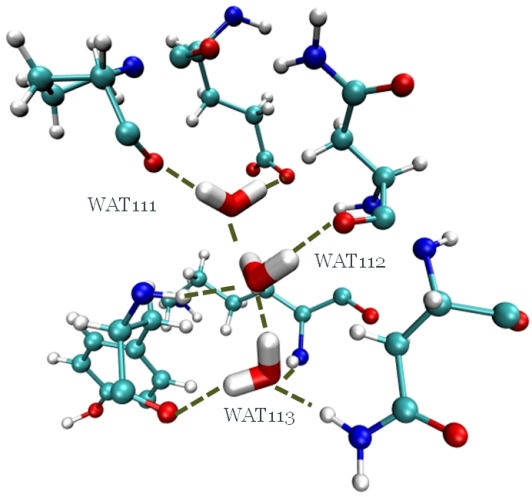


Figure 5.12: The binding pocket of Wat111, Wat112 and Wat113 in 5PTI

A JAWS stage 2 binding free energy calculation was performed upon each of the three located water sites in turn, with the results compared to those obtained by double-decoupling. Table **??** shows that the two methods give excellent agreement with each other.

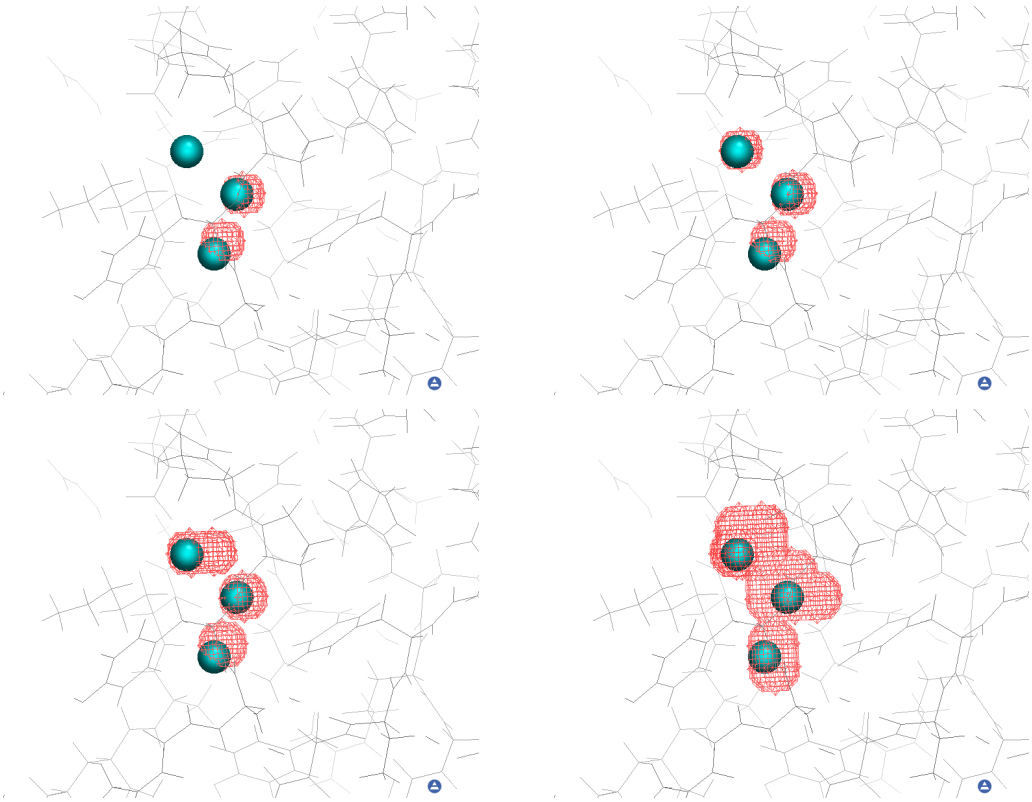


Figure 5.13: JAWS stage 1 clustering density for the Wat111 cavity. From left to right: Top 10 % of data, top 40 %, 70 %, 100 %

Water Molecule	ΔG_{bind}-JAWS (kcal/mol)	Bias Potential (kcal/mol)	ΔG_{bind}-RETI (kcal/mol)
111	-12.07 $\pm$ 1.20	20	-12.70 $\pm$ 0.40
112	-15.60 $\pm$ 1.30	21	-15.77 $\pm$ 0.40
113	-15.70 $\pm$ 1.40	22	-15.86 $\pm$ 0.40

Table 5.5: JAWS and double-decoupling binding free energies for the three water molecules in the Wat111 cluster

The error estimates in table **??** from GCMC and RETI arrive from the calculated standard error across a number of simulations, whilst the JAWS error protocol is described in section **??**.

It can be seen from Figure **??** that the water molecules in this cavity are all hydrogen bonded to both the protein and each other. A GCMC simulation, an-

nealing the value of B from high to low, was performed upon the entire cavity to see whether the method could accurately predict both the occupancy and binding free energy of the water molecules. It was found that the method predicts an occupancy of 3 at high values of B, however as the chemical potential was dropped the prediction changed.

As the chemical potential is lowered to -14.6 kcal/mol (corresponding to a binding free energy of -8.2 kcal/mol), Wat111 leaves the system. This is an unexpected result, since both the JAWS stage 2 and double-decoupling simulations suggest that the binding free energy of Wat111 should be around -12 kcal/mol. When Wat111 leaves the system Wat112, seen as the middle water in Figure **??**, moves slightly and occupies an intermediate site between itself and Wat111 - preventing Wat111 from reinserting back into the cavity. As a result of this reorganisation effect the average occupancy of the cluster is predicted to be 2 at this value of the binding free energy, compared to the crystallographic and JAWS evidence suggesting it should be 3. Once the intermediate water is observed the average cavity population is never subsequently observed to be 3. Consequently, the reorganisation effect on the system results in a different prediction of the binding free energy being obtained when GCMC is used. The calculated free energies of the other waters using the GCMC method will now be indicative of the new intermediate positions, instead of the original positions. A snapshot of this behaviour can be seen in Figure **??**.

It is important to test whether the same behaviour is observed going from high to low chemical potential as it is going from low to high. If this is not the case, then it suggests that the method does not display reversibility. Figure **??** shows the effects of forward and backward annealing on the occupancy of the cavity, and demonstrates that no considerable hysteresis is observed.

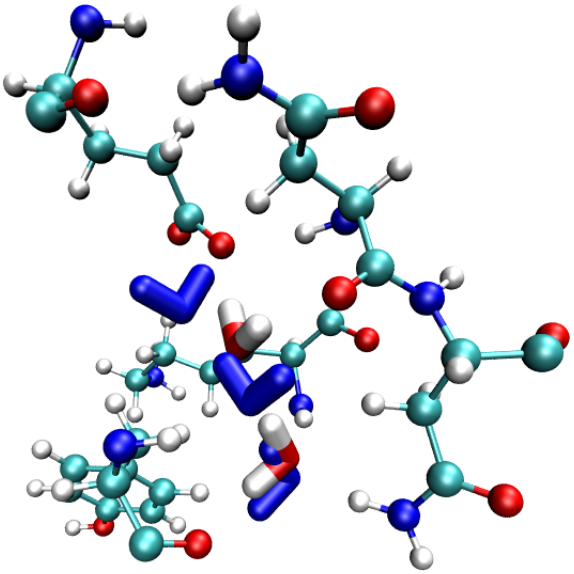


Figure 5.14: The intermediate position of Wat112 in 5PTI. The waters in blue show the original positions, with the intermediate positions in red and white

This result highlights an interesting feature of the GCMC simulated annealing approach. The method is capable of highlighting effects upon the *system* as the binding free energy changes and in this instance it suggests that if a bound water molecule is removed from the cavity, the system reorientates itself to accommodate the change. JAWS stage 2 and double-decoupling simulations cannot simulate this process, since a hardwall potential is applied to the water molecules during the calculations. The GCMC method is capable of looking at the free energy of the entire cluster, whilst both JAWS stage 2 simulations and double-decoupling simulations look at a water molecule in a particular environment. The information obtained from GCMC cannot be obtained easily from a JAWS stage 1 simulation if the entire pocket is studied, since the method will always predict the optimum packing within the cavity.

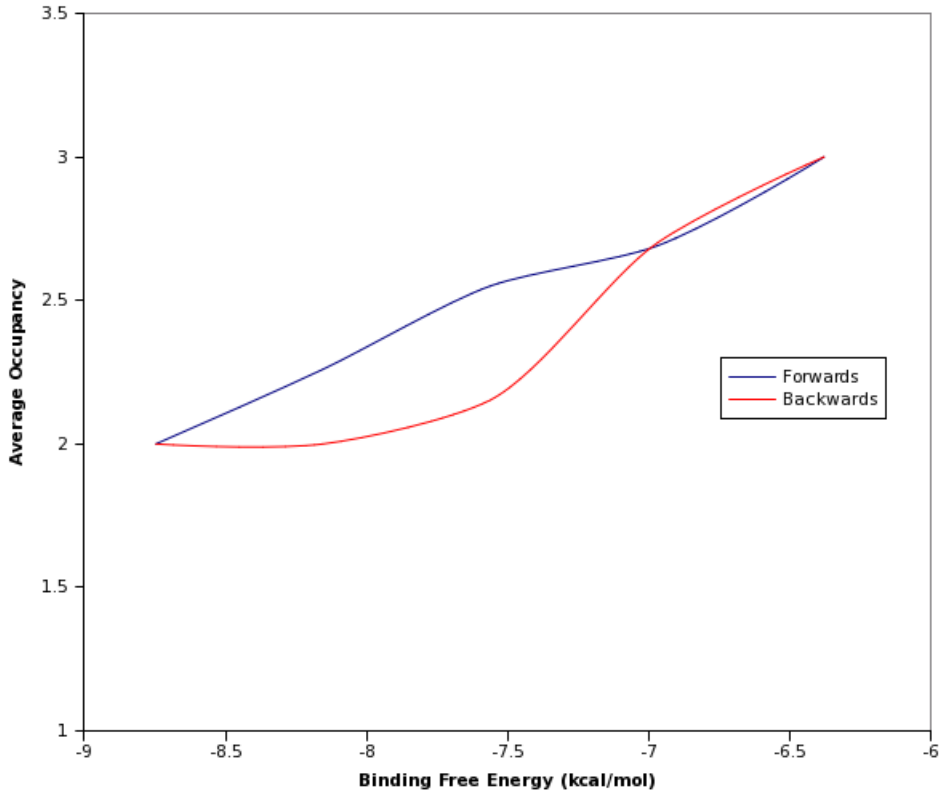


Figure 5.15: Forward and backward occupancies for the Wat111 cavity as a fucntion of the binding free energy

5.6 Conclusions

In this chapter the development of the modified JAWS algorithm was discussed, and this method used alongside double-decoupling Monte Carlo and GCMC to initially locate and calculate the binding affinity of water molecules in N9-neuraminidase. The calculations were consistent with each other and also prior studies, indicating that all three methods can be utilised to calculate the binding free energy of water molecules.

The application of the three methods to two seperate cavities in BPTI was then described, with the methods showing that for a cavity consisting of one water molecule the methods all give the same results. Different behaviour was observed

for the cavity containing three molecules, whereby it was found that the GCMC method can predict hydration changes in the cavity as a function of the binding free energy. This is something which will be exploited in the following chapter, and cannot be obtained from JAWS or double-decoupling simulations which impose a hardwall around waters of interest and look at water molecules in isolation.

The clear indication from these test systems is that although all three methods generally give extremely similar binding free energies, there are definitely cases where certain methods should be used. The double-decoupling method can be viewed as the 'gold standard' of free energy methods, although it is by far the most computationally intensive. As such, it is recommended that this method is used in cases where rigorous free energies are required. Since JAWS stage two simulations require less simulation time than GCMC to calculate binding free energies, this method should be used to initially calculate the binding free energies of waters. Since they can be tuned to specific binding free energies, GCMC simulations should be employed when information on how multiple water molecules behave as a function of the binding free energy of the network. The resultant water locations could then be scored using JAWS or double-decoupling.

Chapter 6

Predicting the location and binding affinity of water molecules - II. Applications

6.1 Introduction

Having established the relative merits of double-decoupling, GCMC and JAWS, the methods can now be applied to novel and more demanding systems. The previous chapter highlighted some of the limitations of the approaches, although the relatively straightforward test systems do not allow a rigorous examination of the methodologies. In this chapter, two different types of problem are examined; the hydration of hydrophobic cavities and the application of the methods to three different kinases. Whilst the hydrophobic cavities have been explored in the literature, the hydration of the kinase systems is a much more novel case. Through careful examination of the apo hydration patterns, strategies for novel ligand design are proposed. The three methods are then used upon the Chk-1 kinase system,

where the role of water network stabilisation upon ligand binding affinity is explained - something which has not been reported in the chemical literature before.

6.2 Hydrophobic cavities

6.2.1 Biological Relevance

The formation of a non-polar core is something which has been established as playing a key role in protein folding and stability.[122] Although significant effort has been made towards understanding the nature of protein folding, methods for determining the hydrophobicity of protein interiors is something which is still of great debate. Some proteins, such as cytochrome P450 [123], require the presence of water in their hydrophobic interior to function, suggesting that the relationship between the nature of hydrophobicity within the protein and its function is still unclear.

A 1998 paper by Hummer *et al.* [122] looked at the relationship between pressure and protein unfolding. Upon the application of pressures > 100 MPa, proteins can undergo denaturation and unfolding. Such behaviour appears to contradict the hydrophobic effect, since the presence of non-polar residues in an aqeuous environment is unfavourable. Hummer instead considered the transfer of water from the bulk into the protein interior, finding that an increase in pressure forces water molecules into the protein interior. This in turn can fill the non-polar core and break apart the protein interior structure, leading to unfolding.

As a case study, Collins looked at the pressure induced filling of the L99A mutant of T4-lysozyme.[59] This cavity is known to be empty under ambient conditions, hence providing an interesting test case to observe the changes in hydration

as the pressure is changed. High pressure crystallography was used alongside a novel free energy approach based upon a grand canonical partition function. MD runs of the system were performed at different pressures and with different numbers of water molecules within the non-polar core. The occupancy probabilities were then calculated using the potential energy change of increasing the number of molecules in the cavity by one, together with the excess chemical potential of water at the desired pressure.

The free-energy results indicated that under ambient conditions the cavity is empty whilst at 200 MPa the cavity can stabily accommodate four water molecules, results supported by the crystallographic electron density. The change in occupancy was attributed to a change in the bulk chemical potential of water rather than a change in the protein interior structure. In this system the 4 water molecules provided a stabilisation effect on the system at higher pressure, suggesting that intermediate states between the folded and unfolded forms might be stabilised by water interactions.

The observed change in chemical potential ($\Delta\mu$ = 0.84 kcal/mol) between ambient conditions and 200 MPa is relatively small, yet it is sufficient to induce filling of the cavity. As such, understanding how water interacts within protein interiors is of great interest. If a small change in the solvent conditions can affect the presence of water within a protein interior, and hence the protein activity, then possible strategies could be designed to directly influence the performance of proteins such as enzymes. The JAWS and GCMC methodologies have already shown to be ideally suited for looking at the behaviour of water within enclosed cavities, and hence lend themselves to be utilised in understanding the behaviour of waters within protein interiors.

6.3 T4-Lysozyme

6.3.1 System Preparation

The same protein preparation and simulation protocols as for N9-neuraminidase were followed, using the pdb structure 2B6T (resolution = 2.10 Å).[124]

6.3.2 Simulations

In order to investigate the occupancy of the hydrophobic cavity, a JAWS stage 1 simulation was performed upon the 210 $Å^3$ cavity. A threshold of $\theta > 0.995$ was employed, with the top 60 % of density shown in Figure **??**. A θ threshold of 0.995 was chosen since it was found by another researcher that the predictions made using this threshold were in better agreement with those made by GCMC for more solvent accessible pockets compared to a 0.95 threshold.[125]

Figure **??** shows that there are 5 possible density regions observed in the simulation. The study performed by Collins *et al.* [59] indicated that the maximum stable occupancy within the cavity at higher pressures was 4, which appears to be in disagreement with the JAWS stage 1 result. The crystal structure deposited from this study however only contained three water molecules, likely to be due to difficulties in accurately resolving the electron density within the cavity. Indeed, the paper acknowledges that the crystallographic evidence alone points to an occupancy between 2 and 4.

The convoluted experimental number density suggests that the waters in this system are likely to be mobile, and can adopt different positions in the cavity. As such the JAWS evidence could be indicating 5 *possible* positions instead of 5 *absolute* positions. This highlights one of the pitfalls in the JAWS placement of

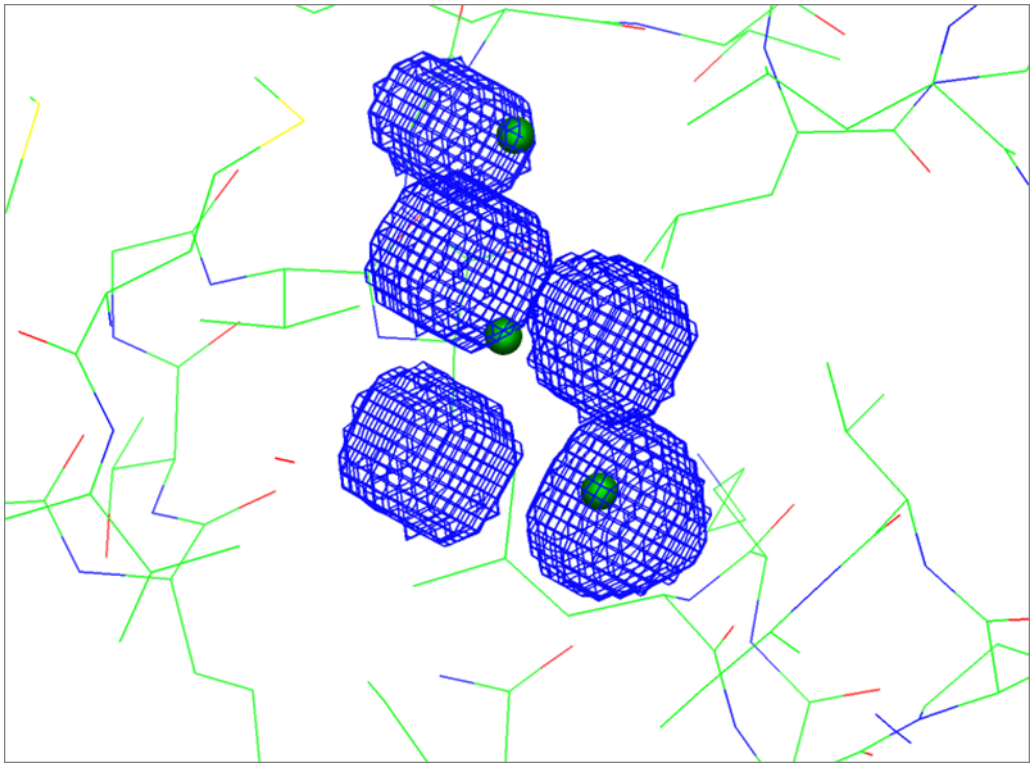


Figure 6.1: Calculated JAWS stage 1 density for the T4-lysozyme cavity. The top 60 % of data is displayed in blue, with crystallographic waters in green

waters; that it cannot identify cooperativity between possible water sites. During a GCMC simulation a water molecule can only be on or off, meaning that identifying cooperativity can be achieved by visual inspection of the simulation output. In comparison, water molecules can adopt θ values between the on and off states in a JAWS stage 1 simulation. This allows water molecules to be stabilised by nearby waters with intermediate θ values, making the assignment of accurate hydration sites a challenge.

In order to learn more about the relationship between waters in this system, GCMC simulated annealing was performed at different values of B. As the value of B is changed, the occupancy of the waters within the cavity should change and highlight the effect of water cooperativity.

Figure **??** shows the effect upon the occupancy as the value of B (plotted as the binding free energy) is changed. Average occupancies above 4 were not observed at any of the B levels suggesting that an occupancy of 5 is extremely disfavoured, corroborating the previous work by Collins.

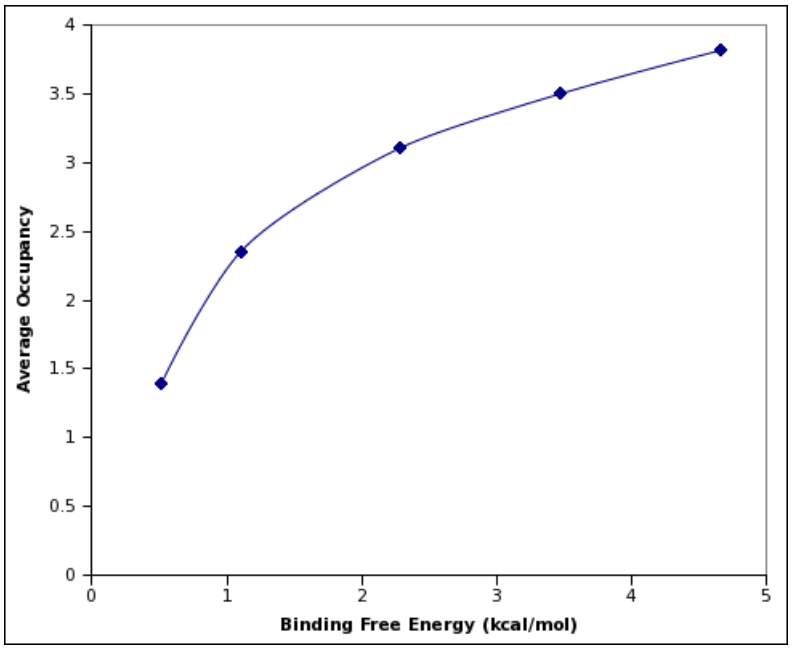


Figure 6.2: Average occupancy of the T4-lysozyme cavity as a function of the estimated binding free energy, found using GCMC

Since both the GCMC simulations and previous work by Collins suggest an occupancy of 4 water molecules, the oxygen-water coordinates from the B = -1 simulation (corresponding to a binding free energy of + 4.65 kcal/mol) were used to create a density map. A representative snapshot from the simulation was then sought which best matched the density plot and had four water molecules. This matching process was done by visual inspection across the simulation snapshots, and allows JAWS stage 2 simulations to be performed. The outcome of this is shown in Figure **??**.

Figure **??** shows that three of the observed water molecules are in good agree-

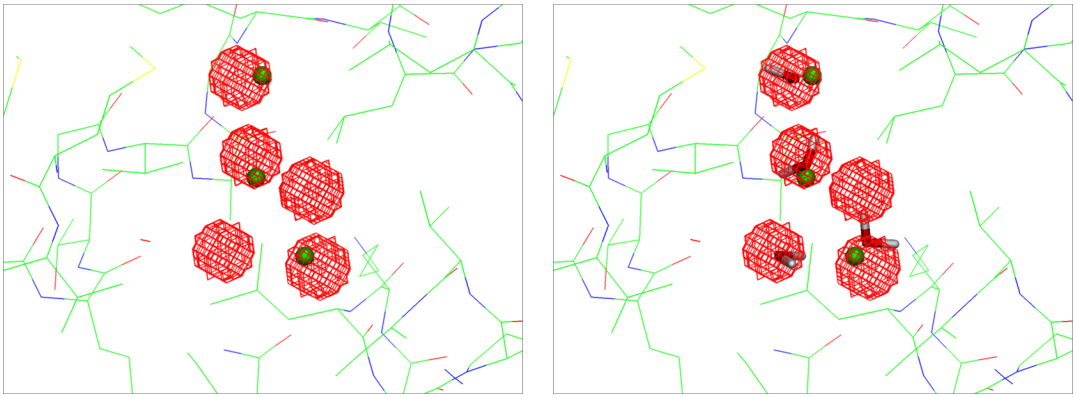
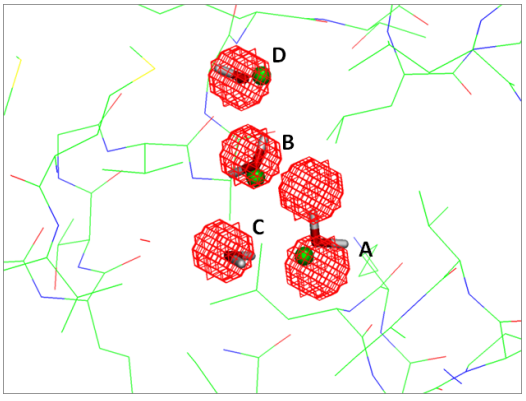


Figure 6.3: **Left:** Density plot for the T4-lysozyme cavity obtained using a B value of -1, corresponding to a binding free energy of + 4.65 kcal/mol. Shown in green are the three crystallographic waters. **Right:** Chosen snapshot locations for the waters

ment with the crystallographic sites. It is interesting to note that the middle crystallographic water sits in a large cloud of density compared to the other two sites, which could explain the presence of the 5th region of density found in both GCMC and JAWS. As expected, this water molecule exhibits the highest temperature factor of the 3.

Having placed the waters, each water was then subjected to a JAWS stage 2 simulation. The same biasing potential was used for all of the simulations, being 10 kcal/mol. The calculated binding free energies can be seen below in Figure **??**.



Water	ΔG_{bind} (kcal/mol)
A	+0.70
B	-2.23
C	+1.36
D	+1.09

Figure 6.4: JAWS binding free energies for the water molecules found at B = -1, corresponding to a binding free energy of + 4.65 kcal/mol

The calculated free energies show that 3 of the water sites are unfavourable, whilst one exhibits a negative binding free energy. It is interesting to note that this molecule, Wat B, is the one which closely matches the crystallographic water with the highest temperature factor. This suggests that there could be an entropic contribution to the binding free energy, in addition to the hydrogen bonding between the water molecules. Since the binding free energy of Wat B is supported by three water molecules with positive binding free energies under these simulation conditions, it is important to understand the effect of removing the positively bound water molecules. It is assumed in this work that water molecules with positive binding free energies will prefer to exist in the bulk rather than the protein cavity, providing a justification for the removal of the waters.

When the binding free energy of Wat B is calculated in the absence of the other 3, the binding free energy changes to +1.92 kcal/mol, suggesting that its occupancy depends solely upon the presence of other 3 molecules. Based upon this it is apparent that the occupancy of the cluster under standard conditions should be zero, corroborating the previous results obtained by Collins. This is also in agreement with the titration plot obtained by GCMC, as shown in Figure **??**, where the number of water molecules at a binding free energy of 0 kcal/mol is expected to be zero.

As previously discussed, the filling of the cavity under higher pressures is thought to be due to a shift in the chemical potential of bulk water, as opposed to a change in the protein structure. Since JAWS stage 2 simulations are performed in the canonical ensemble, changes in pressure cannot be accounted for directly. As a result, a correction term needs to be applied which implicitly takes into account the change in the bulk chemical potential. This was achieved by running double decoupling simulations at 2000 bar to see the effect upon the hydration free energy

of water in bulk water, and observing if this matched the change observed by Collins ($\Delta\mu$ = 0.84 kcal/mol).

A water molecule was decoupled from a box of bulk water in two stages; initially decoupling the electrostatic terms and then perturbing the Lennard-Jones terms on the oxygen atom of the water molecule to zero. The NPT ensemble was used to perform the simulations, with the simulations performed at the two different pressures. The same number of MC moves and λ windows as described in the N9 neuraminidase double-decoupling simulations were used. Solvent moves were attempted with a probability of 99 %, solute moves with a probability of 0.9 % and volume moves with a probability of 0.1 %.

The PMF profiles for the water decoupling at 1 bar and 2000 bar can be seen in Figure **??**.

The calculated hydration free energy is 5.84 $\pm$ 0.4 kcal/mol; a change of 0.56 kcal/mol compared to the standard hydration free energy at 1 bar. This is in good agreement with the value calculated by Collins.

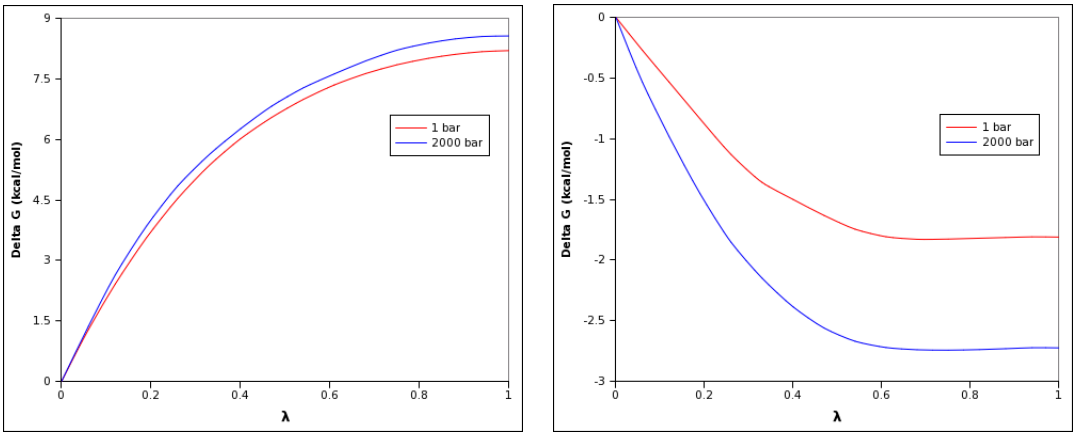


Figure 6.5: PMF profiles for the decoupling of the electrostatic (left) and Lennard-Jones (right) terms for a water molecule at 1 and 2000 bar

The JAWS stage 2 binding free energies for the four waters, corrected for the difference in hydration free energy at 2000 bar, -0.56 kcal/mol, can be seen in table

??. It can be seen that all of the free energies become more favourable, with waters A, C and D all displaying hydration free energies which are indicative of waters which are expected to be intermittently present under standard conditions. The work by Collins showed that, although a cluster of four waters is stable at 2000 bar, it is still not as favourable as having no waters in the cavity. The modified JAWS stage 2 results support this assertion, and provides evidence for the possible existence of water within the T4-lysozyme cavity at elevated pressures.

Water	ΔG_{bind} (kcal/mol)
A	+0.14
B	-2.79
C	+0.80
D	+0.53

Table 6.1: Modified JAWS binding free energies for the water molecules found in the T4-lysozyme cavity

6.3.3 Conclusions

The T4-lysozyme system provides more evidence for the critical role which waters play in a network. As demonstrated through the JAWS stage 2 binding free energies at 1 bar, although a water molecule might have a favourable binding free energy it does not mean that it will be present under standard conditions if its occupancy is supported by nearby, unfavourable waters. As such, the entire network needs to be taken into account - something which can potentially be time consuming for JAWS stage 2 simulations. The GCMC methodology achieves this since it looks at the free energy of the cavity, rather than each individual water molecule. In this system, the method clearly correctly predicts that the occupancy of the cavity is zero under standard conditions.

6.4 Interleukin 1 β

6.4.1 Biological relevance and motivation

The interleukin 1 β protein (IL1β) is a cytokine which is released to coordinate responses to immune challenges.[126] Similar to T4-lysozyme, it contains a hydrophobic cavity with a volume of around 80 Å^3. Early crystallographic studies indicated that the cavity was empty under ambient pressures, however later NMR studies indicated that the cavity could accommodate water, citing the fact that the disordered nature of waters within the cavity masked their presence by X-ray crystallography.[127] A similar debate has arisen when simulation methods have been applied to the system. Original studies by Zhang and Hermans [71] concluded that the cavity was empty, whilst a more recent paper by Somani [128] has suggested that the cavity can accomodate 4 water molecules. As such, the nature of the pocket is unclear.

A later study by Yin utilised the grand-canonical approach used for T4-lysozyme and concluded that the cavity was empty under ambient conditions.[126] In addition, the study noted that the assumptions used by Somani were incorrect; namely that the free energy of the addition of a water molecule into the cavity from 0 to 1 to 2 is sequential. Considering that the JAWS/GCMC methodology was successful in predicting the occupancy of T4-lysozyme, the IL1β system should provide another test case for the methodology. A 2005 study by Adamek looked at introducing mutations into the hydrophobic cavity and noted the changes in stability and hydration.[129] It was found that 11 different mutations destabilised the system, whilst none of the cavities were found to contain water molecules when studied by X-ray crystallography. The mutations observed did not result in significant structural changes to the protein, suggesting a subtle change in the system dynam-

ics. It is possible that some of the mutations could induce a weak and disordered water network within the cavity causing the protein destabilisation, something which lends itself to be studied by GCMC and JAWS.

6.4.2 System preparation

The same protein preparation and simulation protocols as for N9-neuraminidase were followed for the crystal structure 6L1B.pdb.[130] For the JAWS stage 1 and GCMC simulations, a 166 $Å^3$ binding site region was defined to locate and score the waters.

6.4.3 Wild type simulations

As with the T4-lysozyme system, GCMC simulations were performed at a range of different B values, alongside the standard JAWS stage 1 simulation. At a B value of +15, corresponding to a water binding free energy of 14.26 kcal/mol, a network of 4 water molecules was observed. The observed network matches that described by both Somani and Yin in their MD studies and also the JAWS stage 1 simulation. Upon visual inspection, the network was found to be present across the majority of simulation snapshots, and was chosen to be a good match to the JAWS stage one density. The correlation between the JAWS stage 1 density and the water network can be seen in Figure **??**.

As the B value was decreased to lower potentials the average occupancy dropped dramatically, indicating that under standard conditions of temperature and pressure the occupancy is zero. This initial evidence supports the results obtained by Yin that the occupancy of the cavity under ambient conditions is zero. The population graphs for each B value can be seen in Figure **??**.

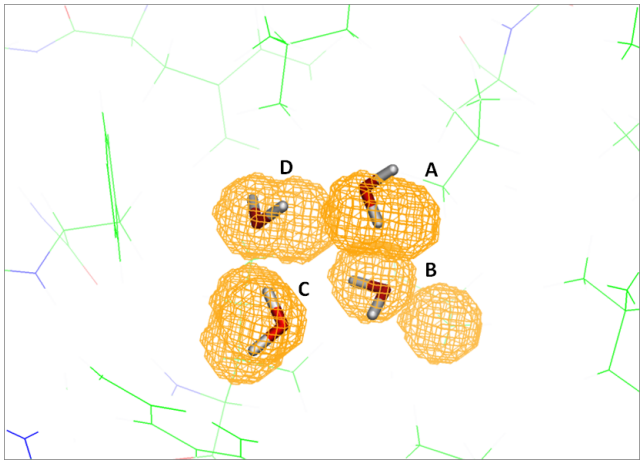


Figure 6.6: Observed water network for IL1β compared to the JAWS stage 1 number density

Using the water network found at B = +15, corresponding to a water binding free energy of 14.26 kcal/mol, each water was then analysed during a JAWS stage 2 simulation. The calculated free energies are shown in table **??**.

Water	ΔG_{bind} (kcal/mol)
A	+1.50
B	+1.78
C	-1.60
D	+3.62

Table 6.2: JAWS binding free energies for the water molecules found at B = +15

Table **??** shows that three of the water molecules in the system display positive binding free energies, whilst another is weakly bound in the system. The positively bound waters suggest unfavourable sites and are hence unlikely to exist in the system. Removal of Wat A, Wat B and Wat D from the system increases the binding free energy of Wat C to +2.18 kcal/mol, suggesting that the water molecule is not stable by itself and hence that the cavity is likely to be empty under standard conditions. This agrees with the population data obtained using

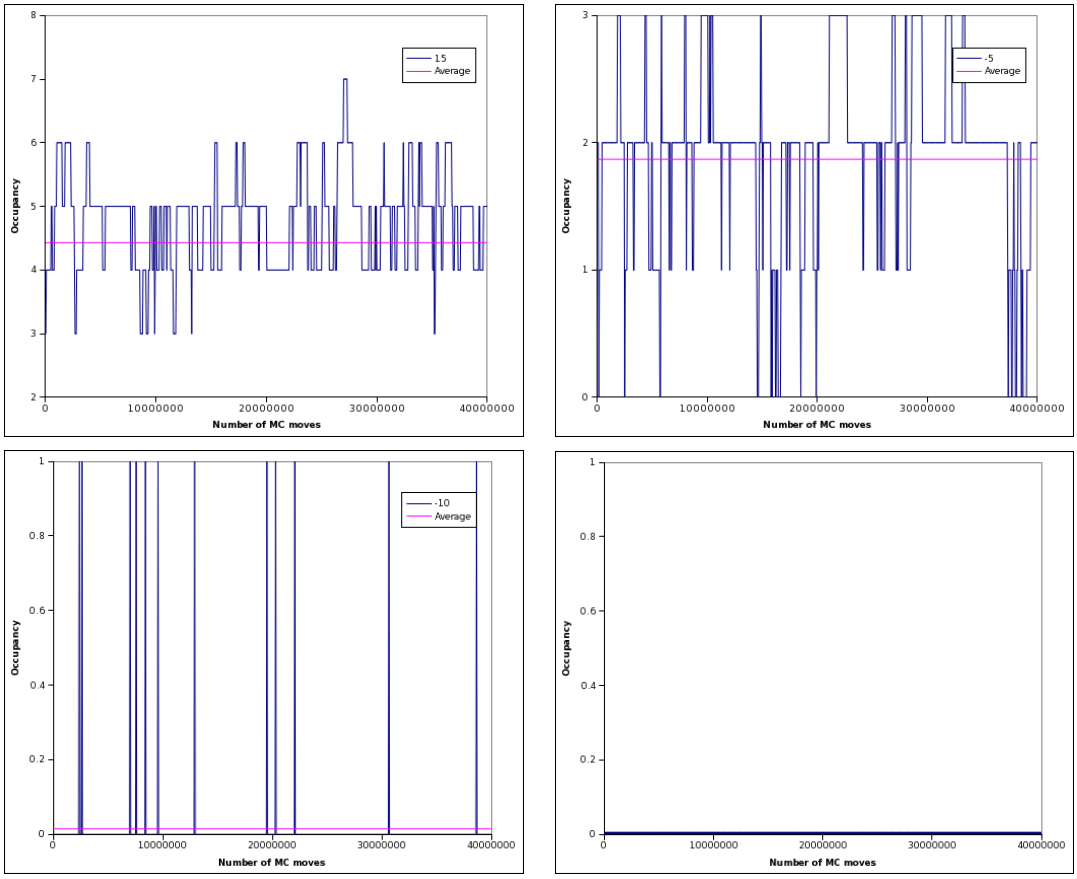


Figure 6.7: GCMC population plots for the IL1β cavity. From left to right: B = +15, B = -5, B = -10, B = -15. These B values correspond to a binding free energy of +14.26 kcal/mol, 2.42 kcal/mol, -0.53 kcal/mol and -3.50 kcal/mol respectively. The value of B can be seen in the legend of each graph

GCMC simulations, shown as the bottom left graph in Figure **??**, where the occupancy of the cavity at a binding free energy of -0.53 kcal/mol is predicted to be approximately zero.

6.4.4 Mutant simulations

Two different mutations studied by Adamek were considered [129]; the F146Y mutation and the WWW mutation, whereby the phenylalanine residues at positions 42, 101 and 146 in the amino acid chain were mutated to tryptophan. GCMC

simulations were performed at three different potentials to assess whether any water molecules observed were positively bound, weakly bound or strongly bound. The same size of binding cavity was identified to allow easy comparison with the wildtype simulations. The average occupancy for these simulations for both mutations in comparison to the wildtype can be seen in table **??**.

B	F146Y	WWW	Wildtype
-5	2.01	1.98	1.87
-10	0.08	0.12	0.02
-15	0.00	0.00	0.00

Table 6.3: GCMC populations for the F146 and WWW mutations as a function of the applied B value, alongside the respective wildtype simulations

Table **??** shows that there is no significant difference between the two mutant structures and the wildtype. As such, the clear indication is that the mutant structures are likely to be empty under ambient conditions. If this is the case, then the destabilising effect which has been observed experimentally is unlikely to be due to water penetrating the hydrophobic cavity of the protein. Although the JAWS and GCMC methods suggest that the cavity is empty, the mutations might be causing subtle polarisation changes within the cavity which cannot be described adequately using a standard MM forcefield. If this is the case then it is of little surprise that the results obtained between the mutant and wildtype structures are extremely similar. One way of testing the polarisation hypothesis is by incorporating the polarisation effects into the water model; effectively tuning the coulombic and Lennard-Jones parameters to account for polarisation.[131]

6.4.5 Conclusions

The studies on the interleukin 1 β protein again highlight the importance of network cooperativity. As with the T4-lysozyme example, the use of JAWS stage 2 on the single water molecules within the cavity suggests that one of the molecules within the cavity has a favourable binding free energy and would be expected to be present under ambient conditions. However the binding free energy of this water molecule is supported by the presence of other water molecules with unfavourable binding free energies and, upon removal of these waters, the previously favourable water is no longer expected to be present. The use of GCMC simulations at different binding free energies reveals that the occupancy at favourable binding free energies is expected to be zero; corroborating the experimental evidence and the modified JAWS stage 2 simulations.

6.5 CDK2 kinase

6.5.1 Biological Relevance

In the human body there are over 500 different kinases [132], with each responsible for a different role. Cyclin Dependent Kinases such as CDK2 are involved in cell cycle regulation, with CDK2 responsible for cell proliferation.[133] When CDK2 is bound to cyclin E an activated form of the kinase is produced, which in turn promotes cell reproduction. Human cancers typically have an overexpression of cyclin E, leading to an abundance of activated CDK2 and hence promotion of tumour development.[133] Targetting CDK2 via drug therapy should render the active conformation inactive, meaning that the cell proliferation pathway fails and tumour growth supressed.

One problem with kinase therapy lies in the fact that 60 % of kinases are sequentially similar.[70, 134] As a result careful control of inhibitor is required for each individual cancer, otherwise undesirable side effects could occur whereby other kinases are also affected. Such a problem is ideally suited to computational approaches, since simulating inhibitor effects *in silico* is advantageous compared to performing expensive *in vivo* experiments. Ligand binding studies have been performed upon CDK2 previously, but the similarity in ligand activities meant that reliable predictions were not achievable.[3] It has been noted that kinase protein-ligand complexes typically contain an unusual interaction; an aromatic CH-O hydrogen bond. This interaction is typically thought to be weak, yet it is found in most structures.[135]

Understanding such an interaction could help to elucidate how inhibitors bind to kinases. In order to look at this problem, JAWS and GCMC simulations were performed to see whether the hydration patterns in the apo form of CDK2 control the binding in the holo form. A range of kinases have been studied by Robinson *et al.* [70], where the WaterMap methodology was used to attempt to rationalise selectivity trends. Although their obtained results helped in justifying certain trends between different ligands, they failed to explain atomistic cooperativity between water molecules and the protein, which is of great interest in this study.

Figure **??** shows the structure of the CDK2 pocket. Of key interest in this study is the nature of the hydration patterns around the binding site. There are three key regions of interest in the pocket; the hinge region, the mouth of the pocket and the activation loop. The positions of water molecules in the pocket will be referred back to these broad positions in the subsequent analysis.

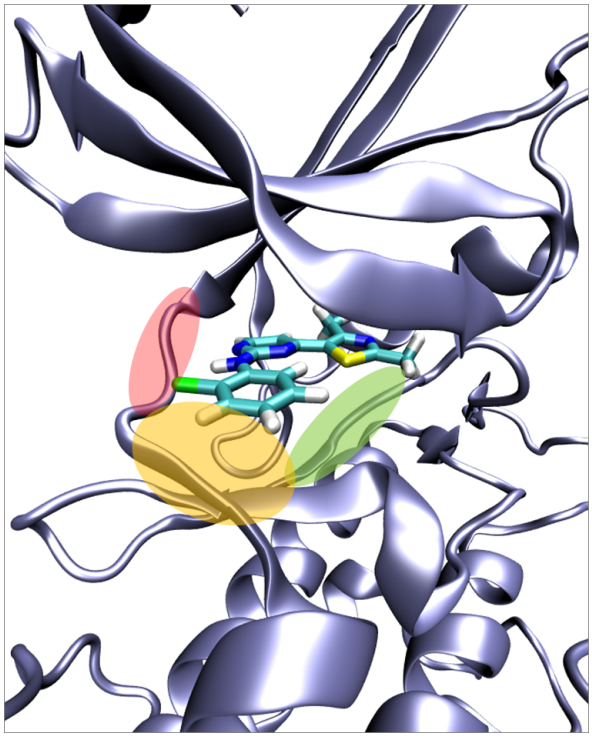


Figure 6.8: Structure of CDK2 kinase, PDB 2WEV, highlighting the key regions around the binding site; The hinge region, mouth of the binding site, and the activation loop.

6.5.2 System Preparation

A previously prepared protein structure of CDK2 was used for this study, which included only protein residues that have a heavy atom within 15 Å of the crystallographic ligand.[3] The same protocols as used for the N9 neuraminidase were followed for the simulations.

6.5.3 Simulations

A JAWS stage 1 simulation was performed upon the CDK2 binding pocket using a θ-threshold of 0.995. The population density, seen in Figure **??** shows that the density spreads across the entire pocket. As a result, three crucial pieces of

information cannot be determined:

1. How many water molecules should be placed in the pocket?

2. Where should these water molecules be placed?

3. Should these water molecules be correlated with each other?

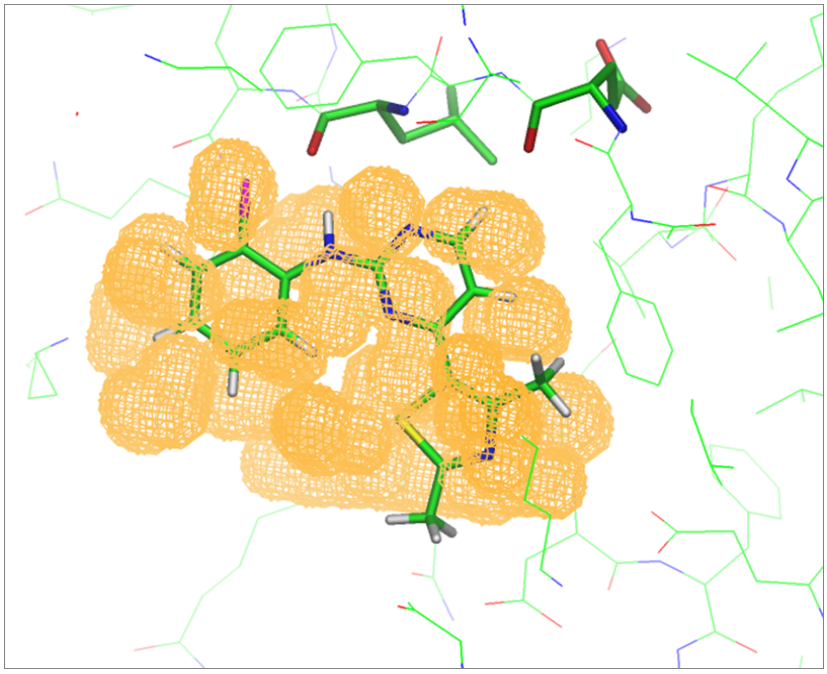


Figure 6.9: JAWS density contours for the CDK2 pocket. The top 90 % of data is displayed, alongside the critical backbone residues E62 and L64

Whilst clustering the JAWS density can allow us to place water molecules on the highest regions of density, it still does not help to answer the first and third questions posed previously. As a result, another approach needs to be applied. As discussed in section **??**, the GCMC method is capable of predicting hydration patterns as a function of the chemical potential. In addition, unlike JAWS, GCMC reveals correlation between water molecules implicitly, since the water molecules can only ever be on or off. As a result, the CDK2 system lends itself to being

studied by GCMC. If the correct chemical potential is chosen then the hydration pattern within the cavity can be easily identified, with the resulting molecules being studied by JAWS stage 2 to ascertain their hydration free energies seperately. GCMC simulations were performed at six different chemical potentials to understand the hydration effects as a function of the applied potential. The normalised number density plots, based upon the oxygen-water coordinates in the simulation snapshots and using a 1 Å grid spacing, can be seen in Figure **??**, with table **??** showing the relationship between the value of B and the estimated binding free energy, alongside the average population collected over 40 M MC moves.

B	ΔG_{bind} (kcal/mol)	$< N >$
-6	+0.28	16.93
-8	-0.90	13.06
-10	-2.09	10.73
-12	-3.27	6.70
-14	-4.45	5.39
-16	-5.64	3.37

Table 6.4: Average populations within the CDK2 binding pocket as a function of the applied B value

From looking at Figure **??**, several patterns can be easily observed. Firstly, the middle of the pocket region is empty at all values of the chemical potential, indicating that hydration in this region is extremely unfavourable since no density is observed. As the chemical potential is lowered to observe more favourable waters the density around the hinge region decreases, with two major water-binding regions observed. The first, found near the mouth of the pocket, is characterised by a high concentration of acidic residues, whilst a second region is found near the thiazoline moiety of the ligand, close to the catalytic lysine of the protein.

Two important chemical potentials studied were those at B = -6 and B = -8. These chemical potentials represent a minimum binding free energy of 0.28

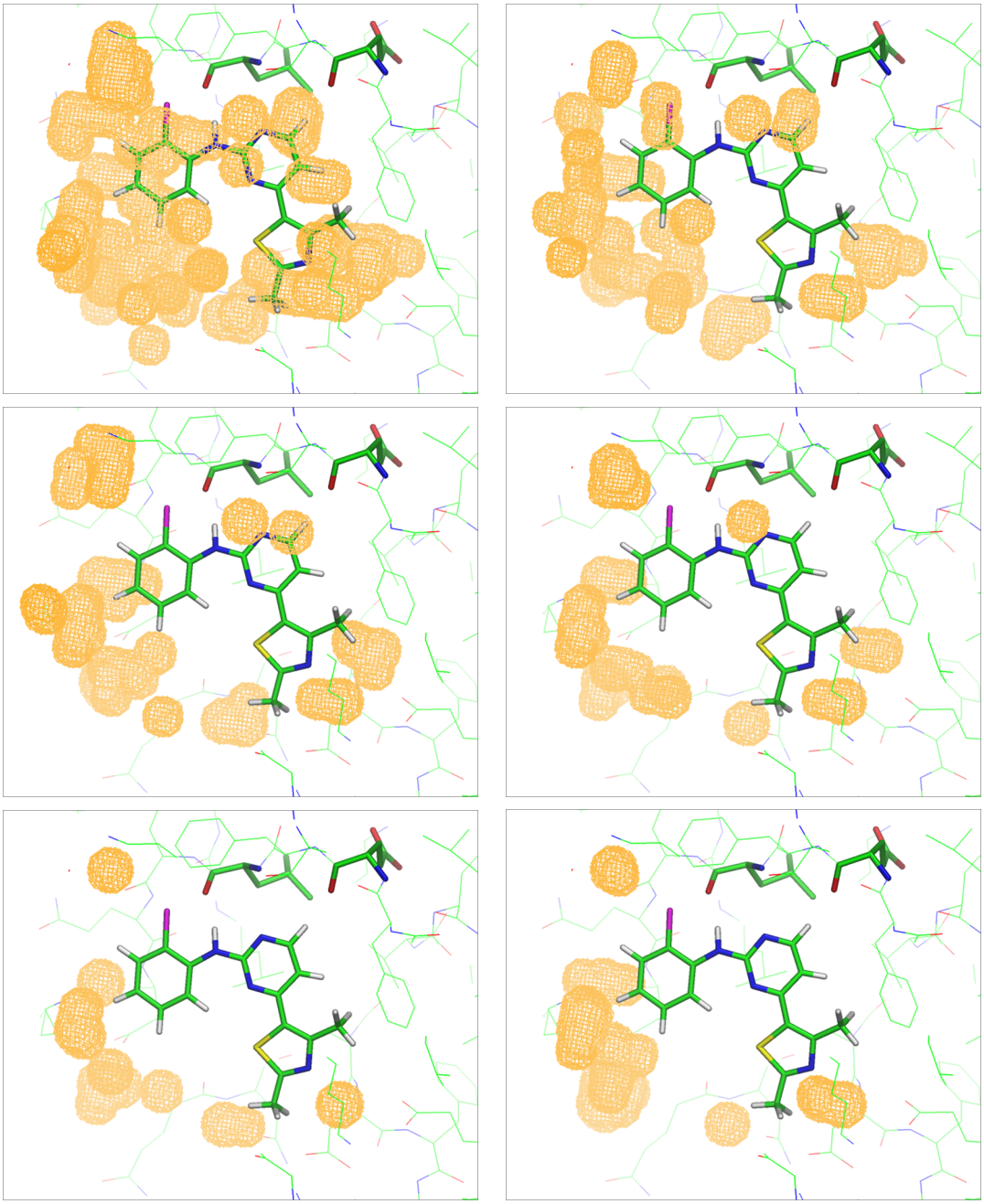


Figure 6.10: Number density plots within the CDK2 pocket as a function of the applied B value. From left to right: B = -6, B = -8, B = -10, B = -12, B = -14, B = -16. The binding free energy associated with these B values can be seen in table **??**. A contouring level of 0.1 is used for all of the plots

kcal/mol and -0.90 kcal/mol respectively, and show the effect upon the system at a slightly positive and negative water binding free energy. By simulating at a slighty

positive binding free energy it can be seen if the cavity accommodates bulk-style waters, whilst simulating at a slightly negative binding free energy should identify water molecules which are expected to be present under standard temperature and pressure conditions.

Having obtained a number density profile, this needs to be matched to a representative snapshot. A clustering based approach is unlikely to be suitable for two major reasons. Firstly, there is a chance that the placed waters could be too close to each other. Secondly, it is important that the waters are in a realistic configuration; something which cannot be guaranteed from clustering. As such each simulation snapshot was examined with respect to the density contours, with the most suitable snapshot chosen for further examination - as performed previously in the T4-lysozyme and interleukin 1 β systems.

Figure **??** shows the placements for the simulations run at B = -6 and B = -8, alongside the population in the pocket as a function of the number of MC moves in the simulation. It is important to note the high degree of variation in the number of molecules throughout the simulations. This is indicative of fluctional behaviour inside the pocket, and suggests that the networks which are formed are likely to be dynamic. Crucially, each water network is a valid representation at the chemical potential at which it is run. As such, it becomes impossible to definitively assign a static network in the system; instead the most we can achieve is examining the network which corresponds best to the density peaks.

Figure **??** shows that running the simulation at a B value of -6 results in an average of 17 molecules in the system, whilst at B = -8 results in 13 water molecules being predicted in the cavity. From Figure **??** it is interesting to note that at more negative, favourable, chemical potentials water molecules tend to be located near the mouth of the pocket, whilst at higher potentials the pocket is more evenly hy-

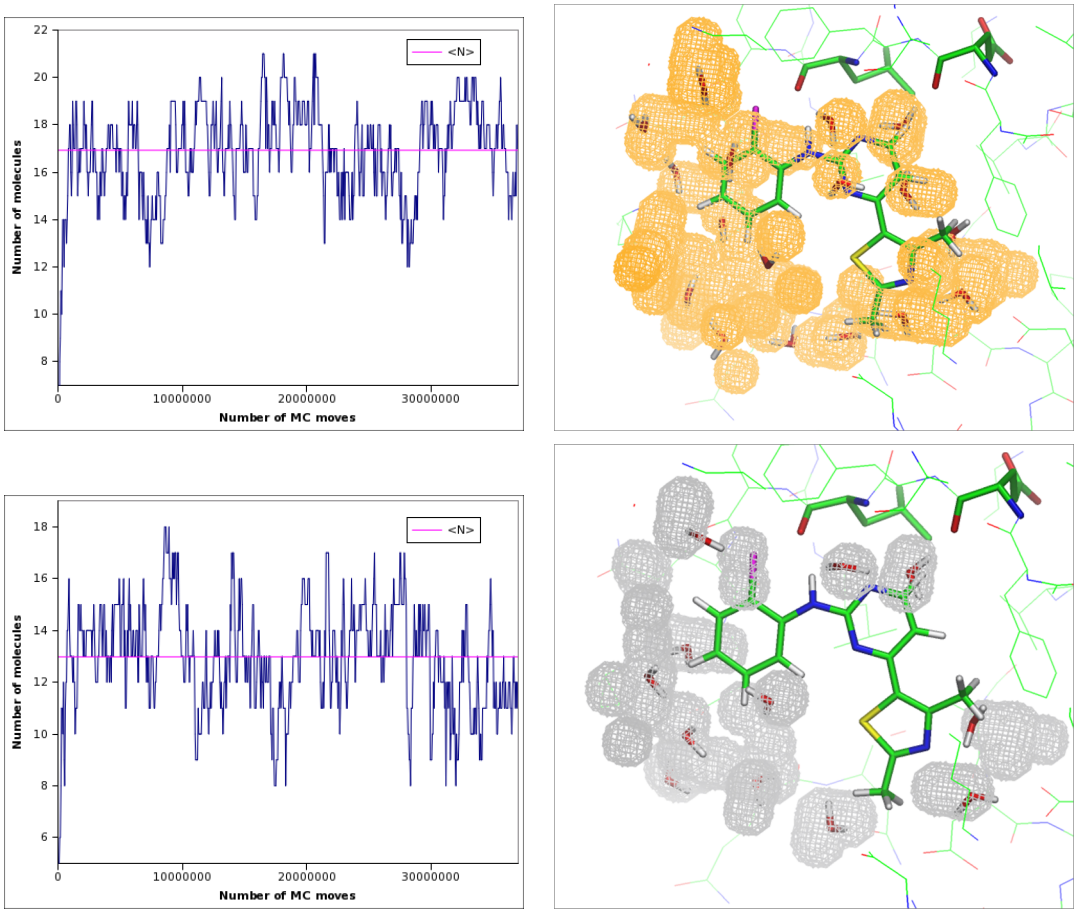
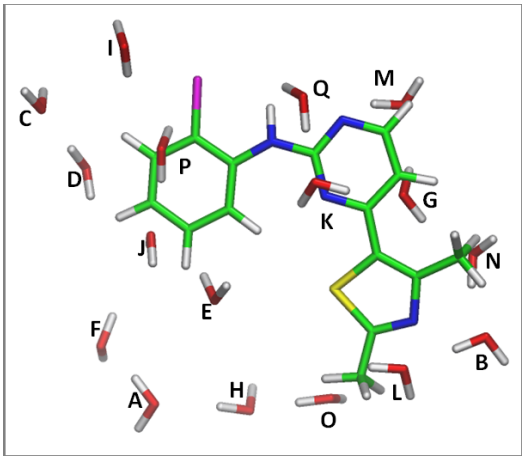


Figure 6.11: Population graphs and representative water locations for B = -6, top, and B = -8, bottom. A contouring level of 0.1 is used for the plots

drated. Crucially, water molecules interacting with L64 and E62 and along the hinge region are identified in the simulations at B values greater than B = -12, corresponding to a binding free energy of -3.27 kcal/mol. Such hydrogen bonding behaviour is found in the vast majority of kinase complexes.[136] The water interacting with L64 acts as an acceptor to a NH interaction from the leucine backbone, whilst the second acts as a donor to the carbonyl oxygen on the glutamate backbone. The latter interaction mimics one which is found in the standard inhibitor, whereby a CH-O hydrogen bond is found between the ligand and the protein.[135] Such an interaction is typically weak, yet it is found in most ligand-kinase com-

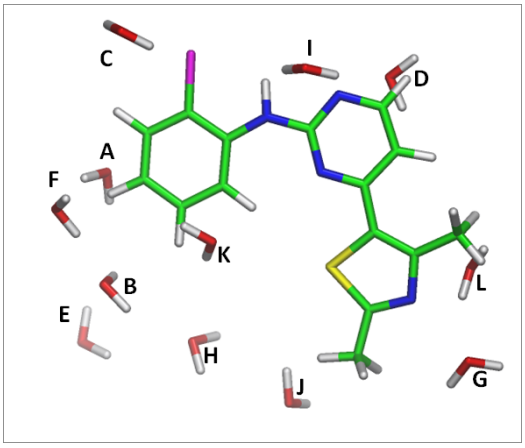
plexes. In order to attempt to understand the hydration patterns observed in CDK2, JAWS stage 2 simulations were performed on the example water configurations found at B = -6 and B = -8. All simulations were run at the same biasing potential of 10 kcal/mol. The results for the B = -6 set can be seen in Figure **??**, whilst those for B = -8 can be seen in Figure **??**.



Water	ΔG (kcal/mol)	Water	ΔG (kcal/mol	Water	ΔG (kcal/mol)
A	-2.71	G	-0.81	M	-5.86
B	-7.97	H	-2.01	N	-1.77
C	-1.73	I	-3.91	O	-1.84
D	-2.59	J	-2.62	P	-1.48
E	-1.55	K	-3.71	Q	-6.81
F	-5.54	L	-1.48		

Figure 6.12: JAWS binding free energies for the water molecules found at B = -6 for CDK2. The error for these binding free energies is ± 0.80 kcal/mol

From comparing the two sets of JAWS stage 2 binding free energies, a clear destabilisation across the hinge waters can be observed as the binding free energy at which the simulation is performed is reduced. The binding free energy of waters M and Q in the B = -6 set are considerably more favourable than the corresponding waters D and I in the B = -8 set, which can be attributed to the stabilisation presence of waters G and K. It is significant to note that the water sites M and Q are not present at a B value of -12 in the GCMC simulations, despite the



Water	ΔG (kcal/mol)	Water	ΔG (kcal/mol	Water	ΔG (kcal/mol)
A	-3.48	E	-5.05	I	-2.69
B	UNCALC	F	-3.88	J	-4.70
C	-1.84	G	UNCALC	K	0.27
D	-2.35	H	-3.19	L	-1.83

Figure 6.13: JAWS binding free energies for the water molecules found at B = -8 for CDK2. UNCALC signifies waters for which free energy estimates cannot be obtained at a biasing potential of 10 kcal/mol due to insufficient $\theta < 0.05$ transitions. The error for these binding free energies is $\pm$ 0.80 kcal/mol

energies of waters in the B = -6 GCMC simulations suggesting that they should be strongly bound to the protein. This suggests that there is a stabilisation effect in the JAWS stage 2 simulations which is not present during the GCMC simulations at different potentials.

It is interesting to note that Wat K exhibits a binding free energy of -3.71 kcal/mol in the B = -6 set, posing the question as to why it is not present in the B = -8 set. In order to understand this, the marginal binder Wat G was identified as a critical water. The binding free energy of -0.81 kcal/mol is indicative of a water with marginal occupancy, and it is probable that this water is not always present under standard conditions. As such the effects of removing this water was investigated, and the binding free energy of Wat K recalculated. This resulted in a drop in the binding free energy of Wat K to -1.27 kcal/mol; suggesting that, in the

absence of Wat G, Wat K is expected to also exhibit marginal occupancy under standard conditions. The removal of waters G and K from the system leaves a configuration of waters around the hinge which is consistent with the B = -8 set and explains why waters M and Q are not present at B values less than -12 in the GCMC simulations.

The destablisation in the network through the removal of a weakly bound water has important consequences for rational drug design. If weakly bound water molecules can be targetted, then the surrounding molecules can be potentially destabilised to such an extent that these too are now easily displacable. As such the removal of one weakly bound water reduces the desolvation cost of many water molecules in the site, meaning that these waters can be easily displaced by hydrogen bonding groups in the ligand. This could potentially lead to improved potency since there is a minimised entropic and enthalpic loss upon displacing such waters in the pocket, allowing strong gains to the binding affinity through protein-ligand interactions.[42]

The change in the binding free energies of the water molecules as a function of the applied chemical potential is also highly significant. A single simulation might give the impression a particular water is not displacable, whilst on further examination this site could be easily displaced due to the presence of weakly bound supporting neighbours. This demonstrates a potential flaw in methods such as WaterMap which employ only a single simulation, and hence the free energies which are obtained from the method are not necessarily going to mirror true biological conditions.

From further analysis of the B = -8 set, the nature of the inhibitor binding mode can be easily understood. As mentioned previously, the majority of kinase inhibitors exhibit a weak CH-O interaction. It can be seen that Wat D directly

mimics this interaction of the inhibitor, suggesting that for a binding event to occur the water interaction must be replaced by the inhibitor. The binding free energy of Wat D is relatively low at -2.35 kcal/mol, which explains why a stronger protein-ligand interaction is not required to displace the water molecule.

It is important to note that all of the JAWS stage 2 simulations were performed at the same biasing potential of 10 kcal/mol. This value was chosen so that the changes upon removing waters could be easily observed. For most of the water molecules this biasing potential induced more $\theta > 0.95$ states than $\theta < 0.05$ states. This ensured that adjacent water molecules did not drift away during the simulation, which was observed when higher values of the biasing potential were utilised. If the hydrogen bonding network is allowed to break apart throughout the simulation then the calculated free energy is not indicative of the original starting configuration, and as such the end points between different sets of simulations are not the same. Such behaviour was observed in analogous RETI double-decoupling simulations, whereby adjacent water molecules to the solute water whose binding free energy was being calculated were observed to drift away as the electrostatic and Lennard-Jones terms were decoupled since the stabilising effect of the water keeping them in place was lost.

6.5.4 Corroborating the results

As mentioned earlier, one drawback of the current placement scheme is that it relies upon visual inspection of the simulation snapshots. For longer simulations this is clearly an intractable task, and other methods need to be sought. One possibility lies in the generic site calculation scheme proposed by Mezei.[137] This approach identifies *n* generic solvation sites, and then iteratively loops over all of the simulation frames to find the snapshot which best represents the solvation

sites.

The MMC program [138] was used to run the generic solvation site calculations upon the output generated from the B = -6 simulations. The placed waters compared to the visual inspection can be seen in Figure **??**.

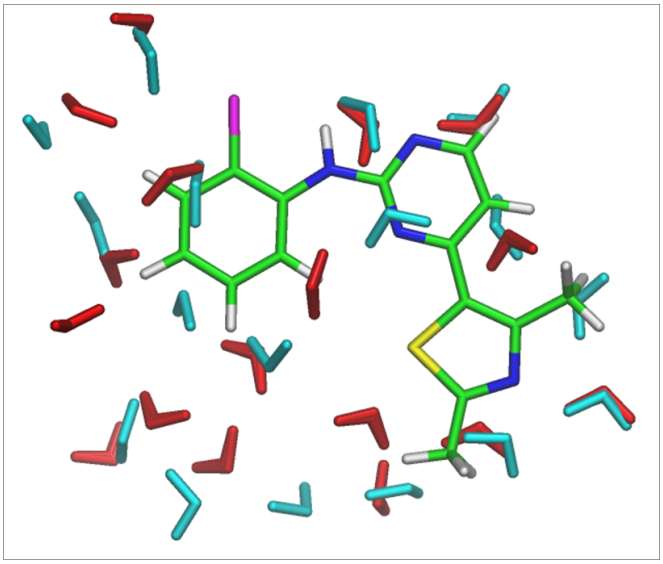
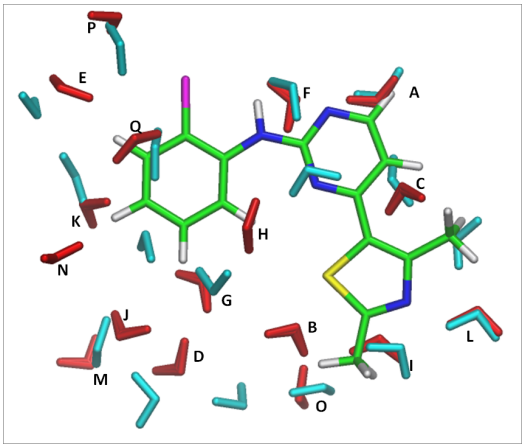


Figure 6.14: Placed waters from the MMC generic site algorithm (red) against a visual snapshot inspection (cyan)

Figure **??** shows that the MMC generic site algorithm predicts a similar hydration pattern to the visual inspection. There are however a few subtle differences. First, only three molecules are found in the upper hinge region in the MMC output, compared to the quartet found in the visual inspection. Second, the position of the waters around the mouth of the binding pocket differ between the methods. In order to understand whether similar binding free energies are obtained between the patterns, JAWS stage 2 simulations were performed upon the MMC-placed waters.

The lack of a fourth water molecule in the hinge region results in a destabilisation in the Wat Q position. Indeed, the binding free energy of waters A and F



Water	ΔG (kcal/mol)	Water	ΔG (kcal/mol	Water	ΔG (kcal/mol)
A	-3.58	G	-2.67	M	-7.41
B	-0.75	H	-1.40	N	-6.65
C	-0.78	I	-0.30	O	-1.11
D	-1.93	J	-3.84	P	-2.59
E	-1.69	K	-4.72	Q	-2.31
F	-1.82	L	-2.42		

Figure 6.15: JAWS stage 2 binding free energies for the MMC generic site water molecules

are now similar to the corresponding waters D and I found in the B = -8 set of molecules. For the vast majority of water molecules, the binding free energies are broadly consistent between the two sets, suggesting that the MMC methodology is a valid method of placing water molecules alongside the visual inspection approach.

As another comparison, and to check that the observed results are realistic, a MD study was performed upon the protein. The simulation was performed using gromacs 4.5.1 [139] using the amber99SB forcefield [140] to allow an accurate comparison with the JAWS and GCMC results. The protein was protonated using the pdb2gmx tool in gromacs.

MD protocol

The MD simulation was performed with cubic periodic boundary conditions, a particle mesh Ewald treatment of electrostatics (with an interpolation order of 4), a 2 fs time step, and a 10 Å cutoff. All bonds were constrained to equilibrium lengths using the LINCS algorithm. The simulation was performed in the NPT ensemble using the V-rescale thermostat and the Berendsen barostat, and run using a temperature of 300 K at a pressure of 1 bar. Water molecules were modelled using the TIP3P model.[44] Three Cl- ions were added to neutralise the system.

A 15 ns simulation was performed with data collected for the last 5 ns. The resulting oxygen coordinates of waters in the region of interest were then analysed in the same method as the GCMC/JAWS simulations. The contoured results can be seen below in Figure **??**.

Figure **??** shows that, like the GCMC simulations in Figure **??**, the major hydration sites are identified in and around the hinge region. It can be seen that the same four membered ring observed in the GCMC simulations, seen in Figure **??**, around the glutamate and leucine hinge residues is observed in MD, indicating that this is a key motif in the hydrogen bonding network. In addition, density is found near D103 and Wat B, indicating that this is likely to be a region of favourable hydration. The similiarity between the GCMC simulations and the MD result is highly encouraging, and serves to show the effectiveness of the GCMC method. Whilst performing a simulation at a set value of B for 40 M MC moves takes approximately 8 CPU hours, collecting the data from a MD simulation takes 3 orders of magnitude longer, requiring 1700 CPU hours. Since a MD simulation does not give any energetic analysis, a GCMC simulation appears to be vastly superior if the hydration of cavities is to be studied. In addition, an MD simulation does not give any information upon the correlation between sites if the data is

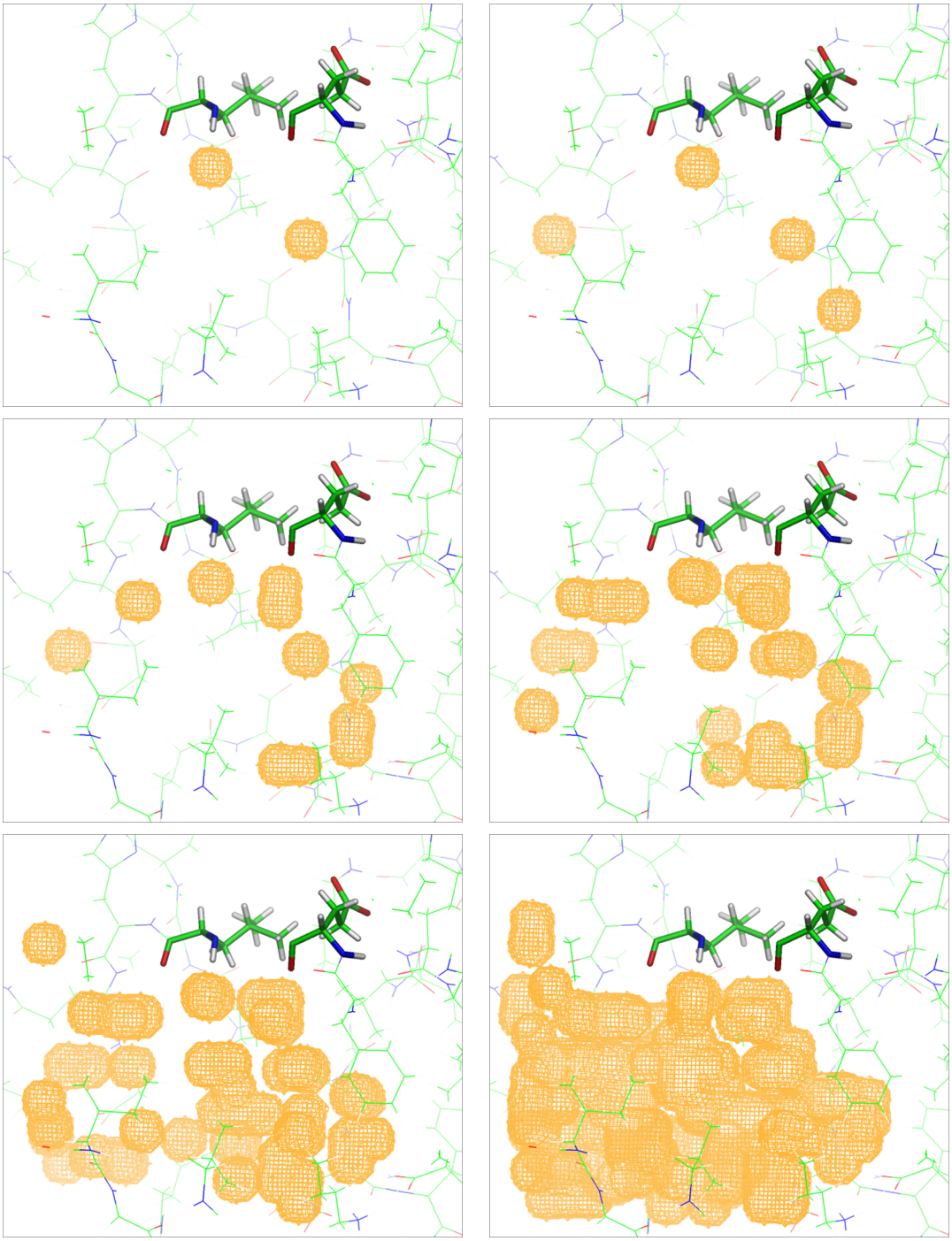


Figure 6.16: Contoured MD positions for a 5 ns simulation of the apo CDK2 cavity. From left to right: Top 10 % of data, top 30 %, top 50 %, top 70 %, top 80 %, top 90 %. The critical hinge residues, E62 and L64, are shown in bold

clustered. Clustering complex data such as the bottom right window in Figure **??** is likely to lead to an erroneous water network, since neither the position or number of waters is known.

Comparison with Src kinase

The kinase study performed by Robinson [70] utilised the WaterMap methodology to look at the hydration in Src kinase. The methodology identified a quartet of waters around the hinge region, analogous to the quartet of waters found when the CDK2 system was simulated at B = -6. Interestingly *all* of these waters were defined as unstable in the WaterMap methodology, with the unfavourable free energies assigned to both enthalpic and entropic considerations. In comparison, the GCMC/JAWS method suggested that two of the molecules are extremely favourable in a quartet, with the other two, Waters G and K, unlikely to be present under standard conditions.

This clearly identifies a difference between the GCMC/JAWS method and WaterMap. In the GCMC/JAWS method unfavourable waters are removed from the system, since it is assumed that these waters will not be present under ambient biological conditions. In comparison, WaterMap retains the unfavourable waters in their calculations. This poses the question as to which method is more reliable. The WaterMap methodology retains unfavourable waters since it is claimed the energetic cost of creating a vacuum is greater than the free energy gain for displacing the unfavourable water molecules. The GCMC/JAWS approach calculates absolute binding free energies, whilst the free energies calculated by the WaterMap method do not have a similar reference state. As such, it becomes challenging to reliably compare the two methods.

6.6 Pim-1 kinase

6.6.1 Biological Relevance

Like CDK2, Pim-1 is a kinase which has been implicated in a wide variety of different cancers, mainly oral and prostate cancers.[136] The normal role of Pim-1 is in prosurvival; the protein binds to the pro-apoptotic protein Bad, causing phosphorylation of Bad on S122 and rendering Bad inactive.[141] As a result the Bad protein is not able to cooperate in the apoptotic cycle, leading to the survival of the cell. Over expression of Pim-1 in cancers means that the cancerous cells are less prone to apoptosis, leading to tumour growth. As a result Pim-1 has started to become an accepted drug target, with inhibitors now being described in the chemical literature.[136]

The Pim-1 kinases have a unique hinge region sequence which seperates them from other kinases. As previously discussed, inhibitors of CDK2 rely on hydrogen bonding interactions with the hinge residues. The unique sequence of Pim-1 kinases places a proline residue along the hinge region. This induces a twist in the hinge region, preventing the formation of hydrogen bonds along the hinge.[136] As a result, inhibitors of Pim-1 typically target the region occupied by the catalytic lysine and acidic residues and experience no interactions with the hinge.

The marked change in hydrogen bonding patterns compared to CDK2 provides an intriguing case for the GCMC/JAWS methodology to predict the water network for a different kinase system. Since the protein structure for Pim-1 is significantly different to that of CDK2, it is hoped that the methodology should pick up the new water network and highlight the lack of hydrogen bonding along the hinge.

6.6.2 System Preparation

The protein structure used for this study was 3DCV. The same protein preparation and simulation protocol as the N9-Neuraminidase system was followed. A pocket of 640 $Å^3$ was identified for the GCMC simulations. The binding pocket was defined as the region in which the native ligand is situated, with an angstrom added in all directions. The binding pocket of Pim-1 highlighting the key binding regions can be seen in Figure **??**.

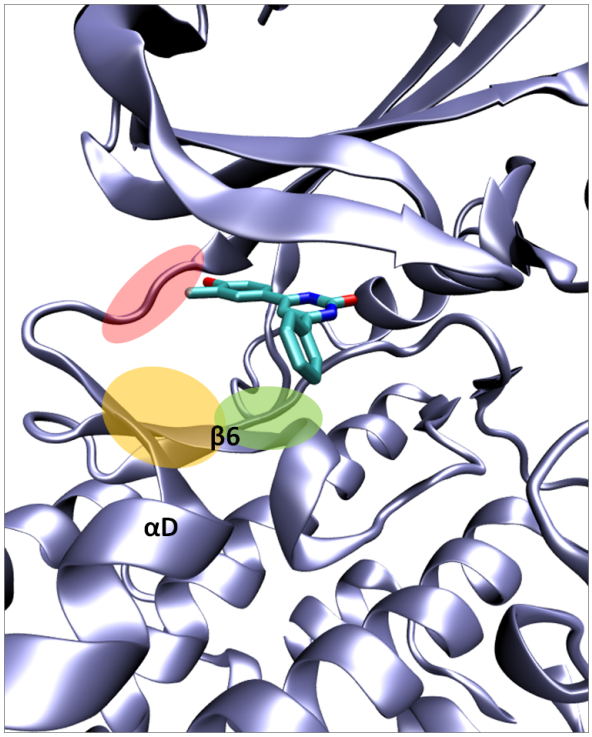


Figure 6.17: Structure of Pim-1 kinase, PDB 3DCV, highlighting the key regions around the binding site; The hinge region, mouth of the binding site, and the beginning of the catalytic loop.

6.6.3 Simulations

Since the JAWS simulation performed upon CDK2 lead to difficulties in analysing the resulting populations, GCMC simulations were preferred for the Pim-1 system. As with CDK2, a range of different chemical potentials were employed upon the system to understand the water network as a function of the applied potential. Table **??** shows the relationship between the value of B and the estimated binding free energy for the different simulations using this B value. The GCMC density plot for 3DCV taken from the GCMC simulations can be seen in Figure **??**. This density plot was based upon the oxygen-water coordinates in the simulation snapshots, using a 1 Å grid spacing.

B	ΔG_{bind} (kcal/mol)
-6	1.03
-10	-1.33
-12	-2.52
-16	-4.88
-18	-6.07
-20	-7.25

Table 6.5: Relationship between the binding free energy and B for the 3DCV Pim-1 system

In the GCMC simulations, the B = -10 simulation corresponded to a binding free energy of approximately -1.30 kcal/mol, meaning that any waters observed at this potential should be, at least, weakly bound to the system. The average population across the simulation for 3DCV was found to be 10.85, with a standard deviation of 1.5. A frequency plots showing the number of molecules at B = -10 can be seen in Figure **??**.

Based upon the frequency plot and the ease of fitting a representative snapshot to the number density, 10 molecules were chosen to represent the B = -10 density.

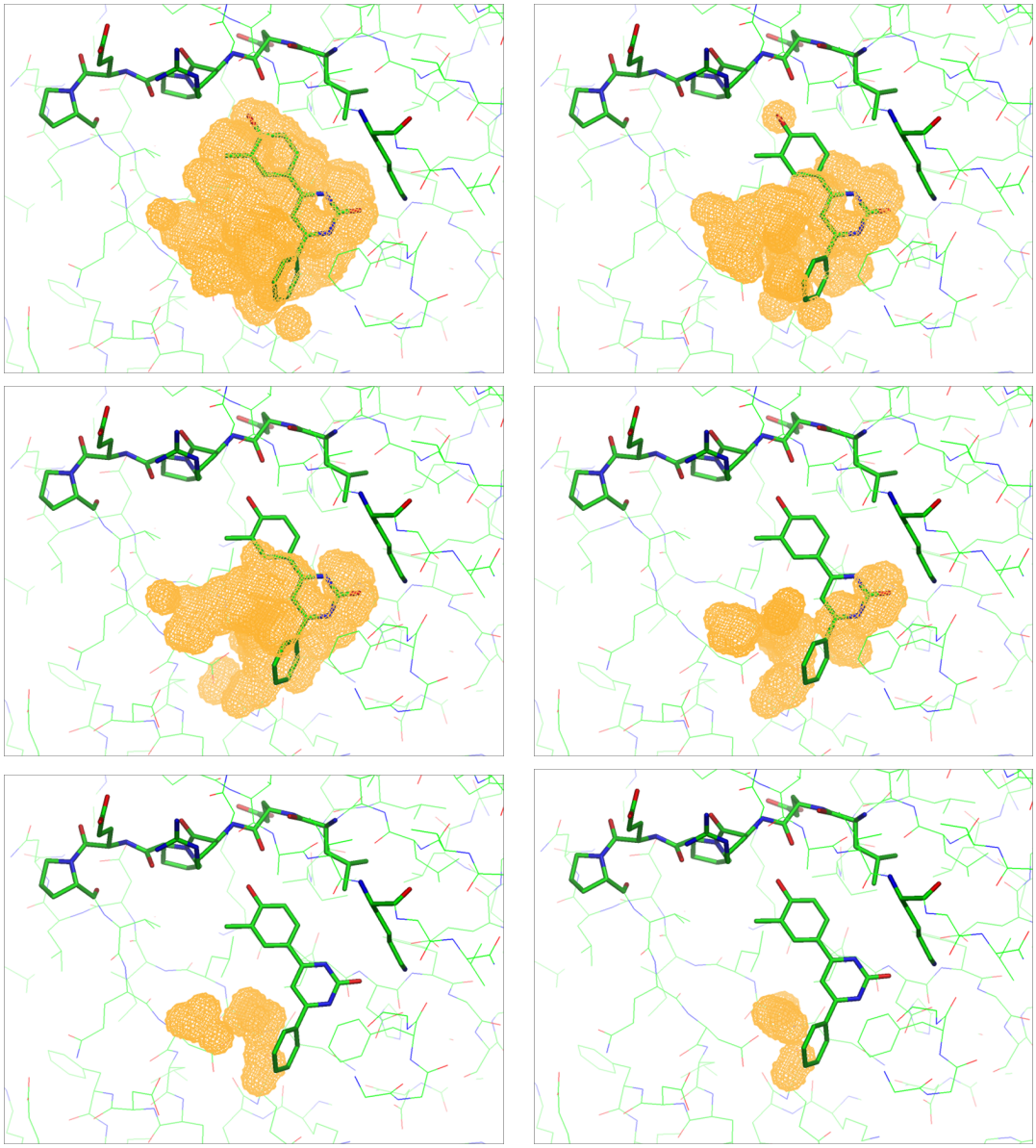


Figure 6.18: GCMC density plots for 3DCV at different B values. From left to right: B = -6, B = -10, B = -12, B = -16, B = -18, B = -20. The binding free energies for these B values can be found in table **??**

The same contouring and placement scheme as used for CDK2 was applied, with the water positions for each system shown in Figure **??** alongside their population graphs.

Qualitative analysis of the water network shows that the majority of the water

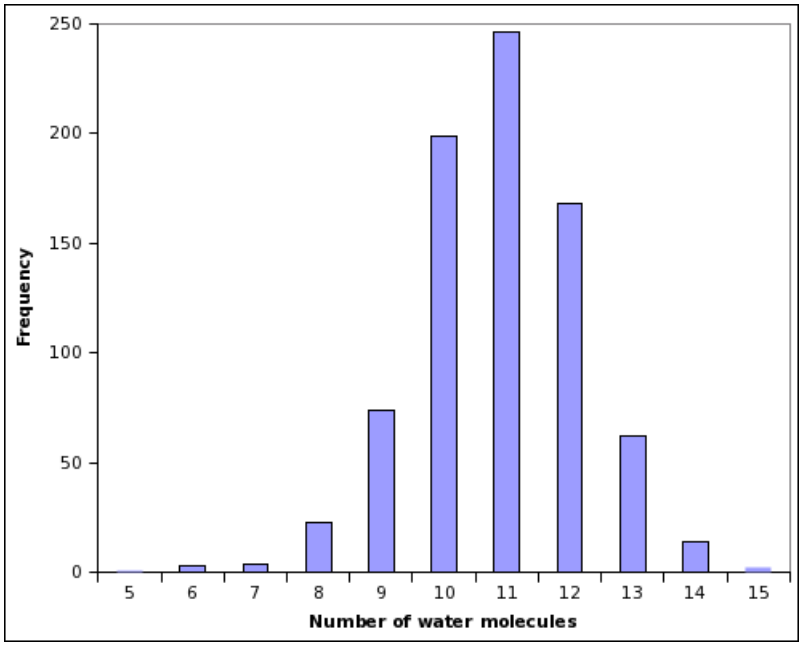


Figure 6.19: Frequency plots for 3DCV at B = -10

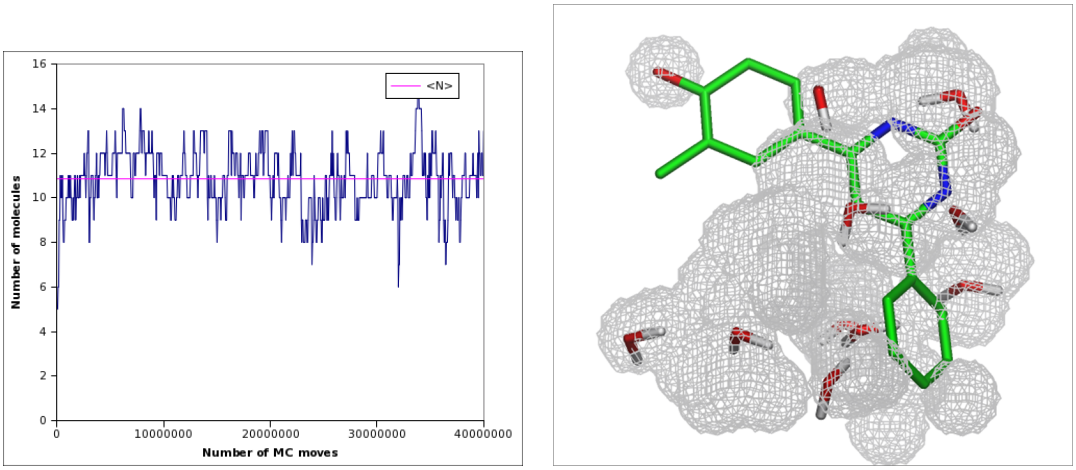


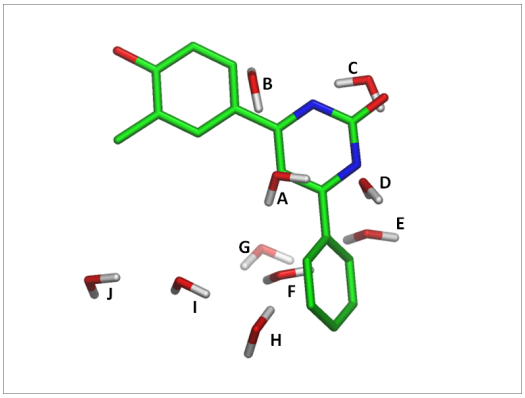
Figure 6.20: Population graphs and placed water locations using the B = -10 simulations for 3DCV. Although a distinct water contour is found close to the hydroxyl group of the ligand, no snapshot was found which adequately described the density of the system and also this site

molecules can be found in three different regions; the eastern region of the ligand close to the catalytic lysine, close to the beginning of the catalytic loop, and at the mouth of the pocket near to the β6 sheet and the αD helix.[136] Crucially, when

analysed alongside Figure **??** it can be seen that no significant water density is found at the hinge region - showing that the method has successfully predicted a different hydration pattern to CDK2. Each of the water sites was subjected to a JAWS stage 2 simulation using a biasing potential of 10 kcal/mol, with the results analysed in the following sections.

JAWS stage 2 results

The calculated free energies for the water molecules in 3DCV, alongside a ligand representation is shown in Figure **??**.



Water	ΔG (kcal/mol)	Water	ΔG (kcal/mol
A	-6.55	F	-8.18
B	-2.67	G	-3.85
C	-7.14	H	-3.59
D	-8.23	I	-1.69
E	-6.59	J	-2.37

Figure 6.21: JAWS binding free energies for the water molecules found at B = -10 using protein structure 3DCV. The error for these binding free energies is approximately ± 0.80 kcal/mol

Figure **??** shows that all of the binding free energies calculated are negative, meaning that no recalculations are required to take into account marginal binders. A study by Qian [136] highlighted the importance of ligand binding to the catalytic lysine in Pim-1. The strong binding free energy of Wat C mimics this in-

teraction, suggesting that the interaction between the ligand carboxyl group and the lysine must be extremely strong to displace the water. Indeed, the Qian study showed that inhibitors which do not have the carboxyl group, or a strong electrostatic interaction, are essentially inactive. Wat D forms part of a strongly bound network between waters A, C and E, and also the sidechain aromatic ring of F18. Waters E, F, G and H are all found close to the catalytic loop / αD helix network of aspartate and glutamate residues, explaining their strongly bound nature. Indeed, Wat H is found in close proximity to W1009 in the crystal structure. Figure **??** shows the locations of the water molecules with respect to the protein, and highlights the key elements of the network.

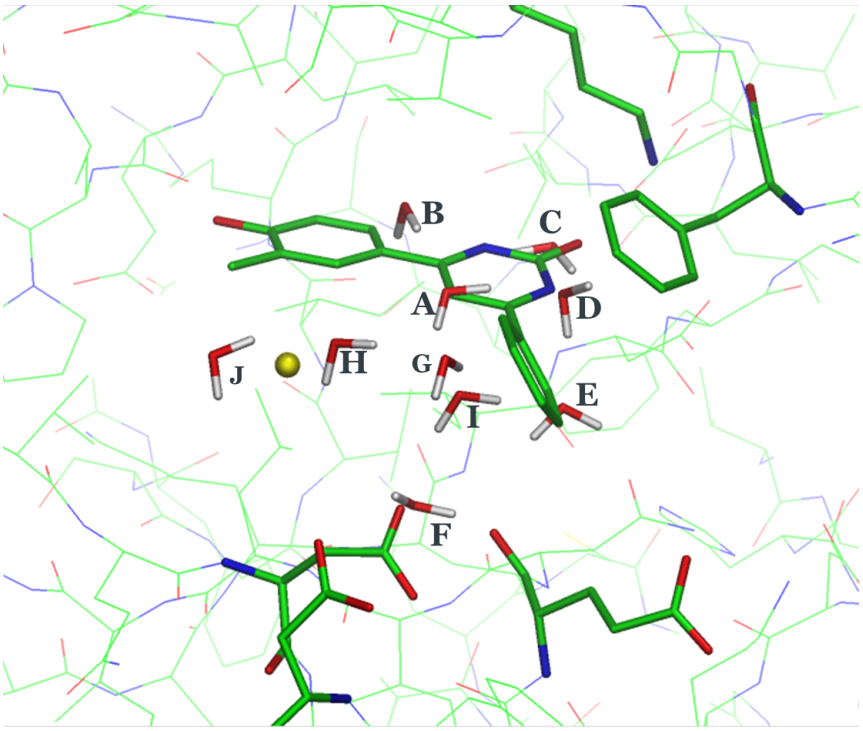


Figure 6.22: The location of the waters in the 3DCV B = -10 network viewed alongside the native ligand and the nearby protein residues. Wat 1009, a crystallographic water, is shown in yellow

Although the JAWS stage 2 energies make intuitive sense and agree qualita-

tively with the GCMC plots, they do not match with the expected average GCMC binding free energies. For example the binding free energy of Wat D is anticipated to be -8.23 kcal/mol by JAWS stage 2, yet a water contour, as seen in Figure **??**, is not observed at a binding free energy of -6.07 kcal/mol in the GCMC simulations. As such, this is suggesting network stabilisation in the system, although this network is stronger than that observed in the CDK2 study. In order to prove this hypothesis Wat B, the weakest binder in the major network of waters, was removed from the calculations and the JAWS stage 2 simulations rerun. The effects this caused on the binding free energies of the nearby waters can be seen below in table **??**.

Water	Old ΔG_{bind} (kcal/mol)	Run without water(s)	New ΔG_{bind} (kcal/mol)
A	-6.55	B	-2.95
C	-7.14	B	-5.33
C	-7.14	A, B	-5.41
D	-8.23	B	-6.79
D	-8.23	A, B	-3.97

Table 6.6: Destabilisation effects upon the 3DCV water quartet

From looking at the recalculated free energies in Table **??**, the GCMC results are now in excellent agreement with the revised JAWS stage 2 binding free energies. When Wat B is eliminated from the system, the knock-on effect is to reduce the binding free energy of Wat A to a comparable figure. As such, it can be deduced that the removal of Wat B from the system will also result in the removal of Wat A. In the absence of Waters A and B, the binding free energies of Waters C and D are now -5.41 and -3.97 kcal/mol respectively, explaining why these waters are not found at GCMC simulations where the binding free energy of the system is -6.07 kcal/mol or less.

It is interesting to note that the binding free energy of Wat H is -3.59 kcal/mol,

yet a contour spot is found at a GCMC binding free energy of -7.25 kcal/mol. In order to explain this, Wat I, a weaker binder with a binding free energy of -1.69 kcal/mol, was removed from the system and the JAWS stage 2 binding free energy of Wat H recalculated. This resulted in the binding free energy of Wat H changing to -6.29 kcal/mol, suggesting that the removal of a nearby water molecule can allow other molecules to adopt more favourable locations. The other major contour in the GCMC plots corresponds to the site occupied by Wat F, with a JAWS binding free energy of -8.23 kcal/mol and characterised by a strong interaction between D97 and the backbone oxygen of E140.

6.7 Using apo hydration site analysis to predict water displacement

Through analysis of the CDK2 and Pim-1 water networks, it can be seen that the observed water networks which correspond to weakly bound or unbound waters generally mirror the location of hydrogen bonded atoms of the crystallographic ligand. Qualitatively this offers a route for shape-based rational ligand design, following a routine analagous to that proposed by Homans in the design of hydrophobic ligands.[42] However, one caveat to this approach is assuming that the waters will be displaced by the ligand. In order to use the hydration patterns to quantitatively decide whether a water is displaced or retained, the effect of the network must be considered.

As exemplified by the CDK2 water network, the binding free energies of water molecules can change as the local environment is changed. When a weakly interacting network is allowed to influence the binding free energies, it can be seen that the two waters appear to be favourable, and display characteristics sim-

ilar to conserved waters.[40] However when the nearby waters are removed from the system, as would be expected under biological conditions upon ligand entry to the binding pocket, the binding free energies of the hinge waters are reduced, suggesting that they are now displacable. It is this shift in the binding free energy, herein referred to as Δ, which gives information regarding how water molecules in networks can be treated in ligand design. Δ can be estimated as the binding free energy before network displacement minus the binding free energy after network displacement.

Through a combination of GCMC simulations and JAWS-2 binding free energies obtained on a network of waters which are weakly bound to the system, the medicinal chemist can gain valuable knowledge of how to treat water networks. Using the CDK2 example, it can be seen that the two hinge waters display a Δ value of approximately 4 kcal/mol upon the removal of the nearby waters - indicating that they are destabilised upon network displacment. This suggests that, upon the removal of the weakly bound waters, the desolvation cost of the hinge waters is significantly reduced. In order to recoup Δ through a binding event the medicinal chemist can choose to either replace these waters through hydrogen bonding with the hinge, or choose to incorporate the waters into the ligand design, but in such a way that the Δ is recouped through interactions with the ligand. In this instance it becomes more advantageous to displace the waters through a protein-ligand interaction, since the interaction energy between the ligand and the protein can easily outweigh the interactions formed between the water molecules on the protein backbone - an interaction which has been shown to be typically weak through the analysis of WaterMap-derived free energies.[142]

In order to incorporate waters into ligand design *a priori*, two features need to be identified. Firstly, the water molecule must be stable in its original net-

work and display a binding free energy indicative of being retained upon ligand binding. Secondly the magnitude of Δ must either be small, suggesting that the water is part of a strong and stable network, and/or the binding free energy of the destabilised state must be sufficiently stabilised by the ligand. In the case of the CDK2 waters, the two hinge waters fail this criterion. The only sites in the CDK2 which are candidates for water retention are the ones close to the mouth of the pocket and at the back of the pocket, since these sites are not affected by network destabilisation.

Using these rules, the nature of the Pim-1 ligands can be easily explained. The network of waters A, C and D in the 3DCV structure are all stabilised in the presence of the weak binder Wat B, and exhibit Δ values of > 1.7 kcal/mol when the water network is disrupted. The resultant free energies when the network is destabilised are all less than 7 kcal/mol, highlighting that these waters can be displaced by a strong protein-ligand interaction. This is achieved in Pim-1 inhibitors through the use of a ligand carbonyl group, targetting the catalytic lysine. The waters along the catalytic loop and the $\beta 6$ sheet are all stable in the absence of other waters due to strong interactions with the nearby acidic residues, and hence exhibit low Δ values with favourable binding free energies. These waters are commonly exploited in Pim-1 ligand design, and incorporation into a protein-ligand complex will help to stabilise these water molecules further.

One question which can arise is since GCMC gives information about water network stability, do we need to worry about the JAWS-2 energies? GCMC simulations at different potentials highlight the favourable regions in the protein pocket as a function of the binding free energy, and therefore allow quick identification of stable hydration sites. Hence, should this be enough of a guide to dictate whether we incorporate waters or not? Although this information is highly useful, it does

not give us an idea about how much particular waters are (de)stabilised, something revealed through Δ. For example, the amount of stabilisation which waters M and Q exhibit in the B = -6 CDK2 network cannot be estimated through GCMC, and relies on the calculation of JAWS waters.

A future, and potentially revealing, application of the GCMC and JAWS methods would be to use the methods on both the apo and holo forms of the same protein. Such a calculation would enable the identification of water locations both pre and post binding. The approach would allow a Δ difference map to be constructed, based on the GCMC simulations, to help identify sites which are affected by the binding process. Such an approach could identify regions where water molecules are stabilised or destabilised upon ligand binding, and will provide opportunities to target these waters in lead development. Similarily to the idea of Δ as a function of the network disruption in the apo phase, Δ could be used to guide whether waters are conserved or displaced during the binding process.

6.8 Understanding the role of water in inhibitor binding to Chk-1 kinase

6.8.1 Biological Relevance

Checkpoint kinase 1 (Chk-1) is a serine/threonine kinase involved in mediating the response to DNA damage in cells.[143] Upon DNA damage Chk-1 is activated and phosphorylates Cdc25C, which then inactivates CDK1. Inactivation of CDK1 results in cell cycle arrest, allowing the cell to repair damaged DNA prior to entering mitosis. In healthy cells the p53 pathway is the major route for inducing DNA-damaged cell apoptosis, however 50 % of cancerous cells contain mutations

in the p53 gene making this pathway unavailable.[144] In such cells only the G2 pathway is able to prevent the replication of damaged DNA, providing a route for targetting cancerous cells over healthy cells. Inhibition of Chk-1, alongside radiotherapy, results in cancerous cells entering the mitosis cycle prematurely with damaged DNA and cell death.

6.8.2 Simulations

Two protein-ligand complexes provided by Vernalis were used as the basis for this study. The same protein preparation procedure and simulation protocols as for N9-Neuraminidase were followed. The first complex 5BT.pdb, exhibits a ligand IC_{50} of 0.07 μM, whilst the second complex 5CH.pdb exhibits a ligand IC_{50} of 0.01 μM. Figure **??** shows the high degree of similarity between the ligands, varying by the switch of the connectivity in the amide group.

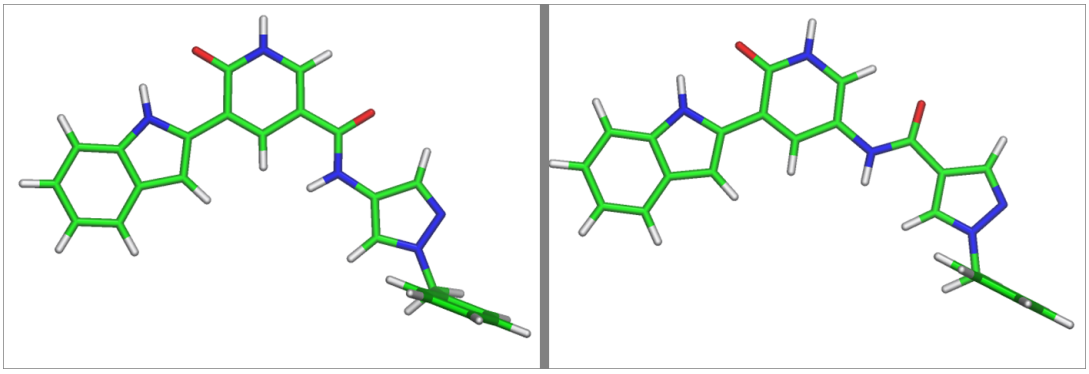


Figure 6.23: Comparison of the two Chk-1 ligands. **Left:** 5BT.pdb. **Right:** 5CH.pdb

Based upon a visual inspection of the protein-ligand complexes the change in affinity is unexpected, since the protein-ligand interactions in both cases are extremely similar, as highlighted in Figure **??**. In order to gain some idea into the reason for the difference in binding affinity, dual topology simulations [145]

were performed between the two ligands, using both crystal structures as reference structures.

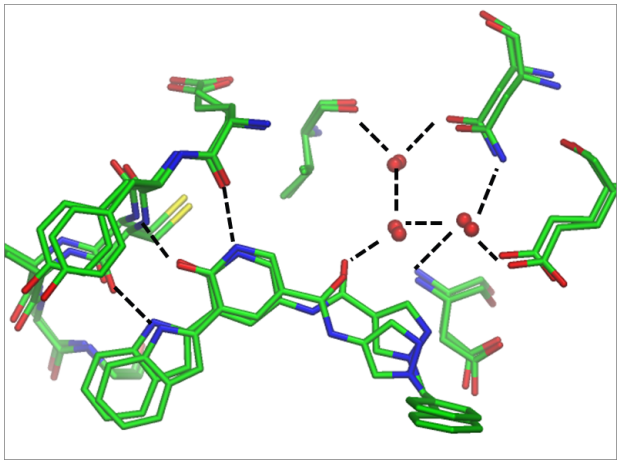


Figure 6.24: Crystal structure overlay of the two complexes. Critical residues and hydrogen bonds are identified by the slashed black lines, with the position of crystallographic waters shown in red

Dual Topology Simulations

Two different sets of RETI simulations were performed upon the bound state.[28, 29] The first set went from the 5CH crystal structure to the 5BT ligand, with the second going from the 5BT crystal structure to the 5CH ligand. Equally, transformations going from the 5BT-5CH and 5CH-5BT ligands were performed in bulk solvent. For the bound simulations, 16 equally spaced λ windows were used with data collected in 600 blocks of 100 K moves following 5 M MC moves of prior equilibration. In the bound leg, solvent moves were attempted with a probability of 85.7 %, protein side-chain and backbone moves with a probability of 12.9 % and solute moves with a probability of 1.4 %. The bond angles and torsions for the side chains and backbone of residues within 10 Å of any heavy atom of the ligands were sampled.

The same number of moves and λ windows were used for the free simulations. In the free leg, solvent moves were attempted with a probability of 98.7 % and solute moves with a probability of 1.3 %. A coulombic softening parameter of 2 and a Lennard-Jones softening parameter of 1.5 were used for the bound transformations, with a coulombic softening parameter of 1 and a Lennard-Jones softening parameter of 1.5 used for the free transformations. The form of the soft-core equations can be seen in equations **??** and **??**.

$$U_{LJ} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{(\lambda\delta\sigma_{ij} + r_{ij}^2)^6}\right) - \left(\frac{\sigma_{ij}^{6}}{(\lambda\delta\sigma_{ij} + r_{ij}^2)^3}\right)\right] \tag{6.1}$$

$$U_{coul} = \frac{(1-\lambda)^n q_i q_j}{4\pi\epsilon_0\sqrt{(\lambda + r_{ij}^2)}} \tag{6.2}$$

In equations **??** and **??**, δ is the Lennard-Jones scaling parameter whilst n is the coulombic scaling factor. q_i and q_j are the atomic partial charges on atoms *i* and *j*. ε_{ij} and σ_{ij} are the Lennard-Jones well depth and collision diameter for atom pair *i* and *j*, with r_{ij} the inter-atomic distance.

Table **??** shows the calculated free energy changes for the transformations.

	5CH-5BT ΔG (kcal/mol)	5BT-5CH ΔG (kcal/mol)
Bound	1.81 $\pm$ 0.48	-1.71 $\pm$ 0.55
Free	-0.24 $\pm$ 0.17	0.23 $\pm$ 0.18
$\Delta\Delta$G	2.04 $\pm$ 0.65	-1.94 $\pm$ 0.73

Table 6.7: Relative binding free energy values between the 5CH and 5BT crystal structures. Standard errors across at least 6 independent simulations are reported

Table **??** shows that the bound and free legs for the two different transformations are the opposite of each perturbation, suggesting that the protein structure is not a factor in calculating the relative binding free energy of the inhibitors. It is encouraging to note that the free energy change of approximately ($\pm$) 2 kcal/mol

is in good agreement with the experimental change of (±) 1.16 kcal/mol. In order to understand the reason for the change in free energy, an enthalpic breakdown of the individual energy terms in the bound legs was performed for each perturbation.

5BT	COU	LJ	SUM	**5CH**	COU	LJ	SUM	% Diff
Sol-Pro	-40	-49	-89(2)	Sol-Pro	-39	-52	-91(1)	1.3
Svn-Pro	-3020	-117	-3137(42)	Svn-Pro	-3036	-122	-3158(45)	0.6
Svn-Sol	-14	-8	-23(2)	Svn-Sol	-17	-9	-27(2)	**14**

Table 6.8: Average energy contributions in the 5BT-5CH transformation. **Key:** Sol-Pro = Solute-Protein energy, Svn-Pro = Solvent-Protein energy, Svn-Sol = Solvent-Solute energy. All values in kcal/mol, with associated errors, found as the block average of the first and last 30 M MC moves, shown in parenthesis.

5CH	COU	LJ	SUM	**5BT**	COU	LJ	SUM	% Diff
Sol-Pro	-38	-52	-89(1)	Sol-Pro	-40	-51	-92(2)	1.3
Svn-Pro	-3125	-92	-3216(97)	Svn-Pro	-3021	-86	-3107(45)	0.6
Svn-Sol	-21	-10	-30(1)	Svn-Sol	-16	-8	-24(1)	**-20.9**

Table 6.9: Average energy contributions in the 5CH-5BT transformation. **Key:** Sol-Pro = Solute-Protein energy, Svn-Pro = Solvent-Protein energy, Svn-Sol = Solvent-Solute energy. All values in kcal/mol, with associated errors, found as the block average of the first and last 30 M MC moves, shown in parenthesis.

Tables **??** and **??** show that the 5CH and 5BT endpoints display extremely similar values to within error, highlighting that the choice of reference crystal structure is not a factor in the change in activity. From comparing the λ_0 and λ_1 values from within the same simulation it can be seen that the solute-protein and solvent-protein energies are similar to within error, suggesting that there is not a change in these energetic contributions as the ligand is changed. This is demonstrated by the low % difference between the two end points. In comparison, a large % difference is seen between the different λ states for the solvent-solute terms, with the 5CH endpoints showing a larger energetic contribution to the system than the 5BT endpoints. This data suggests that, upon a change in the ligand,

the interaction energy between the solute and solvent is affected. This offers a possible explanation for the observed free energy change in the dual topology simulations, although other or additional effects cannot be discounted.

JAWS simulations

Based upon prior discussions with Vernalis, the mostly likely explanation for the apparent change in the solute-solvent interaction energy lied in a cluster of three crystallographic water molecules in the top of the pocket. These waters are in close proximity to the V68 gatekeeper residue and form a network back to the carbonyl amide oxygen of the ligand. Figure **??** shows the position of these waters in the presence of 5BT.

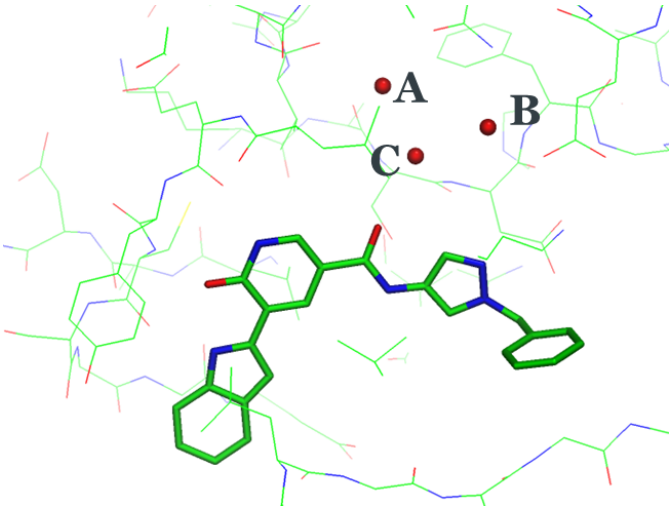


Figure 6.25: Waters A, B and C in the 5BT protein-ligand complex

In order to determine whether the binding free energy of these water molecules is affected by the change in the amide connectivity in the ligand, JAWS stage 2 simulations were performed upon the three waters in turn, using the coordinates previously extracted from the dual topology simulations. The same simulation

protocol as detailed for N9 neuraminidase as used. The results for this can be seen in tables **??** and **??**.

	λ_0 ΔG_{bind}	λ_1 ΔG_{bind}
Wat A	-5.24	-10.96
Wat B	UNCALC	-10.70
Wat C	-6.93	-7.41

Table 6.10: JAWS stage 2 binding free energies in the 5BT-5CH transformation. All values in kcal/mol, using a JAWS biasing potential 10 kcal/mol. UNCALC signifies waters for which free energy estimates cannot be obtained at a biasing potential of 10 kcal/mol due to insufficient $\theta < 0.05$ transitions. The error for these binding free energies approximately is $\pm$ 0.80

	λ_0 ΔG_{bind}	λ_1 ΔG_{bind}
Wat A	-11.29	-5.41
Wat B	UNCALC	-12.43
Wat C	-6.70	-7.26

Table 6.11: JAWS stage 2 binding free energies in the 5CH-5BT transformation. All values in kcal/mol, using a JAWS biasing potential 10 kcal/mol. UNCALC signifies waters for which free energy estimates cannot be obtained at a biasing potential of 10 kcal/mol due to insufficient $\theta < 0.05$ transitions. The error for these binding free energies is approximately $\pm$ 0.80

From looking at tables **??** and **??**, it can be seen that in both sets of simulations Wat A is stabilised when in the presence of ligand 5CH compared to ligand 5BT. This is an unexpected result, since Wat A is the water molecule which is most distal from the ligand modification. As such this could suggest that the change in the ligand connectivity directly affects the network of the three waters, causing a change in the binding affinity of the ligand. In order to establish whether the network is destabilised by a change in the ligand connectivity, GCMC simulations were performed using the simulated annealing method on the entire network region. This method allows the number of water molecules to be tracked as a func-

tion of the applied chemical potential, and is able to help identify changes in cavity occupancy, and hence network stability within the cavity. As with the JAWS stage two simulations, initial coordinates were obtained from the dual topology simulations.

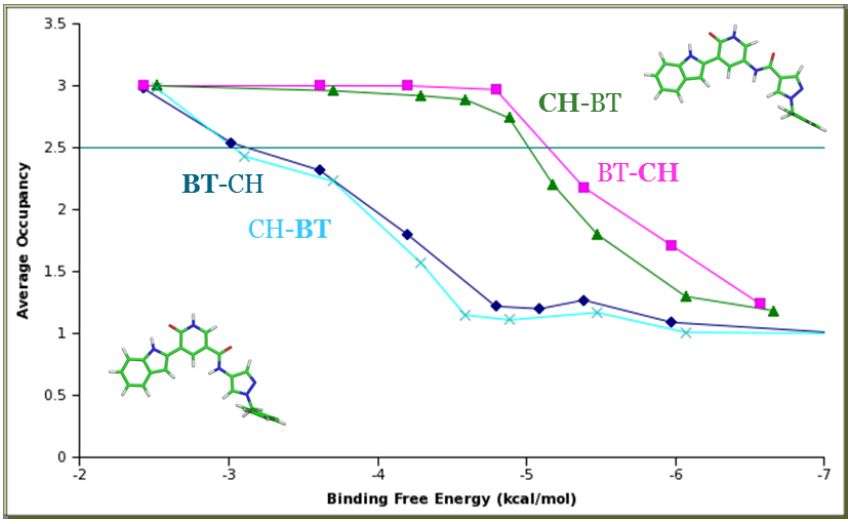


Figure 6.26: GCMC simulated annealing titration plots for the two different Chk-1 ligands, showing the relationship between the average occupancy and the binding free energy. The bold letters in the labels signifies the ligand studied, whilst the first letters in the labels signifies the native protein-ligand crystal structure used

Figure **??** shows that each ligand displays a similar profile in the presence of the two different crystal structures, again highlighting that protein conformation effects are unlikely to be the cause of the change in ligand binding affinity. As expected, and suggested by the JAWS stage 2 binding free energies, the networks which are in the presence of the 5CH ligand are stabilised by approximately 2.3 kcal/mol. Despite showing the expected trend, there are however a few discrepancies in the data. Assuming that the equivalence point at an occupancy of 2.5 is equal to the binding free energy of the weakest bound water, it can be seen that Wat A drops out of the system approximately 2 kcal/mol before the JAWS stage

2 binding free energy would suggest in the 5BT simulations (approximately -5.2 kcal/mol). A similar story is found for the 5CH simulations, where a water drops out of the system at a binding free energy of approximately 5.2 kcal/mol, with the JAWS stage 2 binding free energy suggesting that this should not be the case (approximately -11 kcal/mol).

Visual inspection of the 5CH GCMC simulations confirmed that the water which dropped out of the system first was Wat A, suggesting that there are effects occuring in the simulations which are not consistent between the JAWS stage 2 and GCMC methods. It was noted in the GCMC simulations that when Wat A drops out of the simulations, a nearby phenylalanine residue is able to migrate into the pocket and occupy the nearby space previously occupied by Wat A. As a result the probability of a successful insertion back into the cavity becomes minimal, and the two remaining waters reorganise to satisfy the remaining hydrogen-bonding oppotunities. Snapshots of this behaviour can be seen in Figure **??**, taking from a simulation with 5BT bound in the 5CH crystal structure.

The crucial difference between the JAWS stage 2 and GCMC methodologies in this system is the role of a hardwall potential around the area of interest. The isolated JAWS stage 2 simulations apply a hardwall which applies to every atom in the simulation, and prohibits any other atom from entering the region of interest. In comparison, the GCMC simulations allows protein atoms and residues to migrate according to the unconstrained potential energy function. As a result, in this instance the end points of the GCMC and JAWS simulations are not the same, qualitatively explaining the difference in the binding free energies between the methods.

In order to validate the GCMC binding free energies of Wat A, double-decoupling was used to estimate the binding free energy of Wat A in the presence of the two

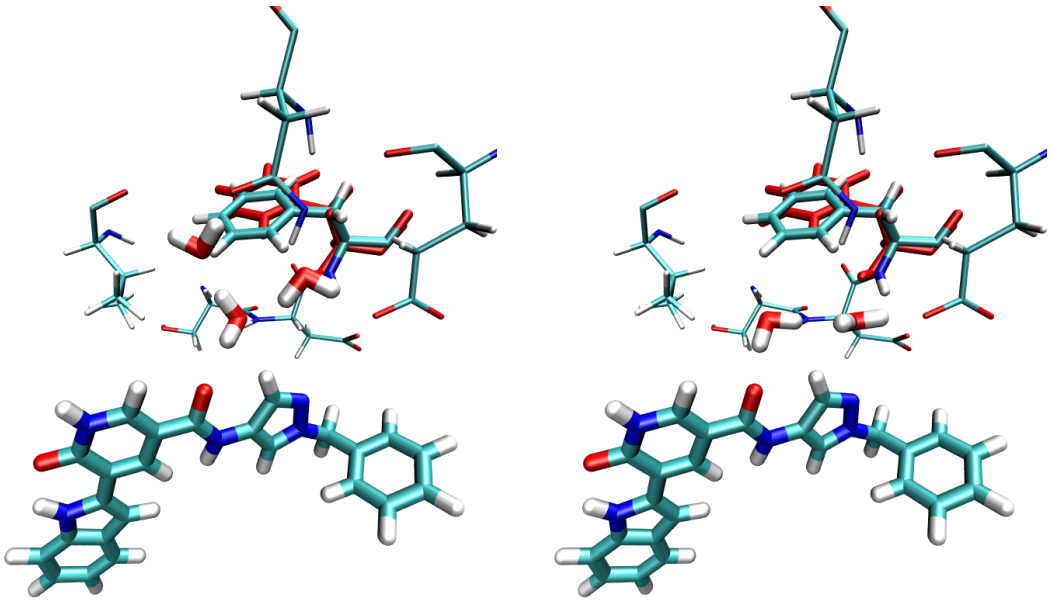


Figure 6.27: Original and modified positions of Waters A, B and C as a function of the relocation of the nearby phenylalanine residue. The original position of the residue is shown in red

ligands. Whereas a hardwall potential was previously used in the N9-neuraminidase, in this case a harmonic potential of 5 kcal/mol/Å^2 was applied to the water of interest. The correction term for this harmonic restraint, ΔG_{rest} is shown in equation **??** and is equal to -2.47 kcal/mol.

$$\Delta G_{rest} = \frac{3}{2} RTln\left(2\pi RT/k_{harm}\right) - RTlnV^{o} \tag{6.3}$$

In equation **??**, R is the gas constant, T is the temperature of the simulation, k_{harm} is the force constant of the harmonic restraint and V^0 is the standard state volume. This correction term is then used to calculate the binding free energy of the water molecule using equations **??** and **??**.

In order to allow the phenylalanine residue to occupy the space left by the water during the LJ decoupling stage a LJ soft-core parameter of 1 was used.

The result of these decouplings for the 5CH-5BT and 5BT-5CH simulations can be seen in Figures **??** and **??** respectively, and show that the harmonic double decoupling simulations are now in excellent agreement with the GCMC titration curves.

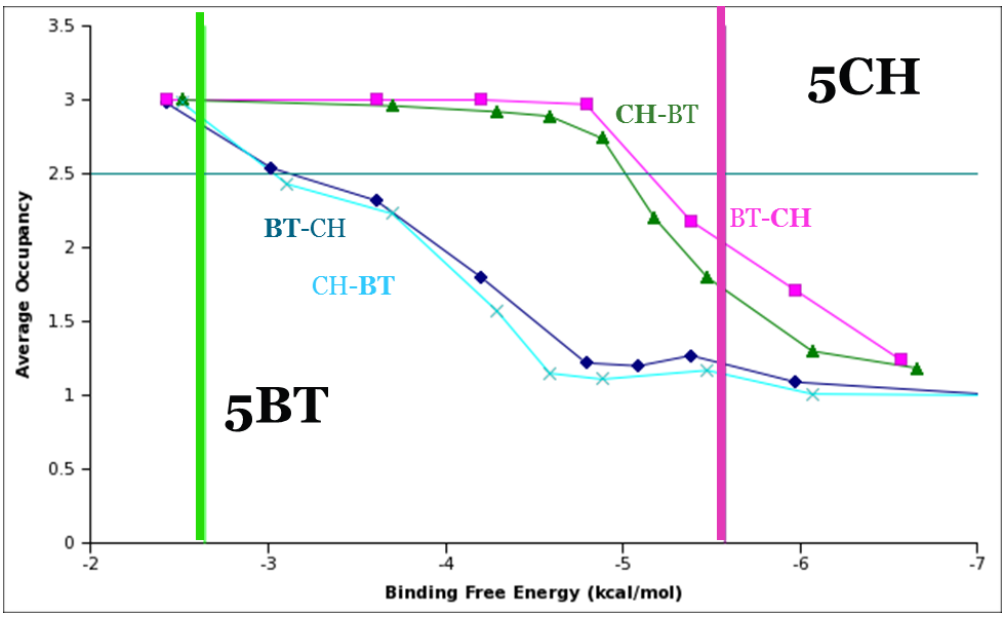


Figure 6.28: GCMC simulated annealing titration plots for the two different Chk-1 ligands, alongside the double decoupling binding free energy calculation, shown as vertical lines, for Wat A based on the 5CH crystal structure. The associated standard error for the double decoupling result is $\pm$ 0.40 kcal/mol

In order to prove that the most likely explanation for the change in affinity is the role of the water network, Waters A and C, the two weakest binders in the system, were removed from the simulation setup and the dual topology ligand perturbation simulations rerun. When this was performed the $\Delta\Delta G_{bind}$ free energy leg for the CH-BT transformation was -0.332 $\pm$ 0.18 kcal/mol and the $\Delta\Delta G_{bound}$ free energy leg for the BT-CH transformation was 0.244 $\pm$ 0.41 kcal/mol, suggesting that there is a neglible change in the relative binding free energy between the inhibitors when the two waters are removed.

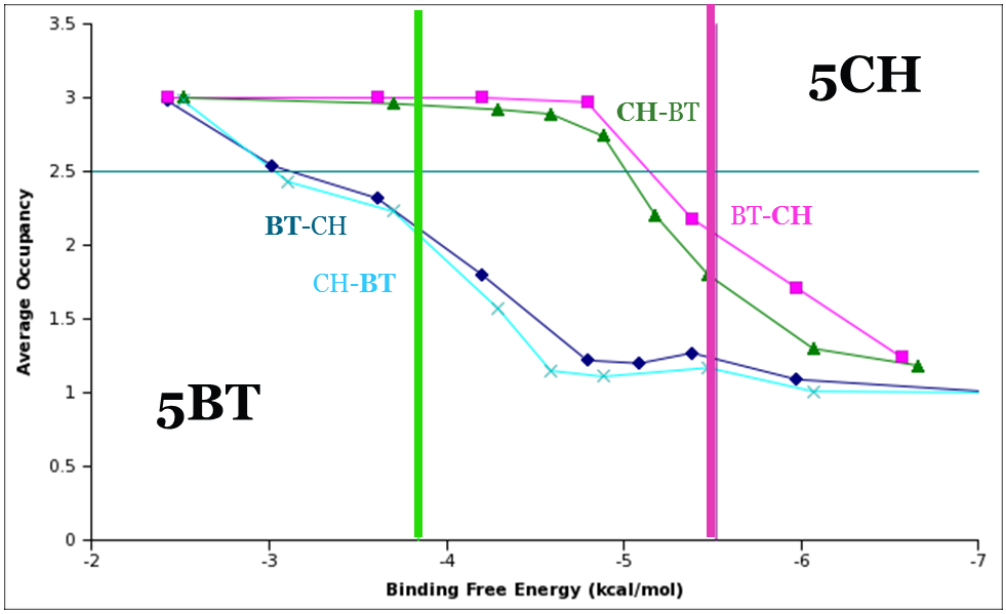


Figure 6.29: GCMC simulated annealing titration plots for the two different Chk-1 ligands, alongside the double decoupling binding free energy calculation, shown as vertical lines, for Wat A based on the 5BT crystal structure. The associated standard error for the double decoupling result is $\pm$ 0.40 kcal/mol

In order to prove that the free energies which have been calculated are correct, a free energy cycle was constructed. The change in binding free energy between the inhibitors should, in principle, be equal to the change in the binding free energies of Waters A and C in the presence of the two different ligands. Starting from the decoupled state of Wat A, with the phenylalanine residue in the vicinity of the pocket, the binding free energy of Wat C in the absence of Wat A was found, again using a soft-core LJ parameter of 1 and a harmonic restraint of 5 kcal/mol based on the oxygen atom of interest. The calculated free energies for Waters A and C using the 5CH and 5BT crystal structures are shown below in tables **??** and **??**, with the free energy cycles shown in Figures **??** and **??**.

Figures **??** and **??** clearly demonstrates that the free energy cycles close satisfactorily to 0.25 and 0.26 kcal/mol respectively, highlighting that the most proba-

Lam0 - 5CH	Elec	LJ	ΔG_{bind}	**Lam1 - 5BT**	Elec	LJ	ΔG_{bind}
Wat A	14.006	0.017	-5.753	Wat A	9.945	1.190	-3.065
Wat C	9.360	1.306	-2.396	Wat C	9.997	1.559	-3.286

Table 6.12: Binding free energies (in kcal/mol) for waters A and C using double-decoupling with a 5 kcal/mol harmonic restraint using the 5CH crystal structure. The table headers denote the free energy for decoupling the electrostatic terms, Elec, the free energy for decoupling the Lennard-Jones terms, LJ, and the corrected binding free energy, ΔG_{bind}

Lam0 - 5BT	Elec	LJ	ΔG_{bind}	**Lam1 - 5CH**	Elec	LJ	ΔG_{bind}
Wat A	12.458	-0.349	-3.839	Wat A	14.529	0.710	-5.519
Wat C	10.356	0.251	-2.337	Wat C	11.131	-0.518	-2.343

Table 6.13: Binding free energies (in kcal/mol) for waters A and C using double-decoupling with a 5 kcal/mol harmonic restraint using the 5BT crystal structure. The table headers denote the free energy for decoupling the electrostatic terms, Elec, the free energy for decoupling the Lennard-Jones terms, LJ, and the corrected binding free energy, ΔG_{bind}

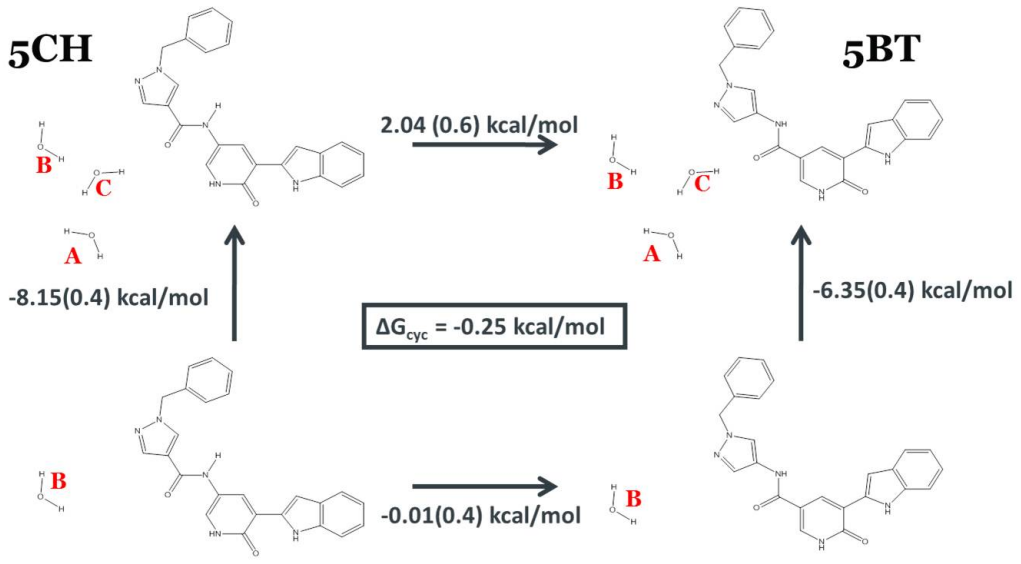


Figure 6.30: Relative binding free energies for the 5CH-5BT transformation in the presence or absence of Waters A and C. Standard errors are shown in parenthesis

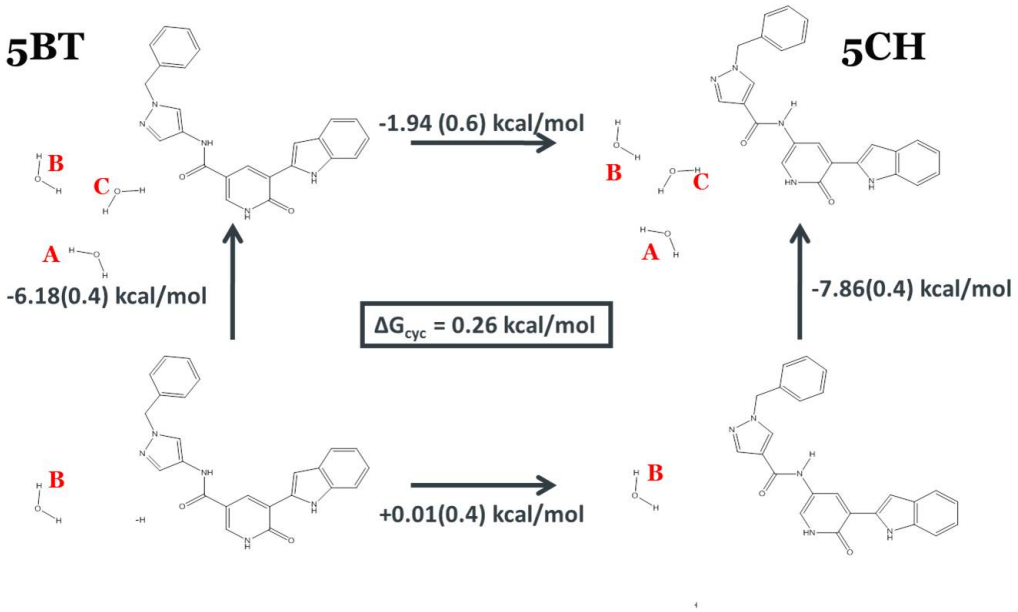
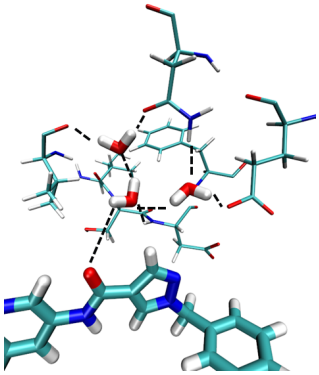


Figure 6.31: Relative binding free energies for the 5BT-5CH transformation in the presence or absence of Waters A and C. Standard errors are shown in parenthesis

ble explanation for the change in ligand affinity is the role of the water network. Crucially the results clearly demonstrate that a change in the ligand scaffold can affect a *distal* water molecule in a network, which in turn affects the binding affinity of the ligand. As such, it provides evidence to the key role which water molecules play in the stabilisation of protein-ligand complexes. Examination of the individual water molecules in the network identifies that the most distal water is affected by the change in ligand scaffold, but the JAWS stage two energies seemingly overestimate the binding affinity of the water molecules, similar to the BPTI cavity in section **??**. This is due to the presence of the hardwall in the JAWS stage 2 binding free energy calculations, which prohibits the pocket from reorganising. As in the previous examples, this highlights that the binding free energy of water molecules in a network are not necessarily additive, and the network should be treated as a whole rather than the individual elements. This, once

again, suggests that GCMC simulations should be used for water networks, since they are able to observe changes and reorganisation in the pocket according to the unconstrained potential energy function.

One question which has not yet been addressed is why Wat A is the water which is destabilised. Could this be identified purely from a single simulation, rather than relying on a rigorous free energy approach? In order to address this question a hydrogen bond analysis was performed for the three waters in the network, using the end points from the 5CH-5BT dual topology simulations. Figure **??** shows the H-bond positions for the three waters in the presence of the 5CH ligand and demonstrates that Wat A only forms 3/4 H-bonds, whilst the other two waters form 4. It was noted that the same configuration is found for the 5BT ligand. Purely based on this count it would be expected that Wat A would be the weakest link in the network, but it does not give indication for why the 5BT ligand destabilises the network.

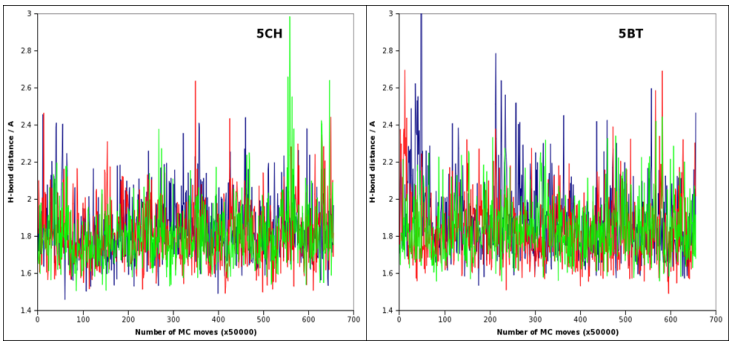


Water	Number of H-bonds
A	3
B	4
C	4

Figure 6.32: Hydrogen bond analysis for the three waters in the presence of the 5CH ligand. These hydrogen bonds were consistently found across the entire simulation

A H-bond distance analysis was performed on the Wat A H-bonds to observe whether there is a considerable difference in the bond lengths between the two ligands. Figure **??** shows that the three H-bond lengths formed by Wat A are not

considerably different between the two ligands, although there is a slight elongation in the presence of the 5BT ligand. Although this could indicate that the interactions formed between Wat A and its local environment are slightly weaker, it does not indicate the large change in affinity which is observed both experimentally and through these simulations. Indeed, the two sets of hydorgen bond lengths are within error of each other. This analysis highlights that the effect of changing the ligand scaffold is extremely subtle, and how rigorous simulations can help to rationalise SAR which are not immediantely obvious to the medicinal chemist.



Bond	5CH /Å	5BT /Å
Blue	1.84 (0.16)	1.90 (0.21)
Green	1.82 (0.16)	1.84 (0.18)
Red	1.81 (0.16	1.84 (0.15)

Figure 6.33: Average H-bond distances for the three Wat A contacts in the presence of 5CH (**left**) and 5BT (**right**). The block average over the first 300 and last 300 frames is shown in parenthesis

6.9 Conclusions

In this chapter the application of JAWS simulations and GCMC to novel problems was discussed. In the T4-lysozyme system, the combination of JAWS stage 2 and GCMC was able to corroborate previous studies and prove that the cavity is empty under ambient conditions. Previous experimental and computational studies have

shown that, upon the application of high pressures, the cavity is capable of hosting 4 water molecules.[59] Since JAWS simulations and GCMC are based upon the canonical and Grand Canonical ensemble respectively, they cannot directly simulate changes in pressure. By changing the JAWS stage 2 reference state, however, the effect can be approximated. The results showed that the complex of 4 water molecules becomes more stable at higher pressures, however, in order to fully understand such a change in the system, a more complex methodology should be employed.[59, 126]

The study upon IL1β corroborated previous work performed by Yin in proving that the cavity is empty under standard conditions. Mutation studies performed by Adamek [129] had previously highlighted destabilisation effects in the cavity; something which could be associated with the presence of water in the cavities. The utilisation of JAWS stage 2 and GCMC appears to refute this study, although it is possible that subtle polarisation effects which are not captured by standard forcefields could change the conclusions.

A study on two kinase structures was then performed; CDK2 and Pim-1. These kinases have a different protein sequence around the hinge region and were expected to have a different hydration pattern to each other. The use of GCMC simulations at different chemical potentials was able to elucidate the hydration patterns and locate the strongly bound water molecules for each system. By calculating the binding free energies of the water molecules in CDK2 using JAWS stage 2, it was found that the hydration free energies of waters are not necessarily additive. This has important consequences for drug design - by targetted a loosely bound water molecule the surrounding water network can collapse, significantly reducing the desolvation cost of the pocket. This allows a large contribution to the binding free energy of the ligand through entropic destabilisation of the waters

and strong enthalpic protein-ligand gains. The idea of Δ as a measure of water stability in a water was also introduced; waters which have low values of Δ are likely to be stable in a network and are candidates for water retention, whilst waters with high Δ values are part of weak networks and are likely to be displaceable by a ligand.

Finally, the lessons learnt from the previous studies were used upon the Chk-1 system. Through the use of dual topology simulations and JAWS stage 2 free energy estimates it was found that, upon a change in the ligand structure, a distal water in a network is destabilised. This destabilisation was identified as a likely explanation for the change in ligand affinity, although the magnitude of the difference in the binding affinity of the waters did not match that of the ligands using this methodology. Through treating the network as a whole it was found that the difference in the network stability was consistent with the dual topology simulations, highlighting again that the binding free energies of water molecules in networks are not additive and should be treated as a whole using GCMC. In order to prove definitively that the water network is responsible for the change in affinity, the two weak waters were removed from the dual topology simulations and the simulation repeated. This yielded a relative free energy change of approximately 0 kcal/mol between the two ligands, highlighting that the water molecules are responsible for the change in ligand affinity. Closed free energy cycles were obtained for both crystal structures by including the binding free energues of the displaced waters, demonstrating the internal consistency in the methodologies.

Chapter 7

Application of JAWS to Fragment-Based Drug Design

7.1 Introduction

The following chapter describes the application of the JAWS algorithm to FBDD. Since the existing methodologies available to FBDD have several deficiencies, the rationale for choosing the JAWS methodology over other methods is initially described. Two different test cases are then described; the anti-cancer target Kinesin Spindle Protein and the coagulant factor Xa. The observed results are compared to existing assay and crystallographic data, highlighting the ability of the JAWS method to predict key structural motifs and highlight oppotunities for lead design. The deficiencies of the method are then described, highlighting the need for extensive testing against real-life assay data which is currently not publically available.

7.2 Choice of method

As described previously, existing computational tools for performing FBDD are all deficient in a few aspects. Amongst other things, the methods typically do not allow for solvent competition, protein flexibility, and are incapable of accurately ranking fragments based upon their free energies. There are six major criteria which need to be addressed for a computational approach to be effective:

1. The method should locate and rank fragments based on their binding free energy;
2. The method should use accurate energy functions, based on a simulation approach;
3. The method must allow competition between different fragments, and critically, water;
4. The method should include an estimate for the desolvation of fragments;
5. The method should allow protein and fragment flexibility;
6. The method should be reliable and efficient.

One possible approach to creating such a method would be to utilise the GCMC methodology. Currently the method does not allow explicit solvation, instead using a post-processing stage where an implicit solvent is used to correct for the binding free energy. An approach can be envisaged where water molecules are allowed to compete against fragments in the same binding site, hence allowing explicit water solvation. However, even with the use of schemes such as cavity-bias [51] and configurational-bias [53], the acceptance rates for such a system setup are likely to be extremely low. Running a GCMC simulation with flexible

sidechains has already been performed for water-based systems, so this is unlikely to be a drawback. However, one other drawback of the method is that there is no easy way to account for fragment desolvation in the process.

As demonstrated in the previous two chapters, the JAWS methodology performs similarly to GCMC in terms of its abilities to both rank molecules according to their free energies, and also to predict the location of molecules in systems. Like GCMC it is capable of allowing protein flexibility, yet the acceptance rates are far higher since the molecules can have θ values which can vary between 0 and 1. In addition, the method should be able to simulate competition between different fragments, as well as including a desolvation term through an modified potential energy term as a function of θ. As such, the JAWS algorithm was chosen as the preferred method for modelling fragments.

7.3 Development of the JAWS algorithm

Whilst employing the JAWS algorithm for water molecules is relatively straightforward and well defined, the application of the algorithm to fragment molecules warrants consideration. Three major issues have been identified in the application of JAWS to fragments:

1. Adequate sampling of chemical space;

2. Incorporation of a desolvation penalty to fragments;

3. Dealing with different standard states.

Approaches to dealing with these issues are now discussed.

7.3.1 Adequate sampling of chemical space

In the original JAWS protocol there is only one type of JAWS solute, being water. Since each water molecule can be essentially thought to be identical to any other JAWS solute in the system we do not need to worry about whether 'Water A' or 'Water B' occupies a specific site in space. However, once different solutes are allowed to explore the binding site this now becomes an issue. If a larger solute than water, such as benzene, needs to pass by other solutes then the likelihood of this occuring is small, since the repulsive van der Waals interactions will prohibit the acceptance of the move. Even though a fragment might have a θ value approaching zero, it will still have a finite coulombic and Lennard-Jones contribution which will drastically affect the acceptance rate for translational moves. As a result, fragments could essentially be trapped in particular regions of chemical space, prohibiting them from exploring other regions of the site and other solutes from sampling that region.

In order to allow fragments to pass by each other and sample the chemical space more effectively, a soft-core potential was incorporated into the JAWS potential energy expression.[94, 146] By including soft-core parameters the coulombic and Lennard-Jones are softened, so that at low θ values the fragments will experience considerably less interaction with the surroundings than previously. This should improve the acceptance rate of fragments passing by each other, and hence improve the sampling of available chemical space. The soft-core equations used can be seen in equations **??** and **??**.

$$U_{LJ,\theta} = (\theta)4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{((1-\theta)\delta\sigma_{ij}+r_{ij}^2)^6}\right)-\left(\frac{\sigma_{ij}^{6}}{((1-\theta)\delta\sigma_{ij}+r_{ij}^2)^3}\right)\right] \quad (7.1)$$

$$U_{coul,\theta} = \frac{(\theta)^n q_i q_j}{4\pi\epsilon_0\sqrt{((1-\theta)+r_{ij}^2)}} \tag{7.2}$$

In equations **??** and **??**, δ is the Lennard-Jones scaling parameter whilst n is the coulombic scaling factor. These were chosen to match the parameters used to decouple the fragments from a box of water, discussed in the next subsection **??**. q_i and q_j are the atomic partial charges on atoms *i* and *j*. ε_{ij} and σ_{ij} are the Lennard-Jones well depth and collision diameter for atoms *i* and *j*, with r_{ij} the inter-atomic distance.

7.3.2 Incorporation of a desolvation penalty to fragments

It is important to account for solvent competition in fragment based drug discovery, since in the body drug molecules need to compete within an aqeuous environment. Whilst simulating fragments with water gives information about whether a fragment is capable of outcompeting water in the binding site, it does not account for whether the fragment would, in reality, actually be in the binding site. Although the interactions between the fragment and the protein might be favourable, the interactions which the fragment makes with the solvent bulk might be greater. In this instance the fragment would remain in the bulk instead of residing in the binding site.

In the standard JAWS algorithm the desolvation cost is accounted for in the second stage, whereby the biasing potential is typically set to the hydration free energy of water.[13] Performing such a simulation for fragments poses one major issue, namely what to surround the fragment solute in question with during the JAWS stage 2 simulation. During the JAWS stage 1 simulation, a mixture of fragments is simulated; each with a set of θ values. As described previously in the study of water molecules, it can be difficult to determine the accurate locations of

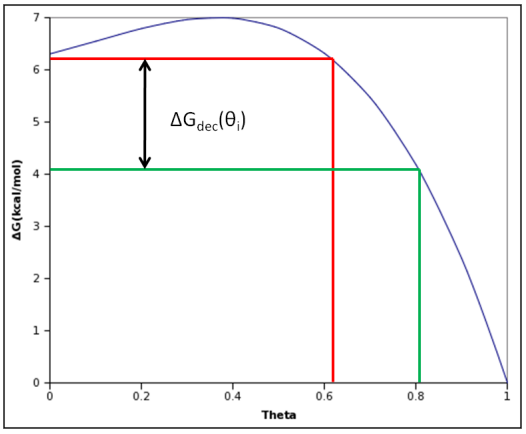
JAWS solutes if the resultant density map is convoluted and hard to interpret. Such a problem can be exacerbated for a mixture of fragments, and reliably scoring the position of a fragment in the second stage of the JAWS protocol is likely to be challenging if the correct environment cannot be defined. As a result, a different approach needs to be applied.

Rather than accounting for the desolvation cost in the second stage of the algorithm, the approach which has been utilised incorporates it into the first stage of the algorithm. Whenever a θ move is attempted an additional term is added onto the potential energy function, which takes into account the corresponding change in intermolcular potential in the bulk. In principle there are two different methods for performing such a procedure. The first is a Gibbs ensemble approach,[147] whereby each fragment in the binding site is simulated in parallel with the corresponding fragment in a water box. Whenever the fragment experiences a θ move in the binding site, the equivalent θ move is performed in the water box. The change in the interaction energies in each box are then totalled. This is then used in the Boltzmann expression to determine whether or not the move is accepted. Whilst this method is rigorous and should allow for an accurate estimate of the fragment desolvation cost, it is also computationally expensive. If a large number of fragments are chosen then each fragment needs to be linked to its own water box, causing a combinatorial expansion of the number of simulations being performed at one time. As a result, another method was sought.

The chosen method relies on generating a PMF profile for each fragment prior to the JAWS simulation, whereby the fragment is decoupled from a water box at a pressure of 1 bar in the NPT ensemble. A dual topology simulation is performed with the fragment gradually perturbed to a dummy atom, assisted by the use of soft-core parameters to prevent repulsive van der Waals' interactions as the frag-

ment is decoupled. 16 equally spaced λ windows were used with data collected in 600 blocks of 100 K moves following 5 M MC moves of prior equilibration. Solvent moves were attempted with a probability of 99 % ,solute moves with a probability of 0.9 % and volume moves with a probability of 0.9 %.

Once the PMF was obtained the profile was fitted to a polynomial expansion, allowing the value of the PMF to be calculated from any value of θ. The JAWS simulation is then performed, with the fragment PMF profile used to account for desolvation as part of the acceptance test. This can be seen in Figure **??**, where the sampled MC move attempts to change the value of θ from 0.6 to 0.8. The values of the PMF at θ = 0.6 and θ = 0.8 are extrapolated, shown in red and green respectively. The difference between these values is then calculated as $\Delta G_{des}\left(\theta_i\right)$. For this attempted move, attempting to increase the intermolecular interactions for θ_i in the simulation would require a loss of favourable interactions in the bulk, and thus the move is penalised by $\Delta G_{des}\left(\theta_i\right)$ in the Metropolis test.



$$\Delta G_{des} = \Delta G_o - \Delta G_n \quad (7.3)$$

Figure 7.1: Sample calculation for the desolvation correction term in the modified JAWS algorithm, for the θ move from 0.6 to 0.8

In equation **??**, $\Delta G_{des}\left(\theta_i\right)$ is the PMF correction term added onto the potential energy function, with ΔG_n and ΔG_o the new and old values of the PMF respectively. Whilst this approach is not as rigorous as the Gibbs ensemble approach it is

significantly faster and easier to implement. The soft-core parameters chosen for the dual topology simulations were varied to see which gave both the smoothest gradient profile and the closest match to the experimental result. In the proceding work four different fragment types were used; water, benzene, pyrazole and acetone. It was found that a coulombic scaling factor of 1 and a Lennard-Jones factor of 1.5 gave close agreement with the experimental result for these fragments.[148] These factors were used in equations **??** and **??**. The PMF plots along with the experimental values are shown in Figure **??**.

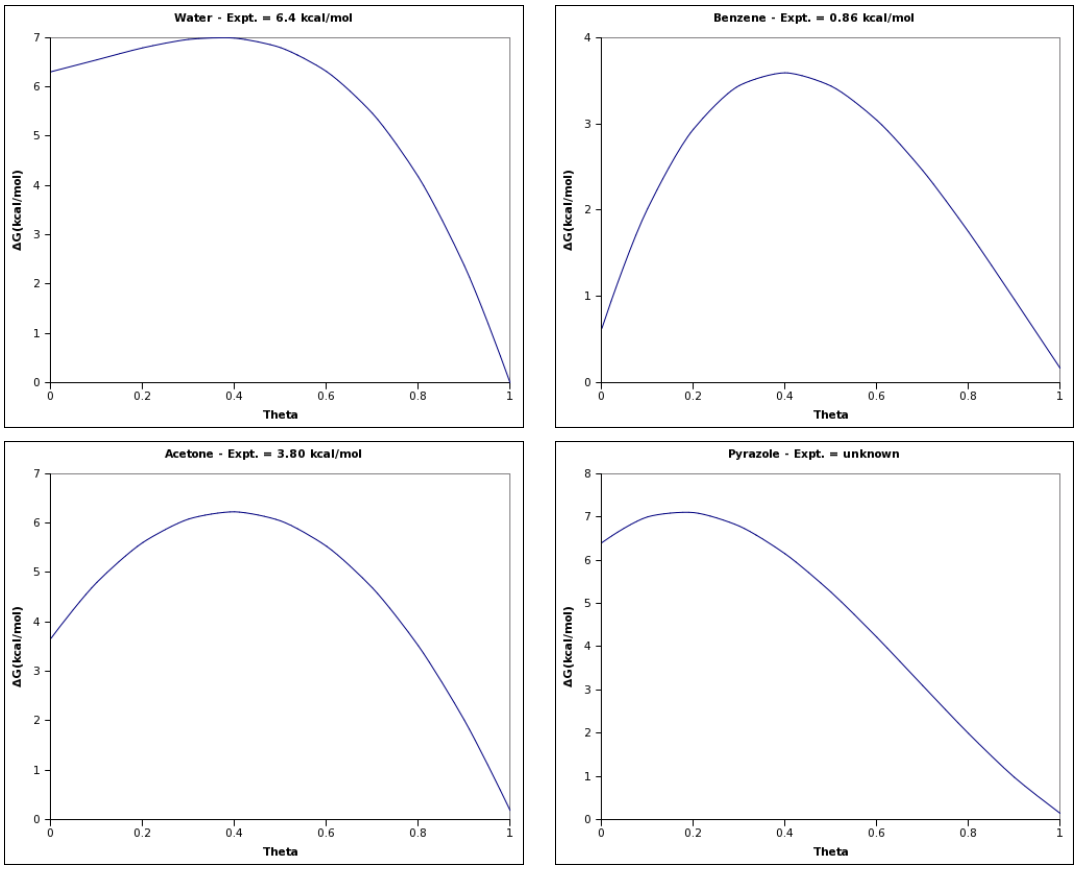


Figure 7.2: PMF profiles for the common fragments in JAWS fragment simulations. The expt value shown in the profiles correspond to the experimental free energy of hydration

The calculated free energies shown in Figure **??** display good agreement with

the experimental results. No data could be found for pyrazole, however the calculated free energy is not unreasonable considering the hydrogen bonding nature of the fragment.

7.3.3 Dealing with different standard states

One final feature which needs to be considered when simulating fragment-water mixtures is the difference in their standard states. Water, by definition, has a standard state concentration of 55 M. In comparison, experimental assays typically screen fragments in the 50 μM range, meaning that this needs to be accounted for in the simulation process. One possible way of simulating this is to have significantly more water molecules in the simulation, although this causes sampling difficulties due to the highly likelihood of condensed water phases forming within the simulation due to the number of waters in the system. Such phases, coupled with the numerical advantage of water molecules, will prevent the adequate sampling of fragments, even with soft-core potentials switched on. In order to account for the fact that water binding to the protein should be preferred for concentration (and hence entropic) reasons, whilst allowing adequate sampling of the fragment chemical space, a volume correction term has been developed. This volume correction term is added onto each fragment θ_i move, and ensures that favourable θ moves for water are preferred to the analogous fragment move. The form of this correction is shown in equation **??**.

$$\Delta G_{corr}(\theta_i) = -kT \ln \left(\frac{V^{sim}}{V^0} \right) \theta_i \tag{7.4}$$

In equation **??**, V^{sim} is the simulation volume and V^0 is the standard state of the fragment.[14] In line with existing fragment based approaches, the volume of

any fragment other than water was taken to be 1 M.

7.3.4 JAWS fragment Metropolis test

Given the corrections made to the JAWS methodology, the acceptance test used for θ moves is as follows:

$$acc\left(\theta_i \rightarrow \theta_j\right) = min\left[1, exp\,\frac{\left(-\Delta E + \Delta E_{des,i-j} + \Delta E_{corr}(\theta_{i-j})\right.}{k_B T}\right] \quad (7.5)$$

In equation **??**, ΔE represents the change in energy between states *j* and *i* as a function of θ. $\Delta E_{des,i-j}$ is the desolvation correction between states *j* and *i*, whilst $\Delta E_{corr}(\theta_{i-j})$ is the volume correction term applied between states *j* and *i*.

7.4 Kinesin Spindle Protein

7.4.1 Biological Relevance

The kinesin spindle protein (KSP) is a motor protein which is involved in mitosis. KSP slides apart microtubules, allowing for spindle assembly which is essential for chromosome seperation.[149] Until recently a common strategy in anti-cancer drugs was to target microtubules and, in particular, tubulin. Whilst this has proved a successful strategy, there is a high risk of side effects in targetting tubulin, in particular neurotoxic effects. The discovery of monasterol by Mayer et al. has helped to provide a new strategy in anti-cancer drugs.[150]

Monasterol has been found to be an inhibitor of KSP, preventing the motor protein from releasing ADP during mitosis and inducing apoptosis. The drug binds to an allosteric pocket of KSP, and has been found to cause distal changes

in the ATP binding site, allowing KSP to bind to ATP but preventing the release of ADP. One advantage of targetting KSP is that it is typically overly expressed in human cancers, meaning that targetting KSP is a much safer strategy than standard tubulin therapy. As a result several drugs are now in advanced clinical testing, one example being ispinesib, shown in Figure **??**.

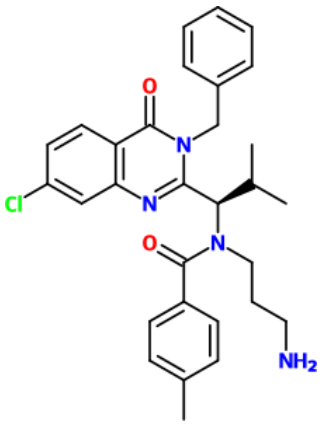


Figure 7.3: The KSP inhibitor, ispinesib

7.4.2 System Preparation

A 22 Å scoop was prepared using the 1YRS pdb, using the same protein protocol in the N9-neuraminidase system. Since the binding site of interest is allosteric, a larger scoop was prepared to include a Mg^{2+} ion and an ADP molecule. A 30 Å watercap was used, and a 594 $Å^3$ allosteric binding pocket defined. Unless stated otherwise, 8 water molecules were used for the simulations, with 4 replicas of each fragment used. Relatively few water molecules were used for the simulations; this was to prevent the aggregation of the water molecules to form a condensed phase during the simulation. It was found that these ratios gave satisfactory results, although no detailed sensitivity analysis into the ratios of fragments was conducted in this study.

7.4.3 Simulation protocol

The JAWS stage 1 simulation was performed upon the entire binding site, encompassing a region of 594 Å^3. The JAWS solutes were allowed to move freely around the grid region for one million moves whilst turned off. Unless stated otherwise, the θ threshold applied for solutes being classed as 'on' was 0.95. Statistics were then collected on the grid region for 40 million MC moves using a grid spacing of 1 Å, in line with the original JAWS study.[13] The resulting data was analysed using AstexViewer, and each fragment position normalised according to the number density of the most frequently observed fragment.[117] During the simulation, the JAWS solutes were allowed to move and sample θ, with sampling of bulk solvent performed. The bond angles and torsions for the side chains of residues within 12 Å of any heavy atom of the binding site were also sampled, with the protein backbone restrained throughout the simulation. For the JAWS stage 1 simulations, solvent moves were attempted with a probability of 23 % and protein side-chain moves with a probability of 3.5 %. Small variations in θ_i, chosen randomly between -0.15 and 0.15, were attempted with a probability of 50 %, in line with the original JAWS study [13], with translations and rotations of the JAWS waters attempted with a probability of 23.5 %.

Soft-core parameters were switched on for all of the JAWS solutes.

7.4.4 Simulations

JAWS simulations were performed upon the KSP binding pocket as a test of the JAWS methodology. Four different fragments were simultaneously simulated; acetone, benzene, pyrazole and water. Benzene and pyrazole were used as examples of an apolar and polar aromatic fragment respectively, whilst acetone was

chosen as an example of a hydrogen bonding acceptor to compete against water. 5 independent simulations were performed upon the 594 Å^3 binding pocket, with the collated results shown in Figure **??**.

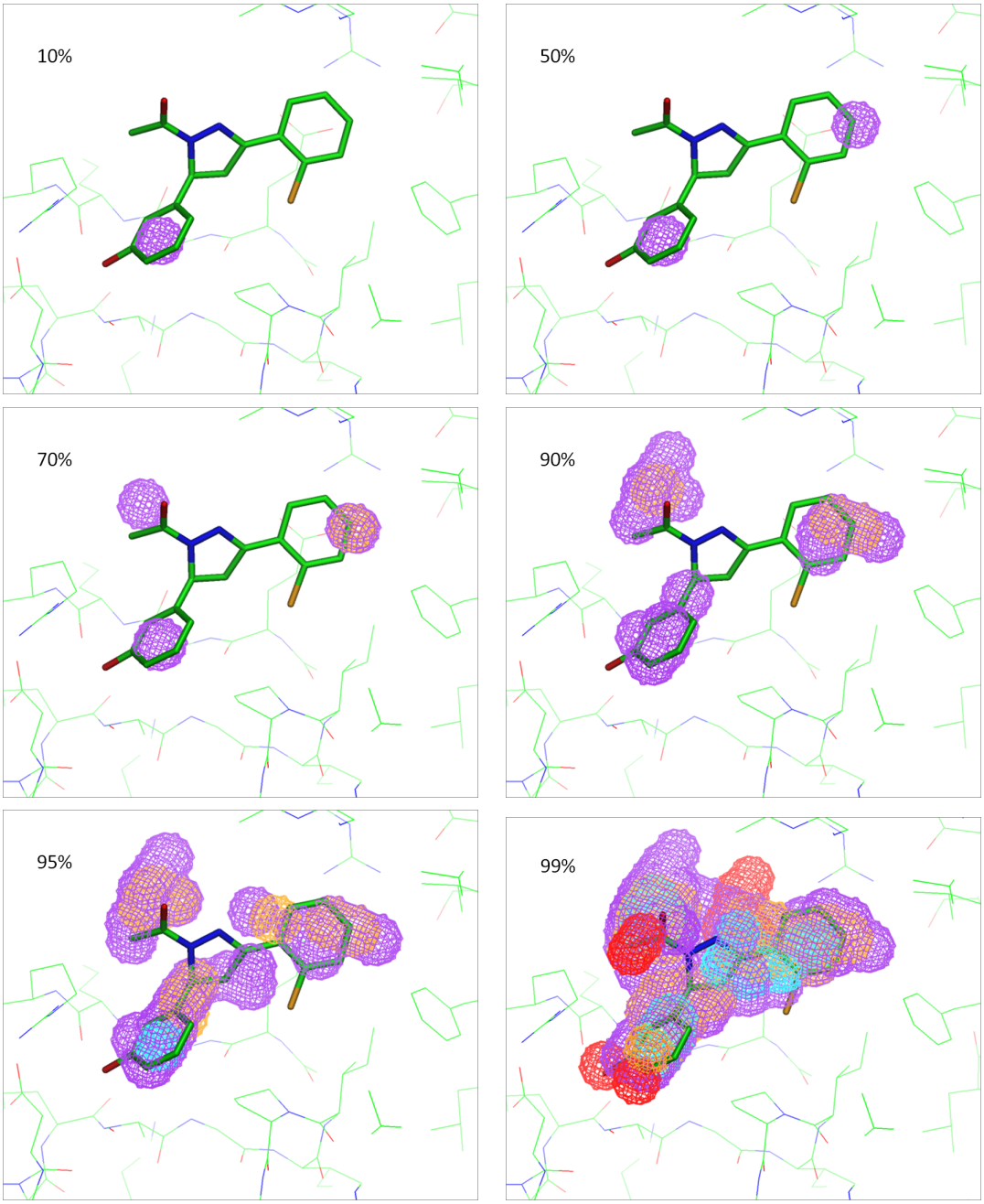
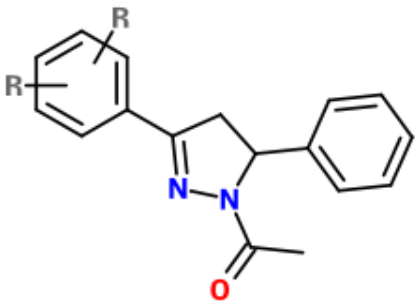


Figure 7.4: JAWS clustering density for the KSP system. The top data percentage for each image is shown in the top left. Key: pyrazole, benzene, water, acetone

Figure **??** shows that the major binding fragment in this system is pyrazole. In particular, three major pyrazole binding regions are observed; two sit on the aromatic rings of the native crystal ligand whilst another is found mimicking the acetyl group of the ligand. These predictions can be rationalised by examining the experimental assay data for the system.[151] An example of this is shown in table **??**, where modifications to the eastern ring are made.



R1	R2	Activity (nM)
2-F	H	3600
2-Br	H	23,300
2-Me	H	46,200
2-F	5-F	94

Figure 7.5: Structure-activity relationships for modifications made on the native ligand

Table **??** clearly shows that there is a strong preference for polar substituents on the right side of the ligand, and hence it is not particularly surprising that pyrazole displays a strong preference to benzene in this region. Whilst there is no particular evidence for why pyrazole outcompetes benzene in the left ring, the JAWS results suggests that there is a large preference for the fragment. One interesting position which is identified in the simulation is the pyrazole ring which mimics the acetyl ring. Experimental assays have shown that exchanging the methyl substituent of the carbonyl to a NH-cyclopropane group increases the affinity of the inhibitor from 4000 nM to 2.0 nM,[152] clearly demonstrating that the incorporation of a slightly more bulky and hydrogen-bonding donor group in this location is hugely advantageous. The fact that JAWS identifies this position is clearly highly

encouraging and helps to validate the methodology.

The dominance for pyrazole is clearly evident since the top 70 % of all data needs to be displayed before any other fragment types are observed. It is interesting to note that water is not observed in the system until the top 99 % of data is visualised. This suggests two things:

- Other fragments outcompete water in the binding pocket, and
- Any water which is observed in the system is likely to have a low binding free energy.

Acetone is observed in right side of the ligand, where the carbonyl group is typically found pointing along the same plane as the halogen on the pyrazole ring. As more and more data is incorporated in to the plots, it can be seen that pyrazole is also found in the linking position between the two aromatic rings. When a high proportion of the data is visualised, it can be seen that a water molecule is found in close proximity to the crystallographic hydroxyl group of the ligand. Experimental assays show that incorporation of the hydroxyl group in this position increases the activity of the ligand by an order of magnitude,[151] hence it is perhaps surprising that this is not observed more regularly. The most likely reason for this is the nearby presence of pyrazole, which itself is capable of making hydrogen bonds whilst also fulfilling the aromaticity of the pocket. Benzene is not found until the top 99 % of data is included, clearly demonstrating that it is outcompeted by the other fragments in this system. However, once it is located, the contours lie on top of the aromatic ligand rings.

Whilst the JAWS output clearly gives results which are in qualitative agreement with the experimental assay data and crystal structures, forming quantitative conclusions is considerably more difficult. Although the fragment populations

are assumed to be indicative of their binding free energy, it would be desirable to acquire more evidence to support this assumption. As previously described in section **??**, JAWS stage 2 simulations cannot be performed on the system since it is impossible to know what to flood the rest of the system with during the binding free energy estimation calculation. Another problem with the JAWS output is the lack of information regarding cooperativity between fragment contours. The lack of correlation occurs since in order to obtain enough data to create the fragment maps several independent different simulations are required. For example Figure **??** shows that an acetone contour is found extremely close to pyrazole, when the existence of both at the same time is an impossibility. When different simulations were analysed independently, it was observed that the position of different fragments can vary significantly, likely to be due to insufficient sampling within the pocket. In order to test the reproducibility of the results, longer simulations need to be performed to assess the convergence of the predictions.

The lack of true correlation between fragment is clearly a drawback to the technique, although there are possible ways around it. One such method would be utilising the assumption that the fragment which appears the most is the one with the most favourable free energy. This assumption has been proved to be the case for water molecules in N9 neuraminidase, demonstrated in Figure **??**.

Figure **??** shows that there is a strong correlation between the JAWS stage 1 normalised density and the calculated JAWS stage 2 binding free energies. This suggests that using the number density as a proxy for the ranking of binding free energies within the same system is a fair assumption in the case of water molecules, and is assumed to hold true for fragment molecules. It should be noted, however, that weakly bound water molecules typically have extremely low JAWS stage one densities, and attempting to estimate the binding free energy for these

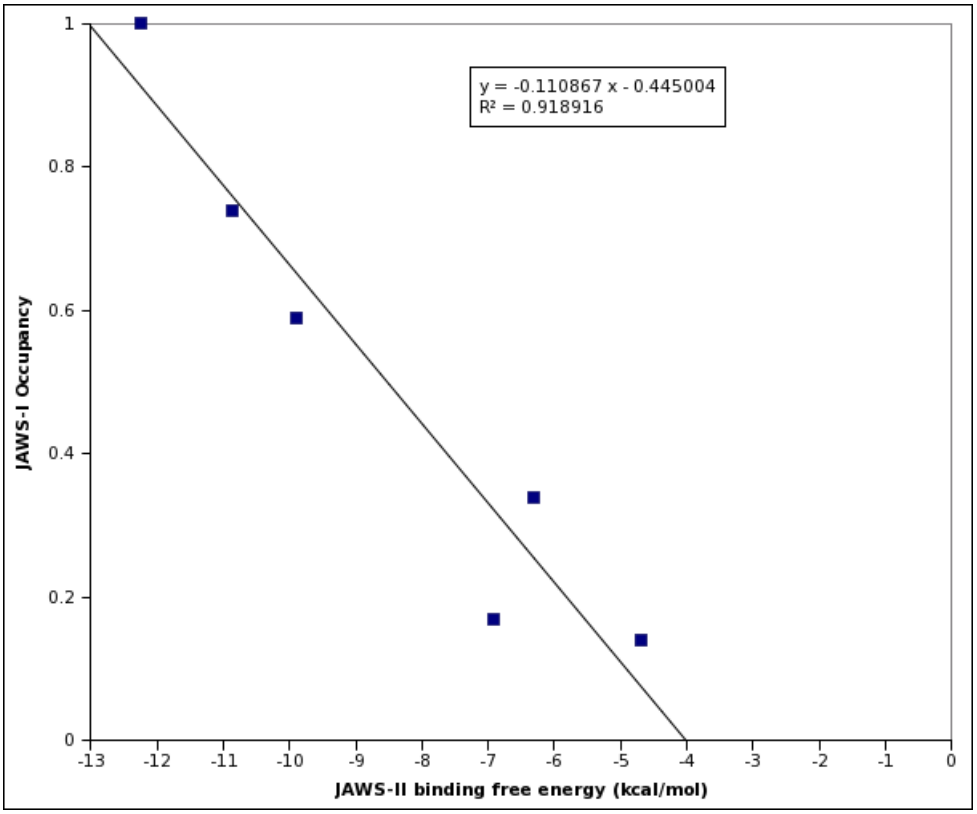


Figure 7.6: Correlation plot between the normalised JAWS stage 1 density for water molecules in N9-neuraminidase and the calculated JAWS stage 2 binding free energy for each site

waters is likely to be unsuccessful.

This fragment could then be restrained in space whilst the other molecules are allowed to sample the chemical space around the restrained fragment. The resulting maps would then give the correlation between the restrained fragments and the rest of the system, although several simulations would still be required to obtain sufficient data to analyse.

Such a procedure was performed for this system. The pyrazole contour located on the western ring was identified as the major site, and hence a pyrazole molecule was placed in this region. This molecule was restrained with a harmonic potential of 5 kcal/mol centred on the geometric centre of the ligand to prevent it from

drifting away during the simulation. A JAWS simulation was then performed on the system, with the results shown in Figure **??**.

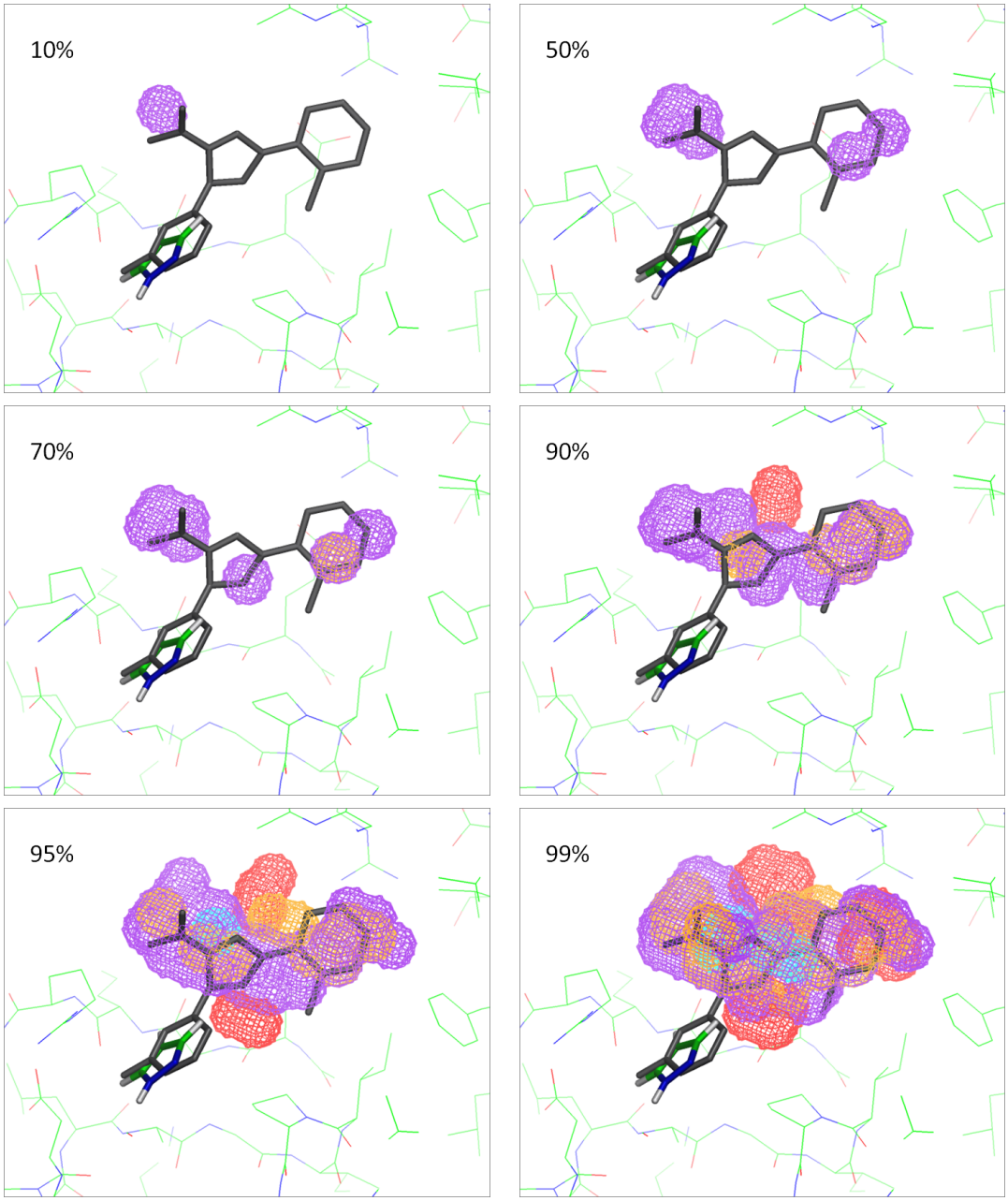


Figure 7.7: JAWS clustering density for the KSP system, using a restrained pyrazole fragment. The top data percentage for each image is shown in the top left. Key: pyrazole, benzene, water, acetone

Figure **??** shows that the major fragment site which is identified is that near the

acetyl group, corroborating the evidence from the other simulation that this is a highly favourable site for pyrazole. As with the standard JAWS run, both acetone and pyrazole are found near the right aromatic ring, whilst a pyrazole contour is now clearly found on the middle linker ring. A slightly higher concentration of water is also found in the pocket, with it occupying a position near the back of the right ring. This site is close to both E162 and R221, and hence its appearance here is not hugely surprising. Overall it can be seen that a similar picture is found when a pyrazole molecule is restrained, identifying the other two major pyrazole sites as occupying the right ring position and the acetyl group.

7.5 Factor Xa

7.5.1 Biological Relevance

fXa is a serine protease which is involved in the blood coagulation pathway. The protein converts prothrombin into thrombin, which in turn promotes the aggregation of platelets. A small amount of fXa produces a large amount of thrombin, and as a result is a more desirable target in the treatment of thrombosis rather than thrombin itself.[153] In addition, targetting fXa should target coagulation specifically, compared to thrombin inhibitiors which have been shown to affect hemostatis.[154] Existing therapies involving warfarin require careful monitoring of blood plasma levels and carry the risk of unfavourable metabolism and drug-drug interactions, and hence demonstrate the need for new drugs.[155]

Crystal structures of FXa show that inhibitors typically bind in an extended 'T' or 'L' shaped conformation, with key binding interactions in the ionic S1 pocket and hydrophobic S4 pocket of FXa. The ionic S1 pocket has been shown

to be capable of hosting a water molecule [153], although inhibitors have recently been developed which displace this water molecule by exploiting Cl-π interactions.[156] Such inhibitors do not have the electrostatic interaction with D189 like the original set of inhibitors. Further studies have shown that the S1 pocket is selective towards aryl-chloride containing inhibitors compared to aryl-bromides,aryl-florides and benzyls.[157]

7.5.2 System Preparation

A 15 Å scoop around the crystallographic ligand was prepared for the protein structure 2W3K using the same preparation as for N9-neuraminidase. A large binding site was identified around the crystallographic ligand of 3762 Å^3. 12 water molecules were used for the simulations, along with 8 fragment replicas.

7.5.3 JAWS simulations

JAWS simulations were performed upon the entire fXa binding pocket to examine the nature of the pocket, using the same protocol as for KSP. Three different fragments were used; water, benzene and pyrazole. Benzene was chosen since it is typical of a non-polar aromatic fragment, whilst pyrazole is a much more polar aromatic molecule. Water was chosen to compete with the other two fragments, allowing for hydration effects within the pocket. By comparing the relative populations of the three fragments, three objectives could be achieved:

1. Probing the FXa pocket to locate the polar and non-polar parts of the pocket;

2. Testing to see whether the methodology can predict the existence of polar moeities in the S1 pocket over less polar fragments;

3. Examining the possibility of hydration within the S1 pocket.

The collated JAWS results across 4 independent runs can be seen in Figure **??**. It can be seen that the major contour across the three fragments is found for pyrazole binding in the S4 pocket, mimicking the pyridone group in the native ligand. The key interaction in this region is a slipped π-π interaction, and it is not surprising that a polar heterocyclic ring is found in this region compared to benzene. As the contouring level is reduced to the top 75 % of data, pyrazole occupies the S1 pocket, as well as locating itself near to the other aromatic moieties in the native ligand. It is interesting to note that benzene is not present at either this contouring level or the next level down, suggesting that pyrazole outcompetes the fragment to a large extent.

As the contouring level is lowered to include the top 90 % of data, it can be seen that both water and benzene are now observed. Benzene is observed in the S1 pocket, as well as two of the other aromatic regions in the native ligand. It is not however seen in the S4 pocket, which still displays a strong preference for the pyrazole fragment. At the base of the S1 pocket a crystallographic water molecule is found, bridging an interaction between D189 and Y228. The JAWS simulation clearly identifies this site. Analysis of further fXa structures show that this molecule is constantly found across fXa structures in the PDB, as well as in related thrombin structures. As the top 92 % of data is observed, another crystallographic water is observed, this time one bridging the interaction between D189 and Y225. This water molecule is not found in all fXa and thrombin structures, and is ligand dependent.

From analysing the JAWS results, it can be seen that no water density is found in the bottom of the S1 pocket. This result corroborates previous work on fXa, whereby neutral substituents are found in the S1 pocket and promote the exclusion

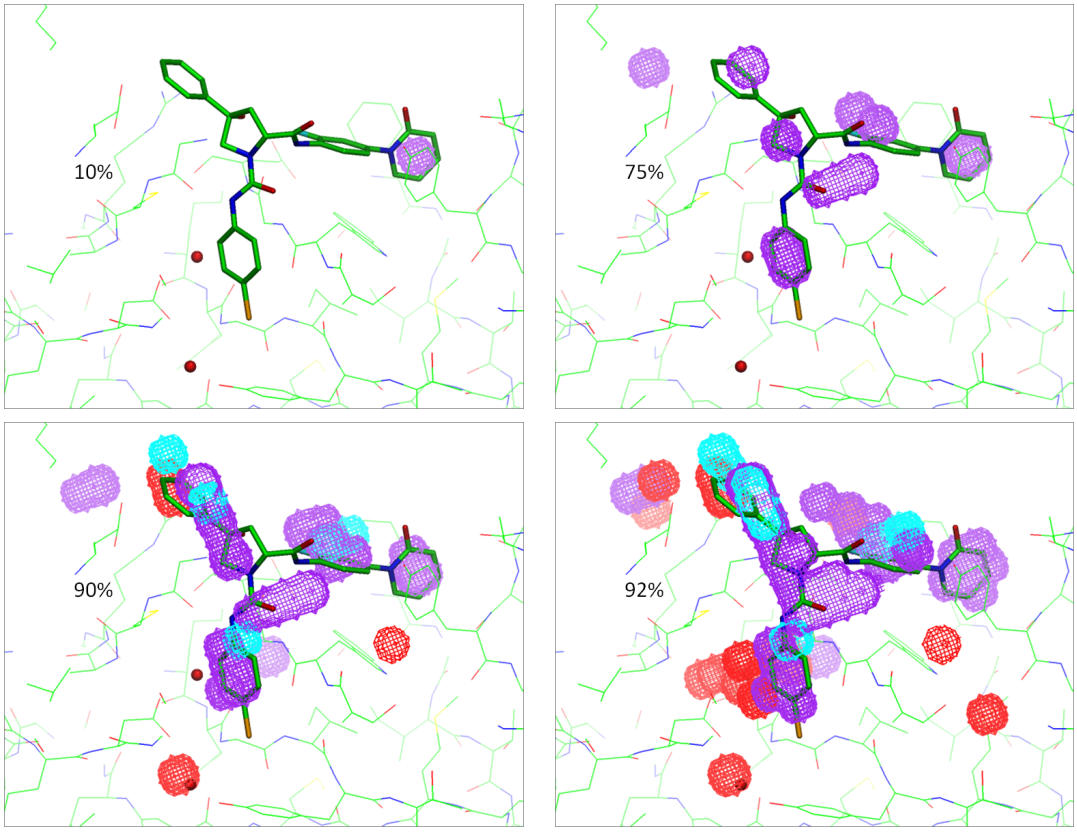


Figure 7.8: JAWS clustering density for the fXa system. The top data percentage for each image is shown on the left. Key: pyrazole, benzene, water

of water.[153, 154]. Since the S1 pocket is known to prefer polar aromatic groups it is perhaps not surprising that this result was obtained. However, if the pocket was challenged with only benzene and water, it is possible that water might outcompete the benzene and be observed in the base of the pocket. In order to investigate this possbility a further simulation was performed, challenging the pocket with benzene and water. The observed results can be seen in Figure **??**.

Figure **??** shows that, despite the pocket being challenged with a less polar fragment, water is still not observed in the base of the S1 pocket. The explanation for this is that volume of the benzene ring excludes the possibility of water. This suggests that the expulsion of water from the S1 pocket is clearly favourable

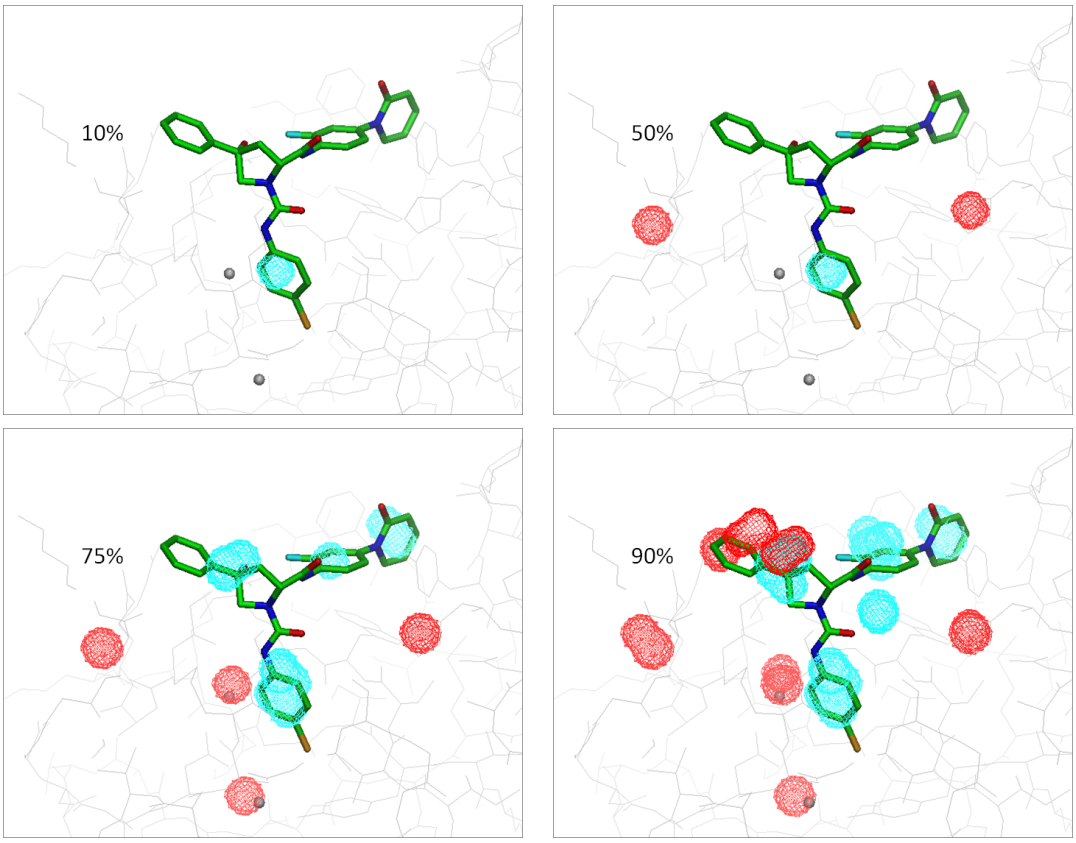


Figure 7.9: JAWS clustering density for the fXa system.The top data percentage for each image is shown on the left. Key: benzene, water

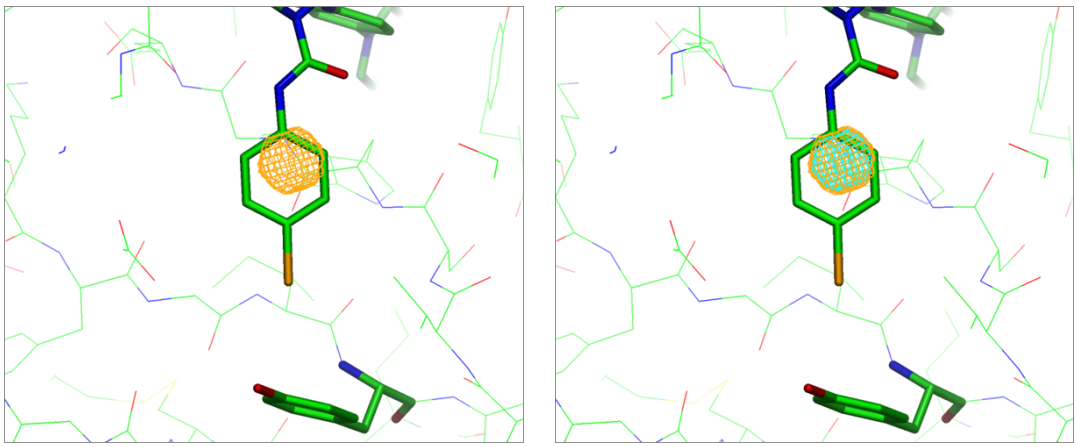


Figure 7.10: JAWS clustering density for the benzene-chlorobenzene simulation. **Left** : Top 50 % of data. **Right** : Top 60 % of data. Key: benzene, chlorobenzene

whenever an aromatic group is allowed to penetrate the entire pocket, and water is only able to coexist in the pocket when a charged electrostatic interaction is made between the ligand and D189.

7.5.4 Aryl Selectivity

As previously mentioned, the S1 pocket of fXa is selective towards aryl-chlorides compared to other aryl-halogens and benzene. In order to see whether the JAWS methodology is capable of identifying this selectivity trend, a series of simulations were performed which focussed solely on the S1 pocket. Only chlorobenzene and benzene was simulated; no water molecules were included. The Lennard-Jones softcore term was increased to 2 for this simulation to accurately describe the chlorobenzene desolvation term, an effect which did not significantly change the behaviour of the benzene desolvation profile. 2 replicas of each fragment were simulated. Four independent simulations were performed, with the results shown in Figure **??**.

Figure **??** shows that benzene only appears once the top 60 % of data is in-

cluded in the analysis, suggesting that chlorobenzene is outcompeting benzene in this pocket. Although the result is clearly encouraging, it is also important to recognise that standard MM forcefields do not describe halogens satisfactorily.[158] As such, the observed results could be fortuitous. Indeed, upon visual inspection it was found that the chloro moiety was pointing into the bulk rather than towards the tyrosine sidechain. This could suggest one of several things:

- The interaction between the chlorine atom and the bulk solvent is more favourable than the interaction between the atom and the sidechain. In the absence of a scaffold to anchor the chlorobenzene in position, this could promote the unexpected behaviour;
- The forcefield fails to adequately describe the halogen-π interaction.

In order to corroborate the result, the system could be simulated using a polarisable forcefield or a QM/MM approach.[159] This should elucidate whether the observed behaviour is indicative of a forcefield artifact or is, actually, correct. However such approaches typically require significantly more simulation time, and were beyond the timeframe of this thesis.

7.6 Limitations of the JAWS approach

Although the fragment based JAWS methodology has displayed great promise in identifying and ranking favourable binding poses for fragments, there are a number of potential drawbacks to the method. These are now highlighted and assessed.

1. **Lack of correlation between the sites** Since the JAWS output generates fragment maps which are collected over several independent simulations,

it is impossible to understand the correlation between different fragments. For example, fragment binding at one position in the system could be facilitated by another, more weakly bound, fragment. An averaged density map will not highlight this, and, if the correlated molecule occurs more rarely, the correlation may be masked by another fragment. Although restraining strongly bound fragments, as in the KSP system, will help to elucidate cooperativity, it is still not ideal and it would be desirable to have a method which can track cooperativity throughout the simulation.

2. **Lack of orientation in contours** During a JAWS simulation, information on the closest grid point to the centre of mass of a fragment is tracked. Whilst this is a relatively minor problem for a small molecule such as water, it becomes much more of an issue for larger fragments such as benzene. If the method is to be applied to more complex fragments where there is less symmetry, then strategies need to be devised to track and map the orientation of each fragment contour. Currently this can be achieved by manually looking through the PDB outputs and averaging the positions, but this is a time consuming and inefficient task.

3. **Assuming that density maxima correlates with binding free energy** Although Figure **??** has highlighted that, for water molecules in N9-neuraminidase, the density maxima can be taken to be a good approximation for the binding free energy, this does not necessarily mean it will hold true for fragments. Since the JAWS fragments are heterogeneous, it is important to establish a relationship between the density and the free energy. The major difficulty with this task lies in the calculation of a reliable free energy. As previously discussed in respect to JAWS stage 2 simulations, it is difficult what to sur-

round the resultant space with if one fragment has been docked. If the correlation and orientation of neighbouring fragments could be reliably found, then JAWS stage 2 and double-decoupling simulations could be performed to calculate the binding free energy. This would then allow the fragments to be ranked reliably due to their calculated free energy, rather than relying upon an approximation.

4. **How many molecules can be simulated simultaneously?** Although the JAWS methodology has been tested on mixtures of 4 fragments, there is the possibility that more fragments could be simulated simultaneously. One potential drawback lies in the sampling of the chemical space, since more fragments will induce more competition in the system. This will prohibit regions of the chemical space being sampled by other fragments if a favourable pose is accepted. Even with the inclusion of soft-core terms, it is possible that a more favourable pose will not be accepted if it cannot adequately sample the configurational space.

One other factor which has not yet been addressed is the overall effectiveness of the method in replicating real life FBDD results. During a fragment-based assay, a large quantity of data is collected. Some of this data will correspond to successful fragment-protein hits, whilst some of the data will correspond to unsuccessful fragment hits. The current method has highlighted cases where fragment contours overlap with known crystallographic ligand positions, but it has not yet been tested upon real life fragment-protein crystal structures.

Ideally, the method should be tested upon systems with knowledge of where certain fragments bind and do not bind. This would allow the robustness of the method to be assessed, and to see whether it is capable of predicting both the

location, pose and binding affinity of fragment-protein hits. This is something which is often a problem for docking approaches, since the force-fields used are often inaccurate and/or not sophisticated enough to discriminate between different binding modes. The ability of the method to discriminate between binders and non-binders would also be assessed through such a test.

In order to truly establish whether the method is capable of distinguishing between binders and non-binders, the method should be used in a blind test. In such a test the method would be tried on a system where 20 fragments are given, some of which are hits and some are decoys, with the knowledge of the binding characteristics of the fragments withheld. This would give information on the effectiveness of the method, and allow for further modifications to the algorithm if required.

Despite the current limitations and work to be done with the fragment-based JAWS algorithm, it is also important to reflect upon the success of the method. Unlike other FBDD approaches, the method is capable of calculating the positions of fragments in protein structures in competition with bulk water, whilst also allowing for an estimate of the desolvation cost of the fragment. This gives information on not only where the fragment binds in the system, but also whether it will leave the bulk aqueous phase to bind there. The method is based on a well described technique, gives rapid results, and is also capable of allowing sidechain motion. These features set it apart from existing computational tools and, by addressing the points above, the method is likely to be become more accurate and useful.

7.7 Conclusions

This chapter described the development of the JAWS methodology to FBDD, and subsequent application to two different case systems in factor Xa and KSP. The JAWS methodology was favoured over the GCMC method for a number of reasons, mainly in that the method is capable of competiting water against fragments during a simulation, and also in that the method displays improved acceptance rates. Since the JAWS method was designed for water molecules, three key factors needed to be considered:

- Dealing with the difference in standard states between a fragment and water;
- Incorporation of a PMF-based fragment decoupling term;
- Introduction of a soft-core potential to enhance sampling.

Having addressed these issues, the methodology was initially trialled upon the KSP system. The methodology predicted the aromatic binding regions of the pocket, identifying that they are likely to favour pyrazole binding over benzene. An additional site was also discovered around the binding pocket. SAR of various ligands suggests that nitrogen based linkers in this region are extremely favourable, corroborating the JAWS evidence of pyrazole binding in this region.

The method was then used upon the factor Xa system. The method identified the four aromatic regions in the system, suggesting a strong preference for pyrazole fragments throughout the system. This is consistent with experimental assays, whereby polar moieties are especially preferred in the S1 pocket. A slipped π-π interaction is found within the simulations in the S4 pocket, as well as two crystallographic water sites. The issue of water being present in the base of the S1 pocket was then explored, by challenging the pocket with benzene and water

only. The results showed that water was not observed, suggesting that the presence of water is disfavoured if benzene is allowed to penetrate the entire pocket. Finally, the issue of aryl-halide selectivity was addressed, with the results proving that chlorobenzene is favoured in the pocket over benzene.

The limitations of the JAWS methodology were then analysed. Although the method has delivered promising results for the two test systems, the method is not capable of predicting correlation between fragments. The method also cannot identify the orientational dependence of fragment binding, something which is important to achieve if the method is to be used for larger and more complex fragments than those currently used. An accurate way of determining the binding free energies of the fragments also needs to be developed, since the existing method of relying upon the fragment density needs to be validated. Mostly important, however, the method needs to be trialled upon real experimental data, which contains information of both successful and unsuccessful binders. The ability to differentiate between hits and decoys is something which would be highly desired, and ultimately the truest test of the method.

Chapter 8

Concluding remarks

The original proposal of this thesis was solely to explore the role of molecular fragments in inhibitor design - the roles of water molecules in protein binding sites was not originally considered. However, as the project progressed, it quickly became apparent that to ignore the role of water molecules would be foolish. Indeed, towards the end of the research, it was water molecules which became the primary focus of the study. This highlights the ever evolving nature of not only a PhD but research in itself, and how a scientist should always be ready to adapt to new challenges in the field.

In chapter two, a brief introduction to the importance of statistical mechanics in computational chemistry was given. Obviously providing full exposure to such a topic is beyond the scope of a thesis, but none the less it is important to provide some context. Computational chemistry has its roots in the Boltzmann distribution, and through a consideration of the partition function, all of the key thermodynamic properties of the system can be extracted. The potential energy of a system is found through the use of a molecular mechanics based force-field, which allows for the calculation of the intra- and intermolecular energy terms. The

ensemble phase space is typically sampled using either Monte Carlo or Molecular Dynamics simulations, whereby the evolution of the system is observed. Depending on the choice of ensemble, the key thermodynamic properties can then be easily analysed and calculated.

Chapter three describes the critical role of water molecules in protein binding sites, and how the ideas have been exploited in computational studies. From the original work by Poornima and Dean, the thinking behind water molecules has expanded significantly in recent years. It is now well recognised that water is far from a passive player in protein binding sites, and as such the incorporation of water molecules in molecular simulations is now beginning to be of significant concern. Indeed there is beginning to be a focus on the role of multiple water molecules in protein binding sites, with evidence showing that groups of waters can dictate protein-ligand specificity.

The four major computational methods for locating and scoring the binding affinity of water molecules were described; Grand Canonical Monte Carlo, Just Add Water Molecules, WaterMap and double-decoupling. Some applications of each method were analysed, alongside the relative merits and drawbacks of each of the methods. It was identified that no study had ever critically appraised and evaluated the methods on the same system; something which is of clear importance. Based upon this, and the availability of the methods, GCMC, JAWS and double-decoupling were chosen for the comparative study. The GCMC and JAWS method were incorporated into the in-house Monte Carlo code, ProtoMS, and contributed approximately 2000 lines of code.

Chapter four describes the fundamental features behind Fragment-Based Drug Discovery, and how it differs from the traditional High Throughput Screening approach. FBDD offers a route for the medicinal chemist to build molecules from

a small fragment, allowing the key physiological parameters such as molecular weight and lipophilicity to be controlled in the fragment to lead development. Experimental FBDD carries considerable financial cost, providing an incentive for computational approaches. The existing *in silico* methods all suffer several drawbacks; in most methods this is due to a poor treatment of solvation in the binding site and the lack of protein flexibility. The key features for an ideal FBDD computational method were identified, in preparation for these to be incorporated into the JAWS algorithm.

Chapter five details the key developments made to the JAWS algorithm and how, by modifying the applied potential bias, the binding free energy of strongly bound waters can be calculated. The application of the 3 methods to N9-Neuraminidase was described, and demonstrated that all three methods give excellent agreement in the calculation of binding free energies. Both GCMC methods; simulated annealing and the interacting particle method, were also shown to be consistent in the calculations. Based upon the N9-Neuraminidase results, the relative merits and drawbacks of each method was assessed. The double-decoupling method is, by far, the most rigorous method for the scoring waters, although the computational time is of an order of magnitude slower than the other two methods. Both GCMC and JAWS deliver an estimate of the binding free energy in a rapid, yet accurate, manner, and are also capable of locating water molecules - something of which double-decoupling is not capable. One important application of GCMC was highlighted in the BPTI study, where it was found that the method predicts changes in hydration patterns as a function of the chemical potential. The idea of network reorganisation and stabilisation as a function of the chemical potential is something which forms the basis of the chapter 5.

Chapter six details the application of the three methods to a range of different

case studies and scenarios. The use of the methods to the hydration of hydrophobic cavities was firstly described, looking at the T4-lysozyme and Interleukin 1β systems. It was found that, for both systems, the cavities were dry, in agreement with experimental evidence. The GCMC methodology was found to be advantageous for the location of water sites in comparison to JAWS, owing to the ease of identifying cooperativity between water molecules. A general scheme was adopted; using GCMC to locate the water sites, then applying JAWS stage 2 simulations to calculate the binding free energy.

The hydration of two different protein kinases was then studied; CDK2 and Pim-1. The two kinases have different sequences along the hinge region, providing a good test of the methods to predict different hydration patterns. GCMC simulations were performed at a range of different chemical potentials to observe the changes in the water locations, with the configurations of waters scored using JAWS-2. It was found that, when weakly unbound waters are allowed to occupy the CDK2 pocket, the binding free energy of two waters along the hinge are stabilised by approximately 4 kcal/mol. This network stabilisation was identified as being highly significant for drug design, since targetting the weakly unbound waters in the network is likely to reduce the desolvation penalty of the protein upon ligand binding. A similar scenario was found near to the critical catalytic lysine residue in Pim-1, where weakly bound waters stabilise the network.

The idea of water network destabilisation can be used to determine how the apo network influences ligand binding. For a ligand to bind to a target, the water molecules in the protein must either be displaced or conserved and incorporated into the protein-ligand complex. Several criteria have been identified for deciding the best strategy for dealing with the water molecules. For a water molecule to be conserved in the complex it must have a binding free energy indicative of a

bound water, and be stable in the absence or partial absence of the network. The magnitude of destabilisation can be expressed as Δ ; waters with high values of Δ are destabilised when the network is disrupted, and are likely to be displaced upon ligand binding. Thus, waters which have low values of Δ are sufficiently stable in the absence of the network and offer opportunities to incorporate them into a protein-ligand complex.

One final application of the methods was used to study the Chk-1 kinase system. Two different ligands, 5CH and 5BT, were studied, varying only in the connectivity of an amide group. This change causes a change in the binding free energy of 1.2 kcal/mol in favour of the 5CH ligand ; a shift which is unexpected since the protein-ligand interactions appear to be extremely similar. Dual topology simulations were performed on the protein-ligand structures, finding that the 5CH ligand is indeed more stable. Energetic analysis suggested that the major change between the ligands is the solvent-ligand energy, with a cluster of three waters identified as the primary cause. JAWS-2 simulations were performed on the different ligand systems, finding that a distal water is stabilised in the 5CH structure compared to the 5BT structure.

When viewed alongside GCMC simulations, it was found that the network stabilisation in the 5CH ligand was in good qualitative agreement with the dual topology simulations. Double decoupling simulations were used to calculate the binding free energies of the weakly bound waters and, by completing a free energy cycle, it provided compelling evidence that the role of the water cluster is responsible for the change in ligand affinity. Such an effect has not been described in the chemical literature, and provides clear evidence for the active role which water molecules perform upon protein-ligand association. Crucially, it is a distal water which provides the key change in the binding affinity, highlighting that molecular

association is far from a straightforward and understood process.

Chapter seven describes the modifications made to the JAWS algorithm to allow for the simulation of fragments. Three different features were added to the algorithm. Firstly a desolvation penalty was applied to fragments, capturing the fact that for a fragment to bind it must move out the bulk and lose its solvation shell. This penalty was incorporated through running dual-topology simulations prior to the JAWS process, and then using the PMF profile to correct for the penalty as a function of θ. The second modification was the incorporation of a standard state volume correction, which corrects for the difference between simulating water at 55 M and fragments at 1 M. Finally soft-core potentials were incorporated into the code, to allow for adequate sampling of the chemical space.

Two different protein systems were investigated; the mitotic cancer target KSP and factor Xa. Both systems demonstrated that the method is capable of predicting where different fragments prefer to locate, with the results validated against known protein-ligand structural data. Crucially the method is capable of locating where specific water molecules locate whilst also allowing fragment competition, with these locations in excellent agreement with crystallographic data. It is important to recognise that, although the results obtained are encouraging, the method requires much more validation. The optimisation of the protocol against real data provides an excellent route for method validation, and efforts have begun to obtain realistic protein systems to test the JAWS methodology upon.

One of the original aims of the thesis was to explore the nature of fragments in drug design, and it can be said that this has been achieved. Although much more validation is required, a method has been developed which takes into account many of the major deficiencies in existing FBDD computational approaches. Perhaps the biggest achievement in this thesis has been the work performed on the

role of water molecules in protein binding sites. Significant lessons have been learnt about how water molecules behave in proteins, and how they can dictate ligand binding and affinity. The idea of using a Δ parameter to gain information on the destabilisation of waters shows great promise, and is something which is likely to be of considerable interest in the field of drug design and development. Efforts have already begun to further understand and validate the idea of the Δ parameter looking at the N9 neuraminidase and Scylatone Dehydratase binding sites, and suggest that waters can be significantly stabilised upon the introduction of a ligand compared to the apo binding free energies.

Bibliography

[1] Caballero, J., Zilocchi, S., Tiznado, W., and Collina, S. *Med. Chem. Res.* **21**, 1912–1920 (2011).

[2] Amaro, R. E., Cheng, X., Ivanov, I., Xu, D., and McCammon, J. A. *J. Am. Chem. Soc.* **131**(13), 4702–4709 (2009).

[3] Michel, J., Verdonk, M. L., and Essex, J. W. *J. Med. Chem.* **49**(25), 7427–7439 (2006).

[4] Erlanson, D. A. *Curr. Opin. Biotech.* **17**(6), 643–652 (2006).

[5] Poornima, C. S. and Dean, P. M. *J. Comp. Aid. Mol. Des.* **9**(6), 500–512 (1995).

[6] Poornima, C. S. and Dean, P. M. *J. Comp. Aid. Mol. Des.* **9**(6), 513–520 (1995).

[7] Poornima, C. S. and Dean, P. M. *J. Comp. Aid. Mol. Des.* **9**(6), 521–531 (1995).

[8] Barillari, C. *The Role of Water in Protein-Ligand Interactions: Implications for Rational Drug Design*. PhD thesis, (2006).

[9] Congreve, M., Chessari, G., Tisi, D., and Woodhead, A. J. *J. Med. Chem.* **51**(13), 3661–3680 (2008).

[10] Hajduk, P. J. and Greer, J. *Nat. Rev. Drug Discov.* **6**(3), 211–219 (2007).

[11] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. *Ad. Drug Del. Rev.* **46**(1-3), 3–26 (2001).

[12] Carugo, O. and Bordo, D. *Acta Cryst. Sec. D Biol. Cryst.* **55**(Pt 2), 479–483 (1999).

[13] Michel, J., Tirado-Rives, J., and Jorgensen, W. L. *J. Phys. Chem. B* **113**(40), 13337–13346 (2009).

[14] Gilson, M. K., Given, J. A., Bush, B. L., and McCammon, J. A. *Biophys. J.* **72**(3), 1047–1069 (1997).

[15] Adams, D. J. *Mol. Phys.* **28**(5), 1241–1252 (1974).

[16] Adams, D. J. *Mol. Phys.* **29**(1), 307–311 (1975).

[17] Voelz, V., Bowman, G. R., Beauchamp, K., and Pande, V. S. *J. Am. Chem. Soc.* **132**(5), 1526–1528 (2010).

[18] Orsi, M. and Essex, J. W. *Soft Matter* **6**(16), 3797–3808 (2010).

[19] Leach, A. *Molecular Modelling : Principles and Applications*. 2nd edition, (2001).

[20] Allen, M. P. and Tildesley, D. J. *Computer Simulation of Liquids*. (1989).

[21] Atkins, P. W. *Physical Chemistry*. 3rd edition, (1988).

[22] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. *J. Comp. Chem.* **25**(9), 1157–1174 (2004).

[23] Oostenbrink, C., Villa, A., Mark, A. E., and Van Grunsteren, W. F. *J. Comp. Chem.* **25**(13), 1656–1676 (2004).

[24] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. *J. Chem. Phys.* **21**(6), 1087 (1953).

[25] Michel, J. *The use of free energy simulations as scoring functions*. PhD thesis, (2006).

[26] Zwanzig, R. W. *J. Chem. Phys.* **22**(8), 1420–1426 (1954).

[27] Mezei, M. *J. Chem. Phys.* **86**, 7084 (1987).

[28] Woods, C. J., Essex, J. W., and King, M. A. *J. Phys. Chem. B* **107**(49), 13711–13718 (2003).

[29] Woods, C. J., Essex, J. W., and King, M. A. *J. Phys. Chem. B* **107**(49), 13703–13710 (2003).

[30] Massova, I. and Kollman, P. *Perspec. Drug Dis. Des.* **18**(1), 113–135 (2000).

[31] Guimarães, C. R. W. and Mathiowetz, A. M. *J. Chem. Inf. Model.* **50**(4), 547–559 (2010).

[32] Graves, A. P., Shivakumar, D. M., Boyce, S. E., Jacobson, M. P., Case, D. A., and Shoichet, B. K. *J. Mol. Biol.* **377**(3), 914–934 (2008).

[33] Cappel, D., Wahlstrom, R., Brenk, R., and Sotriffer, C. A. *J. Chem. Inf. Model.* **51**(10), 2581–94 (2011).

[34] Wallnoefer, H. G., Liedl, K. R., and Fox, T. *J. Chem. Inf. Model.* **51**(11), 2860–7 (2011).

[35] Li, Y., Sutch, B. T., Bui, H.-H., Gallaher, T. K., and Haworth, I. S. *J. Chem. Inf. Model.* **51**(6), 1347–1352 (2011).

[36] Fadda, E. and Woods, R. J. *J. Chem. Theory Comput.* **7**(10), 3391–3398 (2011).

[37] Barillari, C., Duncan, A. L., Westwood, I. M., Blagg, J., and Van Montfort, R. L. M. *Proteins: Struct. Funct. Bioinf.* **79**(7), 2109–2121 (2011).

[38] Wong, S. E. and Lightstone, F. C. *Exp. Opin. Drug Discov.* **6**(1), 65–74 (2011).

[39] Huggins, D. J. and Tidor, B. *Prot. Eng. Des. Sel.* **24**(10), 777–789 (2011).

[40] Barillari, C., Taylor, J., Viner, R., and Essex, J. W. *J. Am. Chem. Soc.* **129**(9), 2577–2587 (2007).

[41] Ross, G. A., Morris, G. M., and Biggin, P. C. *PLoS ONE* **7**(3), e32036 (2012).

[42] Homans, S. W. *Drug. Discov. Today* **12**(13-14), 534–539 (2007).

[43] Setny, P., Baron, R., and McCammon, J. A. *J. Chem. Theory Comput.* **6**(9), 2866–2871 (2010).

[44] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. *J. Chem. Phys.* **79**(2), 926–935 (1983).

[45] Baron, R., Setny, P., and Andrew McCammon, J. *J. Am. Chem. Soc.* **132**(34), 12091–12097 (2010).

[46] Hummer, G. *Nat. Chem.* **2**(11), 906–907 (2010).

[47] Ball, P. *Nature* **478**, 467–468 (2011).

[48] Abel, R., Salam, N. K., Shelley, J., Farid, R., Friesner, R. A., and Sherman, W. *Chem. Med. Chem.* **6**(6), 1049 – 1066 (2011).

[49] Davis, A. M., Teague, S. J., and Kleywegt, G. J. *Angew. Chem. Int. Ed.* **42**(24), 2718–2736 (2003).

[50] Woo, H.-J., Dinner, A. R., and Roux, B. *J. Chem. Phys.* **121**(13), 6392–400 (2004).

[51] Mezei, M. *Mol. Phys.* **40**(4), 901–906 (1980).

[52] Mezei, M. *Mol. Phys.* **61**, 565–582 (1987).

[53] Shelley, J. C. and Patey, G. N. *J. Chem. Phys.* **102**(19), 7656 (1995).

[54] Guarnieri, F. and Mezei, M. *J. Am. Chem. Soc.* **118**, 8493–8494 (1996).

[55] Resat, H. and Mezei, M. *Phys. Rev. E* **62**(5 Pt B), 7077–7081 (2000).

[56] Pan, C., Mezei, M., Mujtaba, S., Muller, M., Zeng, L., Li, J., Wang, Z., and Zhou, M.-M. *J. Med. Chem.* **50**(10), 2285–2288 (2007).

[57] Im, W., Berneche, S., and Roux, B. *J. Chem. Phys.* **114**(7), 2924 (2001).

[58] Deng, Y. and Roux, B. *J. Chem. Phys.* **128**(11), 115103 (2008).

[59] Collins, M. D., Hummer, G., Quillin, M. L., Matthews, B. W., and Gruner, S. M. *Proc. Nat. Ac. Sci.* **102**(46), 16668–16671 (2005).

[60] Kong, X. and Brooks III, C. L. *J. Chem. Phys.* **105**, 2414–2423 (1996).

[61] Michel, J., Tirado-Rives, J., and Jorgensen, W. L. *J. Am. Chem. Soc.* **131**(42), 15403–15411 (2009).

[62] Luccarelli, J., Michel, J., Tirado-Rives, J., and Jorgensen, W. L. *J. Chem. Theory Comput.* **6**(12), 3850–3856 (2010).

[63] Pearlman, D. A. and Charifson, P. S. *J. Med. Chem.* **44**(21), 3417–3423 (2004).

[64] Lazaridis, T. *J. Phys. Chem. B* **102**(18), 3531–3541 (1998).

[65] Lazaridis, T. *J. Phys. Chem. B* **102**(18), 3542–3550 (1998).

[66] Li, Z. and Lazaridis, T. *J. Am. Chem. Soc.* **125**(22), 6636–6637 (2003).

[67] Young, T., Abel, R., Kim, B., Berne, B. J., and Friesner, R. A. *Proc. Nat. Ac. Sci.* **104**(3), 808–813 (2007).

[68] Abel, R., Young, T., Farid, R., Berne, B. J., and Friesner, R. A. *J. Am. Chem. Soc.* **130**(9), 2817–2831 (2008).

[69] Wang, L., Berne, B. J., and Friesner, R. A. *Proc. Nat. Ac. Sci.* **108**(4), 1326–1330 (2011).

[70] Robinson, D. D., Sherman, W., and Farid, R. *Chem. Med. Chem.* **5**(4), 618–627 (2010).

[71] Zhang, L. and Hermans, J. *Proteins: Struc. Func. And Gen.* **24**, 433–438 (1996).

[72] Olano, L. R. and Rick, S. W. *J. Am. Chem. Soc.* **126**(25), 7991–8000 (2004).

[73] Roux, B., Nina, M., Pomès, R., and Smith, J. C. *Biophys. J.* **71**(2), 670–681 (1996).

[74] Helms, V. and Wade, R. C. *Biophys. J.* **69**(3), 810–824 (1995).

[75] Hamelberg, D. and McCammon, J. A. *J. Am. Chem. Soc.* **126**(24), 7683–7689 (2004).

[76] Wenlock, M. C., Austin, R. P., Barton, P., Davis, A. M., and Leeson, P. D. *J. Med. Chem.* **46**(7), 1250–1256 (2003).

[77] Murray, C. W., Verdonk, M. L., and Rees, D. C. *Trends Pharm. Sci.* **33**(5), 224–32 May (2012).

[78] Warr, W. A. *J. Comp. Aid. Mol. Des.* **47**(8), 1–14 (2009).

[79] Hann, M. M. *Med. Chem. Commun.* **2**, 349–355 (2011).

[80] Hesterkamp, T. and Whittaker, M. *Curr. Opin. Chem. Biol.* **12**(3), 260–268 (2008).

[81] Pröll, F., Fechner, P., and Proll, G. *Anal. Bioanal. Chem.* **393**(6-7), 1557–1562 (2009).

[82] Wang, Y.-S., Strickland, C., Voigt, J. H., Kennedy, M. E., Beyer, B. M., Senior, M. M., Smith, E. M., Nechuta, T. L., Madison, V. S., Czarniecki, M., McKittrick, B. A., Stamford, A. W., Parker, E. M., Hunter, J. C., Greenlee, W. J., and Wyss, D. F. *J. Med. Chem.* **53**(3), 942–950 (2010).

[83] Murray, C. W., Carr, M. G., Callaghan, O., Chessari, G., Congreve, M., Cowan, S., Coyle, J. E., Downham, R., Figueroa, E., Frederickson, M., Graham, B., McMenamin, R., O'Brien, M. A., Patel, S., Phillips, T. R., Williams, G., Woodhead, A. J., and Woolford, A. J.-A. *J. Med. Chem.* **53**(16), 5942–5955 (2010).

[84] Mattos, C., Bellamacina, C. R., Peisach, E., Pereira, A., Vitkup, D., Petsko, G. A., and Ringe, D. *J. Mol. Biol.* **357**(5), 1471–1482 (2006).

[85] Ciulli, A. and Abell, C. *Curr. Opin. Biotech.* **18**(6), 489–496 (2007).

[86] Medina, J. R., Becker, C. J., Blackledge, C. W., Duquenne, C., Feng, Y., Grant, S. W., Heerding, D., Li, W. H., Miller, W. H., Romeril, S. P., Scherzer, D., Shu, A., Bobko, M. A., Chadderton, A. R., Dumble, M., Gardiner, C. M., Gilbert, S., Liu, Q., Rabindran, S. K., Sudakin, V., Xiang, H., Brady, P. G., Campobasso, N., Ward, P., and Axten, J. M. *J. Med. Chem.* **1**(6), 1871–95 (2011).

[87] Jahnke, W., Rondeau, J.-M., Cotesta, S., Marzinzik, A., Pellé, X., Geiser, M., Strauss, A., Götte, M., Bitsch, F., Hemmig, R., Henry, C., Lehmann, S., Glickman, J. F., Roddy, T. P., Stout, S. J., and Green, J. R. *Nat. Chem. Biol.* **6**(9), 660–666 (2010).

[88] De Kloe, G. E., Retra, K., Geitmann, M., Källblad, P., Nahar, T., Van Elk, R., Smit, A. B., Van Muijlwijk-Koezen, J. E., Leurs, R., Irth, H., Danielson, U. H., and De Esch, I. J. P. *J. Med. Chem.* **53**(19), 7192–7201 (2010).

[89] Whittaker, M., Law, R. J., Ichihara, O., Hesterkamp, T., and Hallett, D. *Drug Discov. Today : Tech.* **7**(3), e163–e171 September (2010).

[90] Kawatkar, S., Wang, H., Czerminski, R., and Joseph-McCarthy, D. *J. Comp. Aid. Mol. Des.* **23**(8), 527–539 (2009).

[91] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., P, F., and S, S. P. *J. Med. Chem.* **47**(7), 1739–1749 (2004).

[92] Chen, Y. and Pohlhaus, D. T. *Drug Discov. Today : Tech.* **7**(3), e149–e156 (2010).

[93] Verdonk, M., Giangreco, I., Hall, R. J., Korb, O., Mortenson, P. N., and Murray, C. W. *J. Med. Chem.* **54**, 5422–5431 (2011).

[94] Taylor, R. D., Jewsbury, P. J., and Essex, J. W. *J. Comp. Chem.* **24**(13), 1637–1656 (2003).

[95] Miranker, A. and Karplus, M. *Proteins: Struct. Funct. Bioinf.* **11**(1), 29–34 (1991).

[96] Schubert, C. R. and Stultz, C. M. *J. Comp. Aid. Mol. Des.* **23**(8), 475–489 (2009).

[97] Brenke, R., Kozakov, D., Chuang, G.-Y., Beglov, D., Hall, D., Landon, M. R., Mattos, C., and Vajda, S. *Bioinf.* **25**(5), 621–627 (2009).

[98] Stultz, C. M. *J. Phys. Chem. B* **108**(42), 16525–16532 (2004).

[99] Haider, M. K., Bertrand, H.-O., and Hubbard, R. E. *J. Chem. Inf. Model.* **51**(5), 1092–1105 (2011).

[100] Landon, M. R., Lieberman, R. L., Hoang, Q. Q., Ju, S., Caaveiro, J. M. M., Orwig, S. D., Kozakov, D., Brenke, R., Chuang, G.-Y., Beglov, D., Vajda, S., Petsko, G. A., and Ringe, D. *J. Comp. Aid. Mol. Des.* **23**(8), 491–500 (2009).

[101] Ivetac, A. and McCammon, J. A. *Chem. Biol. Drug Des.* **76**(3), 201–217 (2010).

[102] Guvench, O. and MacKerell, A. D. *PLoS Comp. Biol.* **5**(7), 10 (2009).

[103] Raman, E. P., Yu, W., Guvench, O., and Mackerell, A. D. *J. Chem. Inf. Model.* **51**(4), 877–96 (2011).

[104] Imai, T., Oda, K., Kovalenko, A., Hirata, F., and Kidera, A. *J. Am. Chem. Soc.* **131**(34), 12430–12440 (2009).

[105] Stumpe, M. C., Blinov, N., Wishart, D., Kovalenko, A., and Pande, V. S. *J. Phys. Chem. B* **115**(2), 319–328 (2011).

[106] Clark, M., Guarnieri, F., Shkurko, I., and Wiseman, J. *J. Chem. Inf. Model.* **46**(1), 231–242 (2006).

[107] Hermans, J. and Wang, L. *J. Am. Chem. Soc.* **119**(11), 2707–2714 (1997).

[108] Clark, M., Meshkat, S., and Wiseman, J. S. *J. Chem. Inf. Model.* **49**(4), 934–943 (2009).

[109] Michel, J. and Essex, J. W. *J. Comp. Aid. Mol. Des.* **24**(8), 639–658 (2010).

[110] Gubaerva, L. *Vir. Res.* **103**(1), 199–203 (2004).

[111] Varghese, J. N. *Drug Dev. Res.* **46**, 176–196 (1999).

[112] Yen, H.-L., Hoffmann, E., Taylor, G., Scholtissek, C., Monto, A. S., Webster, R. G., and Govorkova, E. A. *J. Virol.* **80**(17), 8787–8795 (2006).

[113] Varghese, J. N., Epa, V. C., and Colman, P. M. *Prot. Sci.* **4**(6), 1081–1087 (1995).

[114] Vriend, G. *J. Mol. Graph.* **8**(1), 52–56, 29 (1990).

[115] Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. *J. Comp. Chem.* **21**(2), 132–146 (2000).

[116] Wang, J., Cieplak, P., and Kollman, P. A. *J. Comp. Chem.* **21**(12), 1049–1074 (2000).

[117] Hartshorn, M. J. *J. Comp. Aid. Mol. Des.* **16**(12), 871–881 (2002).

[118] Woods, C. and Michel, J. (2007).

[119] Zakharova, E., Horvath, M. P., and Goldenberg, D. P. *Proc. Nat. Ac. Sci.* **106**(27), 11034 (2009).

[120] Mannucci, P. M. *New Eng. J. Med.* **339**, 245–253 (1998).

[121] Wlodawer, A., Walter, J., Huber, R., and Sjölin, L. *J. Mol. Biol.* **180**(2), 301–329 (1984).

[122] Hummer, G., Garde, S., García, A. E., Paulaitis, M. E., and Pratt, L. R. *Proc. Nat. Ac. Sci.* **95**(4), 1552–1555 (1998).

[123] Schlichting, I., Berendzen, J., Chu, K., Stock, A. M., Maves, S. A., Benson, D. E., Sweet, R. M., Ringe, D., Petsko, G. A., and Sligar, S. G. *Science* **287**(5458), 1615–1622 (2000).

[124] Collins, M. D., Quillin, M. L., Hummer, G., Matthews, B. W., and Gruner, S. M. *J. Mol. Biol.* **367**(3), 752–763 (2007).

[125] Schopf, P. *Development and application of free energy methods*. PhD thesis, (2011).

[126] Yin, H., Feng, G., Clore, G. M., Hummer, G., and Rasaiah, J. C. *J. Phys. Chem. B* **114**(49), 16290–16297 (2010).

[127] Ernst, J. A., Clubb, R. T., Zhou, H. X., Gronenborn, A. M., and Clore, G. M. *Science* **267**(5205), 1813–1817 (1995).

[128] Somani, S., Chng, C.-P., and Verma, C. S. *Proteins: Struct. Funct. Bioinf.* **67**(4), 868–885 (2007).

[129] Adamek, D. H., Guerrero, L., Blaber, M., and Caspar, D. L. D. *J. Mol. Biol.* **346**(1), 307–318 (2005).

[130] Clore, G. M., Wingfield, P. T., and Gronenborn, A. M. *Biochem.* **30**(9), 2315–2323 (1991).

[131] Jorgensen, W. L. *J. Am. Chem. Soc.* **103**, 335–340 (1981).

[132] Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. *Science* **298**(5600), 1912–34 (2002).

[133] Malumbres, M. and Barbacid, M. *Nat. Rev. Can.* **9**(3), 153–166 (2009).

[134] Vieth, M., Sutherland, J. J., Robertson, D. H., and Campbell, R. M. *Drug Discov. Today* **10**(12), 839–846 (2005).

[135] Pierce, A. C., Sandretto, K. L., and Bemis, G. W. *Proteins: Struct. Funct. Bioinf.* **49**(4), 567–576 (2002).

[136] Qian, K., Wang, L., Cywin, C. L., Farmer, B. T., Hickey, E., Homon, C., Jakes, S., Kashem, M. A., Lee, G., Leonard, S., Li, J., Magboo, R., Mao, W., Pack, E., Peng, C., Prokopowicz, A., Welzel, M., Wolak, J., and Morwick, T. *J. Med. Chem.* **52**(7), 1814–1827 (2009).

[137] Mezei, M. and Beveridge, D. L. *J. Comp. Chem.* **6**, 523–527 (1984).

[138] Mezei, M. (2012).

[139] Hess, B., Kutzner, C., Van Der Spoel, D., and Lindahl, E. *J. Chem. Theory Comput.* **4**(3), 435–447 (2008).

[140] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. *Proteins* **65**(3), 712–725 (2006).

[141] Bachmann, M. and Möröy, T. *Int. J. Biochem. Cell Biol.* **37**(4), 726–730 (2005).

[142] Beuming, T., Che, Y., Abel, R., Kim, B., Shanmugasundaram, V., and Sherman, W. *Proteins: Struct. Funct. Bioinf.* **80**, 871–883 (2012).

[143] Oza, V., Ashwell, S., Brassil, P., Breed, J., Deng, C., Ezhuthachan, J., Haye, H., Horn, C., Janetka, J., Lyne, P., Newcombe, N., Otterbien, L., Pass, M., Read, J., Roswell, S., Su, M., Toader, D., Yu, D., Yu, Y., Valentine, A., Webborn, P., White, A., Zabludoff, S., and Zheng, X. *Bioorg. Med. Chem. Lett.* **20**(17), 5133–5138 (2010).

[144] Prudhomme, M. *Rec. Pat. Anti-Cancer Drug. Discov.* **1**, 55–68 (2006).

[145] Simonson, T., Archontis, G., and Karplus, M. *Acc. Chem. Res.* **35**, 430–437 (2002).

[146] Liu, H., Mark, A. E., and Van Gunsteren, W. F. *J. Phys. Chem. A* **100**(22), 9485–9494 (1996).

[147] Panagiotopoulos, A. Z. *Mol. Phys.* **61**(1), 813–826 (1987).

[148] Mobley, D. L., Bayly, C. I., Cooper, M. D., Shirts, M. R., and Dill, K. A. *J. Chem. Theory Comput.* **5**(2), 350–358 (2009).

[149] Sarli, V. and Giannis, A. *Clin. Can. Res.* **14**(23), 7583–7587 (2008).

[150] Kapoor, T. M., Mayer, T. U., Coughlin, M. L., and Mitchison, T. J. *J. Cell Biol.* **150**(5), 975–988 (2000).

[151] Cox, C. D., Breslin, M. J., Mariano, B. J., Coleman, P. J., Buser, C. A., Walsh, E. S., Hamilton, K., Huber, H. E., Kohl, N. E., Torrent, M., Yan, Y., Kuo, L. C., and Hartman, G. D. *Bioorg. Med. Chem. Lett.* **17**(20), 5677–5682 (2007).

[152] Cox, C. D., Breslin, M. J., Mariano, B. J., Coleman, P. J., Buser, C. A., Walsh, E. S., Hamilton, K., Huber, H. E., Kohl, N. E., Torrent, M., Yan, Y., Kuo, L. C., and Hartman, G. D. *Bioorg. Med. Chem. Lett.* **16**(7), 1775–1779.

[153] Maignan, S., Guilloteau, J. P., Pouzieux, S., Choi-Sledeski, Y. M., Becker, M. R., Klein, S. I., Ewing, W. R., Pauls, H. W., Spada, A. P., and Mikol, V. *J. Med. Chem.* **43**(17), 3226–3232 (2000).

[154] Matter, H., Defossa, E., Heinelt, U., Blohm, P.-M., Schneider, D., Müller, A., Herok, S., Schreuder, H., Liesum, A., Brachvogel, V., Lönze, P., Walser, A., Al-Obeidi, F., and Wildgoose, P. *J. Med. Chem.* **45**(13), 2749–2769 (2002).

[155] Van Huis, C. A., Casimiro-Garcia, A., Bigge, C. F., Cody, W. L., Dudley, D. A., Filipski, K. J., Heemstra, R. J., Kohrt, J. T., Leadley, R. J., Narasimhan, L. S., McClanahan, T., Mochalkin, I., Pamment, M., Peterson, J. T., Sahasrabudhe, V., Schaum, R. P., and Edmunds, J. J. *Bioorg. Med. Chem.* **17**(6), 2501–2511 (2009).

[156] Maignan, S., Guilloteau, J.-P., Choi-Sledeski, Y. M., Becker, M. R., Ewing, W. R., Pauls, H. W., Spada, A. P., and Mikol, V. *J. Med. Chem.* **46**(5), 685–690 (2003).

[157] Matter, H., Will, D. W., Nazare, M., Schreuder, H., Laux, V., and Wehner, V. *J. Med. Chem.* **48**(9), 3290–3312 (2005).

[158] Boyd R. H. and Kesner L. *J. Chem. Phys.* **72**(3), 2179–2190 (1980).

[159] Beierlein, F. B., Michel, J., and Essex, J. W. *J. Phys. Chem. B* **115**, 4911–4926 (2011).