Decompositions of Free Energies in Molecular Simulation

Benedict William John Irwin

Department of Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Clare Hall

September 2018

Thank you to my wife Mingfei for her support.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Benedict William John Irwin
September 2018

Acknowledgements

I would like to acknowledge my supervisor Dr. David Huggins for his direction and supervision throughout the PhD. Also, Dr. Sinisa Vukovič who was a postdoctoral researcher in the Huggins lab during my PhD, Dr. P. L. Chau, based in the Pasteur Institute and Prof. Mike Payne, who offered much guidance and detailed proofreading, especially towards the end of the PhD. The following specific contributions have been made to this work:

- David Huggins had previously written a large repository of code called Solvaware which assists molecular simulation. This was used and improved by the Huggins lab during the PhD and facilitated the running of large calculations on Darwin/CSD3. He also input guidance and review work in his role as a PhD supervisor.

- Sinisa Vukovič and David Huggins worked on an algorithm for finding the best clusters of hydration sites around a protein according to desolvation free energies. This algorithm was applied to new systems during the PhD, namely protein-protein binding problems. The algorithm was also later applied on modifed datasets adapted during the PhD to take additional energetic contributions into account. Sinisa ran the original algorithm on many proteins which created a large dataset of hydration sites which the author could analyse.

- I started a collaboration with P. L. Chau during the PhD which led to a chapter 5 of this thesis. He introduced the problem of predicting the binding of GABARAP the the GABA-A receptor which he had previously modelled. He gave direction during this project and helped in the literature search, and provided me with candidate bound structures from the SwarmDock server.

- Thank you to Dr. Jamie Wynn for help with proofreading and to Prof. Ian Martin for sharing his knowledge on the $GABA_A$ receptor. I acknowledge financial support from the EPSRC Centre for Doctoral Training in Computational Methods for Materials Science under grant EP/L015552/1. This work was performed using the Darwin and CSD3 Supercomputers of the University of Cambridge HPC Service.

Abstract

This thesis describes advances in methods to measure free energy changes in simulations of molecular systems. In each case the free energy is decomposed into local environments which reveal insights about the complex systems being studied. Free energy is a fundamental quantity that can be used to predict whether changes in state are physically favourable. This can be used to predict the solubility of molecules and whether molecules are likely to bind to proteins. There are a handful of methods which measure free energy from molecular simulations. In chapter 3 we show results for an improved endpoint free energy method using inhomogeneous fluid solvation theory (IFST) which takes second order fluid-fluid entropy corrections into account. This is applied to a system of Lennard-Jones particles which show no measurable second order entropy contribution which fits with theoretical predictions. In chapter 4 an adaptation to the Zwanzig equation for path based exponential averaging methods is made. The equation is expanded to give contributions associated with every atom in the system. This method is called atomwise free energy perturbation and is applied to small molecules and ligand-protein binding. In chapter 5, IFST is applied to decompose hydration free energy at the surface of a protein into hydration sites. From these sites, information is inferred about the binding conformation of two proteins GABARAP and the GABA-A receptor. In chapter 6 statistics from hydration sites around hundreds of proteins are analysed. The distributions of free energy are shown and discussed for hydration sites in a range of local chemical environments. Also in chapter 6, the hydration sites decomposition method is augmented with local energy information associated with replacing a water molecule at a hydration site with a probe. The probe represents a ligand, and this is compared to the binding site prediction from the previous method. Further suggestions for improvements are made.

Table of contents

Nomenclature

Roman Symbols

ln	Natural logarithm (to save space)
log	Natural logarithm
$\log_{10}$	Base 10 logarithm
k_B	The Boltzmann constant

Greek Symbols

γ	Euler-Mascheroni constant $\simeq 0.577215\ldots$
$\Gamma(k)$	Euler gamma function with argument k
π	$\simeq 3.14159265359\ldots$
$\psi(k)$	Euler digamma function with argument k, $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$

Acronyms / Abbreviations

μM	Concentration: Micro-molar
.dcd	An MD trajectory file format
.pdb	An MD structure file format
11-mer, undecapeptide	Peptide: with 11 amino acids
13-mer, tridecapeptide	Peptide: with 13 amino acids
14-3-3	14-3-3 Protein
18-mer, octadecapeptide	Peptide: with 18 amino acids

23-mer, tricosapeptide	Peptide: with 23 amino acids
A, Ala	Amino Acid - Alanine
A2AR	Protein - Adenosine A2A receptor
ACE	Protein - Angiotensin-converting enzyme
AChE	Protein - Acetylcholinesterase
ADME	Absorption, Distribution, Metabolism, and Excretion
AFEP	Atomwise Free Energy Perturbation
C, Cys	Amino Acid - Cysteine
c-Abl	Protein - Mammalian Abelson Murine Leukemia Viral Oncogene Homolog 1
C1PE	Conditional 1 Particle Entropy
C2PE	Conditional 2 Particle Entropy
CDK2	Protein - Cyclin-dependent kinase 2
CTSK	Protein - Cathepsin K
D, Asp	Amino Acid - Aspartic Acid
E, Glu	Amino Acid - Glutamic Acid
EC_{50}	Half maximal effective concentration
EXP	Exponential Averaging
F, Phe	Amino Acid - Phenylalanine
FEP	Free Energy Perturbation
G, Gly	Amino Acid - Glycine
$GABA_A$-R	Protein - Gamma-aminobutyric acid type A receptor
GABARAP	Protein - Gamma-aminobutyric acid receptor-associated protein
GCT	Grid Cell Theory
GIST	Grid Inhomogeneous Solvation Theory

GST	Glutathione-S-transferase
H, His	Amino Acid - Histidine
HIV-1p	Protein - HIV-1 Protease
HS	Hydration Sites
I, Ile	Amino Acid - Isoleucine
ICE	Protein - Caspase-1/Interleukin-1 converting enzyme
IFST	Inhomogeneous Fluid Solvation Theory
K, Lys	Amino Acid - Lysine
L, Leu	Amino Acid - Leucine
M, Met	Amino Acid - Methionine
MD	Molecular Dynamics
mM	Concentration: Milli-molar
N, Asn	Amino Acid - Asparagine
NEU	Protein - Neuraminidase
P, Pro	Amino Acid - Proline
PDB	Protein Data Bank
PDE4D	Protein - Phosphodiesterase 4D
PDE5A	Protein - Phosphodiesterase 5A
Q, Gln	Amino Acid - Glutamine
R, Arg	Amino Acid - Arginine
RDF	Radial Distribution Function
S, Ser	Amino Acid - Serine
T, Thr	Amino Acid - Threonine
V, Val	Amino Acid - Valine

W, Trp	Amino Acid - Tryptophan
Y, Tyr	Amino Acid - Tyrosine

Chapter 1

Introduction

1.1 Summary of this Document

This thesis is titled ‘Decompositions of Free Energies in Molecular Simulation’. The underlying theme is working on novel methods to split up free energy changes into contributions which give insight into the system being studied. We specifically consider free energies measured from atomistic simulations of molecules and proteins using molecular dynamics (MD). This chapter will introduce the topics to be covered in the thesis and motivate the development of the techniques.

Chapter 2 will introduce MD, the types of systems simulated and a general overview of ‘path based’ free energy calculations which are used in chapters 3, 4 and 6. It will also introduce the theory of inhomogeneous fluid solvation theory (IFST) which is used in chapters 3, 5 and 6. Additional theory and background relevant to individual chapters will be covered in that chapter.

Chapter 3 discusses the breaking down of free energies of hydration or solvation. These are broken down into entropic and enthalphic contributions. It discusses advances in the calculation of entropic terms as a mutual information expansion which takes the form of an integral series. These developments are applied to the solvation of Lennard-Jones particles using free energy perturbation (FEP) calculations as references. The decomposition allows competitive calculation of free energy changes by only using the endpoints of the path used in FEP. This chapter will also discuss methods to speed up the calculation of the higher order entropy integrals by efficient k-nearest neighbours techniques.

Chapter 4 introduces a new and original technique developed during the PhD called atomwise free energy perturbation (AFEP). AFEP decomposes a free energy change across a molecule by assigning weights to each of the atoms. The weights are calculated from the energies associated with each of the atoms in the MD simulation. A full derivation of the method is given as a main result of this thesis. AFEP is then applied to small molecules and a full absolute binding affinity calculation for lopinavir binding to HIV-1 protease. Lopinavir is a molecule that is known to bind to this protein and is currently used in the treatment of HIV.

Chapter 5 introduces and advances a technique that has been developed and improved in the Huggins Lab during the course of the PhD. The hydration free energy around the surface of a protein is decomposed into hydration sites (HS), each representing the time average of a single water molecule. The method then takes connected clusters of these sites and finds the cluster with the lowest displacement free energy. This method was originally applied to predict binding hotspots for small drug-like molecules on proteins. This thesis shows the first application of this method to protein-protein binding, and the classification of protein surfaces. It is applied to the binding of the γ-aminobutyric acid receptor-associated protein (GABARAP) to the intracellular helices of the γ-aminobutyric acid type A receptor (GABA$_A$-R). A map of hydrophobic sites is made around both proteins using the decomposed hydration free energies. This map is found to highlight protein-protein binding activity with GABARAP and other proteins, along with dimerisation and trimerisations with itself.

Chapter 6 continues the application of IFST in the form of hydration sites. The algorithm was applied to 380 proteins generating hundreds of thousands of HS. Statistics are taken of the distributions of decomposed free energy associated with the HS around specific chemical motifs in the protein. This reveals the differences in the behaviour of water around different chemical groups in the protein. A rescaling scheme is developed for the original hydration patch finding algorithm as used in chapter 5. This rescaling attempts to combat a bias in the algorithm when judging hydration patches with highly charged motifs nearby. Further improvements are suggested which should lead to an improved tool for ligandability prediction.

Chapter 7 concludes and summarises the results of the thesis and discusses future related work.

1.2 Motivation and Background

1.2.1 Why Use Simulations?

Some of the hardest and most interesting problems left to study in physics and chemistry arise purely from complexity. For example, most biological problems involve proteins on some level. Proteins have very little 'order' in the sense that a crystal does, but on the other hand are extremely modular, with relatively easily encoded sequences of amino acids. One might be able to calculate some property of a crystal with a pen and paper by exploiting symmetry and periodicity. For the case of proteins, there is little hope and to study them we must use either experiments or *simulations*.

Simulations and experiments each have their pros and cons. The experiment (if performed correctly) will give the 'correct answer', up to an uncertainty tolerance. A simulation will only ever give an *approximation* to the correct answer, and will also have statistical uncertainties surrounding such values. Experiments can involve many copies of the same phenomenon, for example in a chemical reaction the experiment might sample many paths to complete the reaction simultaneously, whereas a single simulation will generally provide one such path, and may need to be repeated under different initial conditions until convergence. The concept of time is also different between simulation and experiment. There are different timescales associated with different phenomena being modelled. For a drug floating in solution, approaching a protein to find or be guided to the binding site and bind might take seconds of real time. Although this sounds short in real terms, the step length associated with integrating most models used in simulation might be of order femtoseconds. Many steps will be required to model the phenomenon via simulation, but the process is very quick in a lab.

However, simulations allow us to probe the system deeply and provide much more control than an experiment. High levels of automation and flexibility can often overcome the logistical problems associated with experiments. We can query individual atoms and bonds rapidly and even step outside the laws of physics if it is to our advantage. With this, simulation and experiment go hand in hand, offering together a deeper insight and understanding of problems.

1.2.2 The Problems to Be Studied

In the scope of this thesis, ‘the hardest and most interesting problems’, as referred to above, are those of drug discovery and to some extent the fundamental nature of liquids. Drug discovery is a field based around a problem: Finding a molecule or combination of molecules with therapeutic value for a given disease. If such a molecule (the drug) is to act as a medicine it must have adequate absorption, distribution, metabolism, and excretion (ADME) properties in the host, an appropriate level of toxicity, but also importantly must have some kind of favourable interaction with the biological mechanism that causes the disease. The magnitude of such a favourable interaction could be called efficacy. A simple example of an interaction could be that the drug inhibits a specific protein by binding to a specific binding site on that protein which somehow causes a favourable response in the wider scheme of things. Things are not always this simple, some proteins are very flexible and some diseases are caused by cascades of events that operate on length scales much larger than individual proteins.

The problem of drug discovery certainly has a degree of complexity. One of the ways of dealing with complexity is to break the problem down into smaller steps and handle each step individually. The protein itself is also a complex object, and this too can be broken down into units which are easier to understand, for example dealing with the ‘binding site‘ of the protein, which is comparatively smaller. The goal of this thesis is to work on a few methods which also break down complicated calculations into smaller and more manageable, meaningful or understandable parts.

The reason for referring to ‘the fundamental nature of liquids’ above is that all of the interactions between a drug and a protein are occurring in *solution*. For a biological system this means a lot of water, and potentially some ions to recreate the cellular environment. The water acts as a filler and due to its polar nature, naturally screens bare charges. Water will crowd charges, with molecules getting unusually close to each other near the charge, and more distal water molecules will then stay back from this cluster. This kind of information can be measured using radial distribution functions and higher order correlation functions which all describe the structure of water. Because water is dynamic around the protein, it is meaningful to measure the time averaged behaviour of water in various locations. This structure will have a direct impact on the biology and the chemistry of drugs near proteins and should be considered if any reasonable predictive power is desired. Decompositions of this information around the protein will lead to local understanding of this chemistry.

1.2.3 The Tools to Study the Problems

If we are to have a hope of understanding the complicated protein-drug-fluid system on the molecular level using simulation, we will need to have models of the water molecules, the ions, the drug molecule and the protein and all of the interactions between these components. This can amount to many thousands of atoms, so the model will need to be easily evaluated, such that it can be computed in a reasonable time. If we want to decompose the theory on top of this, we may need to keep track of many atoms throughout the simulation. The long time-scales and large system sizes associated with these kinds of problems generally rule out quantum mechanical methods which involve expensive calculations that scale unfavourably with the number of atoms (or electrons) in the system. Only very recently have linear scaling density functional theory codes been able to simulate proteins and it will still be some time before they can be used for this kind of work. At the cost of accuracy, classical models are much better suited for this kind of simulation. Specifically molecular dynamics (MD) with a classical, empirical force field is used in this thesis. Some details behind these methods will be explained in Chapter 2.

A simulation with a protein requires knowledge of the structure of the protein. Currently many protein structures are available online in the Protein Data Bank (PDB). Although most of these structures come from experimental data, there are a number of problems associated with this data. One is that to collect the data using X-ray diffraction, the proteins must be crystallised. Upon crystallisation, the state of the proteins is not the same as the state in solution and there will be some differences. Other than this, there may be damage, averaging errors and missing or unresolved parts of the protein. The resolution of some structures is not enough to distinguish between oxygen and nitrogen atoms, and the locations of hydrogen atoms are unknown. Any simulation involving such a structure will have been processed, generally to fill in the missing or ambiguous pieces with a best guess. This is another source of simulation uncertainty.

Throughout molecular simulations, the energy can be extracted from the parametrised force field, and from this properties associated with the energy can be calculated. For comparisons between states, the *free energy* of the states, and the change between the states is a fundamental and important quantity. The goal of this thesis is to understand complex phenomena which rely on free energy changes by *decomposing* the free energy changes. Decomposing is breaking them down into smaller parts, generally separating terms into sums of different terms. The decomposition can also refer to distributing the free energy in a spatial manner, either around the surface of a protein, or across the atoms of a molecule, or around the simulation cell. Through these decompositions more can be learned about local parts of a large and complex system, each of the parts is less complex and this can facilitate understanding. The background of free energy changes is also covered in chapter 2 and the decomposition methods will be introduced in chapter 2 and covered in more detail individually throughout the chapters.

Chapter 2

Theory, Molecular Dynamics and Free Energy Calculations

In the previous chapter the scene was set for the concepts that will be important throughout the thesis. This chapter will describe these concepts in further detail, often with equations and references to the literature. The chapter introduces molecular dynamics (MD) and free energy methods which are the cornerstone of data collection and methods validation in the thesis. This section will also describe inhomogeneous fluid solvation theory (IFST), and the associated entropy calculations.

2.1 Ligand Binding

One of the goals of this thesis is to calculate decompositions of the binding free energies associated with drugs binding to proteins. The exact methods used to perform the decompositions will be explained in detail in subsequent chapters. However, it is worth explaining what these binding free energies are, why these binding free energies are important and how they are calculated. The usual way to quantify the binding of a drug is by a rate constant. If we have a solution containing a protein P and a ligand L and their complex PL, then there will be a reversible 'equation' of the form

$$P + L + \text{solution} \rightleftharpoons PL + \text{solution}', \tag{2.1}$$

where the $'$ indicates that the solution of water and ions surrounding the protein and ligand has changed or reconstructed around the new complex. Such a reversible reaction will have forward and backward *rates* denoted k_{forward} and k_{backward}. If the forward rate is large and the backward rate is small then one will expect to find many bound complexes. The affinity between the ligand L and the protein P, is then described by K_{d} and K_{a}, which are called the dissociation and association constants respectively. These are given by

$$K_{\text{d}} = \frac{1}{K_{\text{a}}} = \frac{[L][P]}{[LP]} = \frac{k_{\text{backward}}}{k_{\text{forward}}}, \tag{2.2}$$

where $[\cdot]$ is the molar concentration of the species $\cdot$ in the bracket. Large affinity means a large association constant K_{a}. Many research articles and online databases quote pK_{d} scores for protein-ligand combinations, where the p means $-\log_{10}$, i.e. p$K_{\text{d}} = -\log_{10} K_{\text{d}}$. This is an easier number to process because concentrations span many orders of magnitude. It also relates directly to changes in the standard state *Gibbs free energy G* by

$$\Delta G = -RT \log K_{\text{a}} = RT \log K_{\text{d}}, \tag{2.3}$$

(here log is the natural log, and will be for the rest of the thesis).

Equation 2.3 shows that the rate of dissociation or association, relies on a free energy change. This means that the binding affinity of a ligand and protein relies on the free energy change associated with the binding ΔG_{bind}. If we can then simulate the binding using a suitable tool, and calculate the free energy from that simulation, we can begin to estimate the affinity of the ligand binding to the protein.

2.1.1 Hydration Free Energy

As well as binding free energies there are also solvation free energies which describe the free energy change between a bulk solvent and the solvent plus some solute. If the solvent is water then this is called the hydration free energy. Sometimes it makes more sense to think about the inverse process and desolvation will be used to denote the inverse of solvation which changes the sign of the free energy contribution. These quantities are of interest in biological and pharmaceutical contexts because proteins and drugs in the body are surrounded by a solution. If a drug transitions from the solution into the protein there will usually be a desolvation penalty associated with breaking hydrogen bonds between the solvent and the drug molecule. Some drugs are quite hydrophobic and this penalty will be smaller. This information is contained in the hydration free energy.

2.2 An Overview of Molecular Dynamics

Molecular dynamics is a well established simulation technique with a long history [1]. The premise is that all atoms exist as classical particles in a simulation cell arranged in the desired initial configuration which represents the system under study. A potential energy function $U(\mathbf{x})$ (usually called a force field) is defined for the desired system, it takes in the positions of the particles and outputs a scalar energy, which is the system potential energy. The particles are given initial velocities according to a Maxwell-Boltzmann distribution defining the temperature T of the system. At this step, the forces $\mathbf{F}_i$ on atom i can be calculated by taking the gradient of the potential energy function in 3-dimensions,

$$m_i \frac{d^2\mathbf{x}_i}{dt^2} = \mathbf{F}_i = -\nabla U(\mathbf{x}_i), \tag{2.4}$$

here m_i are the masses of the particles. Using this formalism the system of particles is integrated forward through time by updating the positions according to the velocities and the velocities according to the accelerations. Molecular dynamics simulations are carried out in a specified thermodynamic ensemble, examples being the canonical (NVT) ensemble where there is a constant number of particles N, constant volume V and temperature T and the isothermal-isobaric (NpT) ensemble where there is a constant number of particles N, constant pressure p and constant temperature T. Simulations in this thesis were either run in the NVT or NpT ensembles depending on the situation.

2.2.1 Force field Parameters

A force field encodes all of the interactions between atoms in an MD simulation. Although one could define just about any force field, by either equations or numerical data, there are a number of standard terms used in many force field parametrisations, including the CHARMM force field used in this thesis. The functional forms for these terms have simple interpretations and a long history. These terms are defined between pairs, triplets and quadruplets of atoms:

$$U_{\text{elec}}(\mathbf{x}_1,\mathbf{x}_2) = C_{\text{Coulomb}}\frac{q_1 q_2}{\eta||\mathbf{x}_1-\mathbf{x}_2||}, \tag{2.5}$$

$$U_{\text{LJ}}(\mathbf{x}_1,\mathbf{x}_2) = -\sqrt{\varepsilon_1\varepsilon_2}\left(\frac{r_{\text{min}}^{12}}{||\mathbf{x}_1-\mathbf{x}_2||^{12}} - 2\frac{r_{\text{min}}^{6}}{||\mathbf{x}_1-\mathbf{x}_2||^{6}}\right), \tag{2.6}$$

$$U_{\text{bond}}(\mathbf{x}_1,\mathbf{x}_2) = k_{\text{bond}}(||\mathbf{x}_1-\mathbf{x}_2|| - r_{\text{eq}})^2, \tag{2.7}$$

$$U_{\text{angle}}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = k_{\text{angle}}(\text{ang}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) - \theta_0)^2, \tag{2.8}$$

$$U_{\text{UB}}(\mathbf{x}_1,\mathbf{x}_3) = k_{\text{ub}}(||\mathbf{x}_1-\mathbf{x}_3|| - r_{\text{ub}})^2, \tag{2.9}$$

$$U_{\text{dihedral}}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4) = k_{\text{di}}(1+\cos(n\,\text{dih}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4)+\phi_0)), \tag{2.10}$$

$$U_{\text{improper}}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4) = k_{\text{imp}}(\text{imp}(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3,\mathbf{x}_4) - \phi_0)^2, \tag{2.11}$$

$$\text{for } \mathbf{x} = (x_1,x_2,x_3), \qquad ||\mathbf{x}-\mathbf{y}|| = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2}. \tag{2.12}$$

The parameters used in these equations are described in table 2.1 which is located at the end of this chapter. Here η is the dielectric constant, this is set to 1, except when water is not explicitly modelled in the simulation. For water at room temperature it is usually set to 80. All simulations in this thesis use explicit water, but some energy evaluations are made in Chapter 6 which use $\eta = 80$. The functions $\text{ang}(x_1,x_2,x_3)$, $\text{dih}(x_1,x_2,x_3,x_4)$ and $\text{imp}(x_1,x_2,x_3,x_4)$ give the angle between three atoms, and dihedral and improper angles between four atoms respectively.

2.2.2 Integrator Parameters

There are a number of advanced schemes in place to regulate the stability of an MD simulation. For the NVT simulation the temperature must be held constant. This involves using a 'thermostat'. For NpT, both pressure and temperature are held constant, this involves using both a thermostat and a barostat. These are specified as input options into the molecular dynamics code.

There are a number of parameters associated with the pressure and temperature control. The MD software NAMD which is used for all simulations in this thesis uses a Langevin piston dynamics assisted Nosé-Hoover thermostat or barostat [2, 3]. The system is connected to a heat-bath of noise at the desired system temperature and this allows heat to flow in or out of the system in a controlled manner. There are parameters associated with this control including target temperature, target pressure, damping period and oscillatory period. The time constants are usually a few hundred femtoseconds.

2.2.3 Equilibration and Good Practice

A simulation must be equilibrated before data are collected. This involves running the dynamics and allowing the temperature, pressure, volume and energy of the system to stabilise, depending on the ensemble. The density of the system is also important, and a reasonable starting cell should be picked to facilitate easy simulation. After this an NpT equilibration will usually be run to allow the simulation cell to find an appropriate volume.

If the coordinates of the system are imported from an experimental structure or simulation using a different simulation method (e.g. density functional theory), or even a different force field, it is unlikely that the bonds/angles/distances between atoms are at their equilibrium values. An energy minimisation can first be performed on the structure. This involves moving all degrees of freedom against the gradient in energy, this is performed by calculating the derivative in energy with respect to all parameters (i.e. the Jacobian).

2.2.4 Files, Inputs, Outputs and Workflow

The inputs to a molecular dynamics simulation will include a configuration file, which specifies all of the parameters and settings and the locations of the input files. The input files will either be the starting system or if the simulation is being continued from a previous simulation, the initial coordinates and velocities of all atoms. The output will contain the energy, volume, pressure and temperature of the system sampled over time and the coordinates of all of the atoms arranged in ‘snapshots’. The list of snapshots can be called a trajectory, and example file format for this .dcd which is used by the NAMD simulation software. Many types of post processing analysis can be done on these trajectory files including IFST entropy calculations, and the clustering algorithms for defining hydration sites which will be used later in the thesis.

2.2.5 Using MD for Calculations

MD in general is appropriate for simulating many types of system. This could range from bulk fluids, molecules in a solvent, peptides or large proteins or solid crystals and polymer lattices. As described above, each atom will be given parameters including mass, charge and repulsion/Van der Waals (VDW) parameters. Bonds are defined and also parametrised, angular interactions between three atoms and dihedral interactions between four atoms are also added. If a suitable set of parameters are chosen then the simulation will be a good approximation of the real physical system.

The MD force field described in section 2.2.1 is relatively simple and classical in nature, so may fail to capture all of the details of systems with complicated interactions. Some of the harder parts to capture arise from quantum mechanical phenomena and MD does not usually allow for details such as electron/proton transfer. The main advantage of MD on modern computers is speed, and system size. MD is suitable for modelling biological molecules including very large proteins which can contain thousands of atoms. The methods developed in this thesis relate to calculations involving proteins, and therefore MD is the primary simulation method used in this thesis.

There are many implementations of MD available. The main differences are the force fields used and some of the methods implemented in the MD packages. Some force fields are more suitable for modelling crystals and inorganic systems for example the 'Large-scale Atomic/Molecular Massively Parallel Simulator' (LAMMPS) package [4]. Others specialise in biological systems such as CHARMM [5], GROMACS [6], NAMD [7], AMBER [8] and OpenMM [9]. The predicted free energies across software packages are quite consistent, with alchemical hydration free energies reproduced to within 0.2 kcal/mol amongst major codes (AMBER, CHARMM, GROMACS, SOMD) [10]. 'Nanoscale Molecular Dynamics' (NAMD) is the package used for all simulations in this thesis. This implements the Chemistry at Harvard Macromolecular Mechanics (CHARMM) force field which is broken into parts parametrising the protein atoms [11], the solvent atoms using one of many water models and atoms commonly found in small organic molecules using a generalised force field [12].

The output energies, pressures, volumes and temperatures of the MD simulation are functions of time, even if being held constant. There will be fluctuations in parameters controlled by the barostats or thermostats metioned in section 2.2.2 because the control is a dynamic process. Time is parametrised by a step length which reflects a small shift in system time. This step length is usually around 2 femtoseconds. Smaller time-steps can be used for unstable systems, particularly if the input coordinates are not well adjusted to the MD force field parametrisation which is common for structures derived from quantum mechanical simulations or taken directly from experimental data (i.e. nuclear magnetic resonance (NMR) or X-ray crystallography data). Forces, velocities and positions of particles are also output as a function of time. These quantities are not usually output every step because neighbouring data points or frames in the output will be highly correlated and the added information content is low. Also, writing to disk often will affect the speed performance of the simulation, and writing many frames will create very large files over long trajectories of large systems. The goal of MD is to sample the states the system visits in an unbiased manner that reflects the equilibrium state of the system. Because of this samples may be saved around every 2 picoseconds (the exact value will be system dependent).

Free energy is not amongst the output quantities described above. Calculating free energy changes from MD simulations requires further techniques that combine these outputs. Free energy and entropy are thermodynamic variables that are not attached to a given state, but to an ensemble of states. This explains the need for a simulation which adequately samples this ensemble.

2.3 Free Energy Calculations

First it will be useful to define free energy. Free energy is the energy in a system which could theoretically produce work. The quantity that describes this notion will depend on the thermodynamic ensemble the system is being viewed in, i.e. which variables are being held constant. For the two ensembles considered in this work, NVT and NpT there are two free energies, the Helmholtz free energy, A for NVT, and the Gibbs free energy G for NpT. We have

$$A = U - TS \tag{2.13}$$

and

$$G = U + pV - TS = H - TS \tag{2.14}$$

where U is internal energy, p pressure, V volume, T temperature, S entropy and H enthalpy.

When calculating ensemble properties it is often easier to calculate a change rather than the total value [13]. The free energy calculations referred to in this work will calculate the free energy difference between two systems A and B. At constant T, we have

$$\Delta A = (U_A - U_B) - T(S_A - S_B) = \Delta U - T\Delta S, \tag{2.15}$$

with constant T and p we have

$$\Delta G = (U_A - U_B) + p(V_A - V_B) - T(S_A - S_B) = \Delta H - T\Delta S, \tag{2.16}$$

this already reveals a decomposition of sorts; changes in free energy can be decomposed into an energetic or enthalpic term and an entropic term. This concept will be useful in inhomogeneous fluid solvation theory (IFST) which is described later in this chapter. The concept is also used in grid inhomogeneous solvation theory (GIST) which is similar to IFST and the Grid Cell Theory (GCT) method [14]. There are other methods for calculating free energies such as nested sampling, umbrella sampling and potential of mean force methods. These will not be covered in the thesis.

2.3.1 Free Energy Perturbation

Free energy perturbation (FEP) is a method for measuring free energy differences from simulation. Unfortunately, free energy perturbation can be a confusing term in the literature because different authors have used it to refer to different concepts. It can generally refer to a class of methods which approximate a free energy difference from energetic differences across a path. The name comes from the traversal of this path, which often contains neighbouring states, each very similar to the last such that they are approximately the last state with an added perturbation. If FEP is used to refer to a class of methods, these methods include exponential averaging (EXP), Bennett acceptance ratio (BAR) and thermodynamic integration (TI) which will be described below. Some authors use the term FEP to refer to the EXP or BAR methods directly.

As mentioned above, these methods rely on a parameter $\lambda \in [0,1]$ (sometimes called the coupling parameter), which defines a path between the initial and final systems. If $\lambda = 0$ we are in the initial system, if $\lambda = 1$ we are in the final system. This concept will be shown in more detail in chapters 3 and 4 where these so called 'lambda schedules' are used.

2.3.2 Exponential Averaging

The fundamental equation for the exponential averaging (EXP) method, is the so-called Zwanzig equation [15] for the Helmholtz free energy difference ΔF_{AB} between two thermodynamic states A and B

$$\Delta F_{AB} = F_B - F_A = -\beta^{-1} \log \left\langle e^{-\beta(U_B - U_A)} \right\rangle_A, \tag{2.17}$$

where $\beta = (k_B T)^{-1}$ and where the notation $\langle X \rangle_A$ denotes the ensemble average of X, over system A, which can be written

$$\langle X \rangle = \frac{1}{Q_A} \int X e^{-\beta U_A(\vec{q})}\, d\vec{q} \tag{2.18}$$

where Q_A is the partition function of system A

$$Q_A = \int e^{-\beta U_A(\vec{q})}\, d\vec{q} \tag{2.19}$$

here the energy of the system A is parametrised by a coordinate vector $\vec{q}$, which corresponds to all the degrees of freedom in the system, i.e. the atomic positions. Equation 2.17 is a statement that the change in free energy between the two states can be calculated from the energies only if a suitable averaging process is made. The free energy has an energetic/enthalphic part and an entropic part. The entropic part can be linked to the number of different states the system can access from fundamental thermodynamics. In the case of equation 2.17, this part is swept away into the ensemble average, it appears to be a formula which takes the energy as an input and gives out a free energy. It should be remembered that the complexity of calculating the entropy has now been converted to a sampling problem, if the simulation is not run for a sufficient length of time to sample the state(s) accurately the prediction will be wrong [13].

It should also be noted that if systems A and B are *very* different, the sampling will take a long time to converge. This is normally overcome by inserting lambda windows between the two states, and equation 2.17 is applied between each contiguous pair of windows. This is where the λ parameter mentioned above is used.

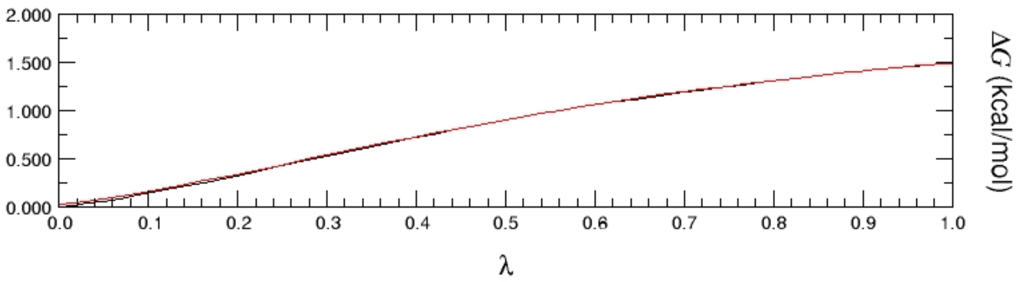


Fig. 2.1 An example of a well converged forward and backward calculation of a free energy change through the ParseFEP plugin for VMD [17].

2.3.3 Bennett Acceptance Ratio

A more sophisticated version of the EXP method is the Bennett Acceptance Ratio (BAR) method [16]. In the EXP algorithm a state average was taken with respect to A, we moved from state A to B, but due to reversibility, we could have also moved from B to A. This can be called forward/backward sampling, and BAR offers a way of combining the two results simultaneously for a more accurate prediction of the free energy change. BAR is implemented in NAMD [17] and was used for all FEP style free energy calculations in the thesis.

Figures 2.1 and 2.2 show the free energy change of an example FEP calculation as a function of lambda, and the associated energy sampling probability distributions for some of the lambda windows. In each of these curves there is a forward (black) and backward (red) value. In this examples the forward and backward curves are strongly overlapping, which implies the simulations were converged and effective energy sampling has taken place. Inspection of such plots is necessary but not sufficient to check for effective sampling. There is no way to guarantee the simulation has sampled the representative space.

2.3.4 Thermodynamic Integration

A related and subtly different method is thermodynamic integration (TI), in which the inverse functions exp and log in equation 2.17 are replaced with inverse operators $\int d\lambda$ and ∂_λ. EXP dealt with differences in energy, and TI deals with *differentials* of energy by writing the change in free energy as

$$\Delta F = \int_0^1 \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{2.20}$$

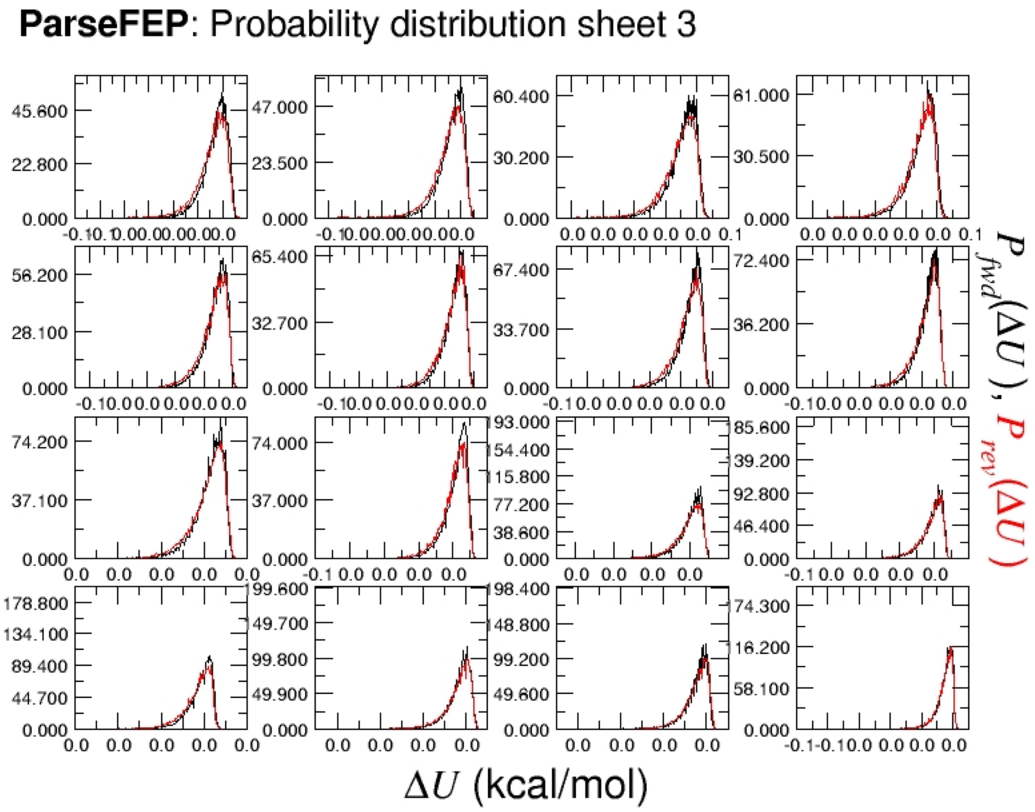


Fig. 2.2 ParseFEP gives out the probability distribution for system energies for each lambda window [17]. Here the forward and backward distributions are almost overlapping which indicates convergence of the calculation.

where the energy is now a function of the parameter λ. Often the energy is defined as a function of λ, and the integration and ensemble averaging is computed numerically from multiple simulations sampling different values of λ. In Chapter 4 TI is used to calculate the free energy associated with restraints on a ligand in a binding site.

2.3.5 Direct Calculation of Free Energy

There are also methods which directly calculate the changes in energy and entropy that together make the change in free energy. One such method is IFST, which will be covered in detail below. From molecular dynamics, energies and enthalpies can be calculated directly. The calculation of entropies requires sampling an ensemble of states that the system could occupy.

2.4 Entropy and Information Theory Concepts

In the IFST method changes in entropy will be calculated directly by considering the distribution of atoms in the solvent. To calculate entropy changes directly some consideration should be made of the wider definition of entropy. One of the best terms to describe entropy is *missing information.* This interpretation crosses all domains, and generally works in both information theory and in thermodynamics. Entropy is essentially a statistical concept and can be viewed in the broadest sense as a functional which acts on a probability distribution to give a scalar quantity.

The discrete formalism will give an information entropy H_{disc} (note that H is also used for information entropy as well as enthalpy)

$$H_{\text{disc}} = -\sum_i p_i \log p_i \tag{2.21}$$

which obeys a number of nice properties, one of which is $H_{\text{disc}} \geq 0$. This can be viewed as the expectation value of $-\log p$, which is written $\mathbb{E}[-\log p]$. Here p_i is a probability distribution. This resembles the Gibbs entropy from thermodynamics

$$S = -k_B \sum_i p_i \log p_i \tag{2.22}$$

In these terms the physical variants of each type of entropy with either have a factor of k_B or R, the gas constant, in front depending on the units of the expression. k_B gives the entropy per particle, and R the entropy per mole.

The continuous analogue of $\mathbb{E}[-\log P(x)]$ is a differential information entropy H_{cont} for a distribution $P(x)$ across support S

$$H_{\text{cont}} = -\int_S P(x) \log P(x)\, dx \tag{2.23}$$

and does not always follow the rules of the discrete case, this continuous entropy can be negative. Also, continuous probability distributions can have values greater than 1; if a particle was uniformly distributed between 0 and $1/2$, $P(x) = 2$ across the support.

2.4.1 Types of Entropy

There are various types and extensions to the concept of entropy which will be covered briefly.

Joint Entropy

For two random variables X and Y the joint entropy is denoted $H(X,Y)$, this is simply related to the pair probability density function $P(x,y)$ by

$$H(X,Y) = -\iint_S P(x,y)\log P(x,y)\ dxdy \tag{2.24}$$

this concept extends to a distribution with a vector of random variables

$$H(\mathbf{X}) = -\int_S P(\mathbf{x})\log P(\mathbf{x})\ d\mathbf{x} \tag{2.25}$$

Every distribution will have an entropy of this type. This entropy will work for the probability distribution associated with an n-body system. The distribution with zero entropy is a delta function, as there is no missing information.

Conditional Entropy

Denoted $H(Y|X)$ this is the entropy associated with Y given X, which can be phrased as the joint entropy minus the marginal entropy

$$H(Y|X) = H(X,Y) - H(X) = \iint P(x,y)\log\left(\frac{P(x)}{P(x,y)}\right)\ dxdy \tag{2.26}$$

in general, a multi-dimensional entropy can be expressed as a sum of partial conditional terms [18]

$$H(X_1,\cdots,X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1,X_2) + \cdots = \sum_{i=1}^{n} H(X_i|X_1,\cdots,X_{i-1}) \tag{2.27}$$

This expansion will be used in IFST to break a high dimensional integral like $H(X_1,\cdots,X_n)$ into a series of lower dimensional terms. The origin is the generalised product rule for multivariate probability distributions

$$P(X_1,\cdots,X_n) = P(X_1)P(X_2|X_1)\cdots P(X_n|X_1,\cdots,X_{n-1}) \tag{2.28}$$

Mutual Information

Denoted $I(X,Y)$, this is defined by the discrete and continuous expressions

$$I(X;Y) = \sum_x \sum_y P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right) \tag{2.29}$$

$$I(X;Y) = \iint P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)\, dxdy \tag{2.30}$$

In the context of IFST, the expansion relating to equation 2.27 is often called a mutual information expansion. The expansion can be phrased either in terms of conditional entropies or in terms of the mutual information defined above. There are identities relating the quantities between entropies, joint entropies and conditional entropies:

$$I(X;Y) = H(X) - H(X|Y) \tag{2.31}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{2.32}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2.33}$$

$$I(X;Y) = H(X,Y) - H(X|Y) - H(Y|X) \tag{2.34}$$

The entropy can be seen as the mutual information of a variable with itself, $I(X,X) = H(X)$. In the following section and chapter 3 the entropy of a fluid will be calculated. There are analogous integrals in this theory with the types of entropy shown above. Most of the identities shown arise from the transformation properties of logarithms and the linearity of the integration operator. Mutual information also has conditional and multivariate types which will not be covered here.

2.5 Inhomogeneous Fluid Solvation Theory

IFST was created from the theory of traditional homogeneous fluids. The idea was to introduce inhomogeneity via a field around a static solute and to express key quantities (energy and entropy) as integrals over the correlation functions of the fluid. As mentioned before the free energy can be split into an entropic and energetic part

$$\Delta G_{\mathrm{IFST}} = \Delta E_{\mathrm{IFST}} - T\Delta S_{\mathrm{IFST}}. \tag{2.35}$$

each of these quantities is a functional of the various correlation functions of the fluid.

2.5.1 Radial Distribution Functions and Correlation Functions

Much of the structure of a simple liquid is described by the pair correlation function. This is often called the radial distribution function (RDF), often written simply as $g(r)$ and some examples will be shown in the next chapter for liquid neon. This function describes the probability of finding a fluid molecule at a distance r from another fluid molecule. This description works well for spherically symmetric homogeneous fluids, for example a monatomic gas. If there are different species of atom present there will be correlation functions describing each pair of types of atom. There also exist higher order analogues of the RDF, for example triplet and quadruplet distance correlation functions. A triplet distance correlation function might be written $g(r_1, r_2)$ and describes the probability of finding a particle at r_2, given that there are particles at the origin and at r_1. As the number of degrees of freedom increase there are also orientational correlation functions to consider. The orientation correlation functions are parametrised by angles between specific atoms and will follow the symmetry of the molecules in the fluid. For a system with liquid water 5 angles are used to parametrise the orientation between just two molecules [19]. This can be written in shorthand as $\omega^2 = (\theta_1, \theta_2, \phi, \chi_1, \chi_2)$, where ω^2 means the orientational variables between two molecules. Here, each of the θ are angles between the dipole vector and the intermolecular axis between the pair of molecules, the χ are the rotational angles of the molecules around their own dipole vector and ϕ is the dihedral angle between the planes defined by the two dipole vectors.

It can be seem that this kind of treatment quickly gets complicated, and a method to keep track of the terms involved is greatly desired. The main modern developments are given by Lazaridis [20, 21] and Karplus [22]. This treatment is more recently discussed by Wang et al. [19]. All of these treatments use the generalized Kirkwood superposition approximation (KSA) which allows higher order densities to be written as ratios of lower order densities in a recursion relation. To generalise this treatment the n-body correlation functions of both distance and orientation the functions are written as $g^{(n)}(\mathbf{r}^n, \omega^n)$. These functions are extremely complex, as can be approximated using the pair distribution (including both distance and orientation) $g^{(2)}(\mathbf{r}^2,\omega^2)$, and the marginal parts of the triplet and quadruplet distributions (and so on): $\delta g^{(3)}(\mathbf{r}^3,\omega^3)$ and $\delta g^{(4)}(\mathbf{r}^4,\omega^4)$. Here the δ symbolises the marginal distribution, such that $\delta g^{(3)}$ is the part of $g^{(3)}$ which is not described by $g^{(2)}$ or combinations thereof. The first of the KSA relations is simply

$$g^{(3)}(\mathbf{r}^3,\omega^3) = g^{(2)}(1,2)g^{(2)}(2,3)g^{(2)}(3,1)\delta g^{(3)}(r^3,\omega^3) \tag{2.36}$$

where more shorthand is introduced; $g^{(2)}(1,2)$ is a pair correlation function, but references atoms 1 and 2 of the three atoms in the $\mathbf{r}^3$ and ω^3 variables. It is likely that the functions $\delta g^{(n)}(\mathbf{r}^n,\omega^n)$ approach the value of 1 fairly quickly as n grows above 4 [23]. However, there is recent evidence that this kind of expansion expansion does not work in systems with long range correlation which has called this broad assumption into question with studies that test the MIE for hundreds of Gaussian distributed variables [24]. Specifically, the correlations must decay exponentially, otherwise specific orders of truncation exist to minimise the error form the expansion method. It is expected that the correlations will decay relatively quickly in molecular systems due to the short range nature of repulsion forces and charge screening type effects for the more long range Coulomb terms in the molecular forcefield.

It can be seen from equation 2.36 that an entropy like integrand

$$g^{(3)}(\mathbf{r}^3,\omega^3)\log\left(\delta g^{(3)}(\mathbf{r}^3,\omega^3)\right) = g^{(3)}(\mathbf{r}^3,\omega^3)\log\left(\frac{g^{(3)}(\mathbf{r}^3,\omega^3)}{g^{(2)}(1,2)g^{(2)}(2,3)g^{(2)}(3,1)}\right) \tag{2.37}$$

looks like the continuous mutual information in equation 2.30. The right hand side has the logarithm of the full three body distribution divided by the marginal distributions.

2.5.2 Homogeneous Fluid

Historically, the theory for homogeneous fluids was extensively studied by Kirkwood [25, 26], Nettleton [27], Raveche [28], and Wallace [29–31]. Major achievements during that period were finding ensemble invariant forms for the resulting series. According to a more modern summary by Lazaridis and Karplus [22] this expansion allows the entropy of a molecular fluid to be written (in the canonical and grand canonical ensembles) as a sum of terms $S_{\text{fluid}} = s_1 - s_2 - s_3 - \cdots$ given by

$$\begin{aligned} s = & s_{\text{id}} - \frac{k_B \rho}{2\Omega^2} \int [g^{(2)} \log g^{(2)} - g^{(2)} + 1] \, d\mathbf{r} d\omega^2 \\ & - \frac{k_B \rho^2}{6\Omega^3} \int [g^{(3)} \log \delta g^{(3)} - g^{(3)} + 3g^{(2)} g^{(2)} - 3g^{(2)} + 1] \, d\mathbf{r}^2 d\omega^3 - \cdots \end{aligned} \tag{2.38}$$

where s_{id} is the entropy of an ideal gas, Ω is the integration of the rotational parameters of one molecule for simple molecules this will be $4\pi^2$ for molecules with symmetries an additional discount will be included. ρ is the average number density of the fluid. One can notice the $g^{(3)} \log \delta g^{(3)}$ term and see how this expansion could be connected to mutual information. In this case the energy can be expressed as

$$e = \frac{3k_B T}{2} + \frac{\rho}{2} \int g^{(2)}(r) u(r) \, dr \tag{2.39}$$

where r is the relative position and $u(r)$ is the pair interaction potential which is defined to capture all of the interaction energy in this way.

2.5.3 Binary Solution

Once a solute is introduced concentration distribution functions need to be created for three kinds of interaction; those between solvent atoms, those between solute atoms and those between the two types. The main difference is the addition of labels s or w denoting the solute and 'water' respectively. This leads to a threefold increase in terms in the number of expressions for entropy and energy in the above section. The most useful step after this stage is to take the concentration of solute atoms to be very small, approaching infinite dilution. The resulting expressions for entropy and energy are given by Lazaridis [20]. These form an intermediate step and will not be written here.

2.5.4 Inhomogeneous Fluid

The next development is to take the binary homogeneous fluid to the inhomogeneous reference picture. For the inhomogeneous fluid, there is a single fixed solute included in the centre of the box at an infinite dilution. This solute molecule creates a 'field', or external potential which leads to a reconstruction of the fluid around it. Morita and Hiroike solved the full inhomogeneous fluid expansion in terms of a terse but somewhat arcane diagrammatic series [32]. Lazaridis gives expressions for the resulting series truncated to interactions between the solute and water (sw) and the water and water (ww) level [20]. The final version of the theory used in the thesis is described in the next section.

2.5.5 Local Treatment

The IFST treatment can also be split into local regions of volume. This works when the original series is truncated to the sw and ww level [33–35]. The two energy and entropy terms are broken simply into respective solute-water and water-water terms, this gives the energy

$$\Delta E_{\text{IFST}} = E_{\text{sw}} + \Delta E_{\text{ww}} \tag{2.40}$$

and the entropy

$$\Delta S_{\text{IFST}} = S_{\text{sw}} + \Delta S_{\text{ww}} \tag{2.41}$$

we have

$$\Delta S_{\text{sw}} = -R\rho \int g_{\text{sw}} \log g_{\text{sw}} - g_{\text{sw}} + 1 \, dw \tag{2.42}$$

where dw is the volume element for one solvent molecule integrated across the entire simulation cell. Formally the next term is given by

$$\Delta S_{\text{ww}} = -\frac{R\rho}{2} \iint g_{\text{sw}} g_{\text{sw}'} \left[g_{\text{ww}'} \log g_{\text{ww}'} - g_{\text{ww}'} + 1 \right] dw dw' \tag{2.43}$$

where w and w' are the positions of different solvent molecules. Using the generalised Kirkwood approximation which is the same as equation 2.36 with the new subscript notation

$$g_{\text{sww}'} = g_{\text{sw}} g_{\text{sw}'} g_{\text{ww}'} \delta g_{\text{sww}'} \tag{2.44}$$

and assuming $\delta g_{\text{sww}'} = 1$ allows this to be written as

$$\Delta S_{\text{ww}} = -\frac{R\rho}{2} \iint g_{\text{sww}'} \log \left(\frac{g_{\text{sww}'}}{g_{\text{sw}} g_{\text{sw}'}} \right) dw dw' - \frac{R\rho}{2} \iint g_{\text{sw}} g_{\text{sw}'} \left[1 - g_{\text{ww}'} \right] dw dw' \tag{2.45}$$

the first integral is a mutual information term, the second is a volume exclusion term.

2.5.6 Non-local Contributions

The hydration entropy is expressed in local and non-local terms

$$\Delta S_{\text{hydration}} = \Delta S_{\text{local}} + \Delta S_{\text{nonlocal}} \approx \Delta S_{\text{IFST}} - R \tag{2.46}$$

the local terms are the fluid expansion terms as covered in section 2.5.4, the non-local terms are additional terms that affect the whole simulation cell. This includes entropy from volume expansion ΔS_{ve} and a liberation entropy ΔS_{lib} which is associated with fixing the solute in the cell which would normally be free to wander the whole volume of the cell. We can express this as

$$\Delta S_{\text{nonlocal}} = \Delta S_{\text{ve}} + \Delta S_{\text{lib}} = R(\alpha T - 1) + R\rho \int [1 - g_{\text{sw}}]\, dV \tag{2.47}$$

where α is the thermal expansion coefficient of the solvent. The change in volume contributes an entropy

$$\Delta S_{\text{ve}} = -RT \log\left(\frac{V_1}{V_2}\right) \tag{2.48}$$

the theory for an extension to IFST is derived in the next chapter in more detail. This treats the solvent-solvent terms in more detail and provides details on the summation methods used to calculate the integrals in the above equations.

2.5.7 Summation

The principle behind the summation methods is that the entropy is an expectation value of a density function $\rho(x)$

$$H[\rho(x)] = \mathbb{E}[-\log \rho(x)] \tag{2.49}$$

this can be approximated by sampling

$$H[\rho] \approx \frac{1}{N}\sum_{i=1}^{N} -\log(\rho_i) = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{N}{V_i}\right) \tag{2.50}$$

here the density sample is treated as a number per unit volume $\rho_i = N/V_i$. If each particle has a local density value associated to it, N is constant and the variation arises by varying the volume available to that particle. This can be phrased as a kernel function of the distance to the nearest neighbour. If this kernel is chosen as a 3-ball this volume is

$$V_i = \frac{4}{3}\pi r_i^3 \tag{2.51}$$

where r_i is the distance to the nearest neighbour of particle i. This is a ball in 3 dimensions but in the case of higher order correlation functions the density is a higher dimensional object. The same spherical notion can be applied to n-dimensional spaces using the volume of the n-ball

$$V_{in}(r) = \frac{\pi^{n/2} r_i^n}{\Gamma(\frac{n}{2}+1)} \tag{2.52}$$

where $\Gamma(n+1)$ is the Euler gamma function, which interpolates the factorial function $n!$ for non-integer values. This gives

$$H[\rho] \approx \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{N\pi^{n/2} r_i^n}{\Gamma(\frac{n}{2}+1)}\right) \tag{2.53}$$

which will be used in the next chapter to build estimators for the entropy.

Some Limitations of the Estimators

It should be noted that if r_i is reported as 0, then the logarithm will diverge. This would be equivalent to a nearest neighbour distance that is exactly the same as the current sample point. for this reason, the above estimator will struggle to measure the entropy of very dense or solid systems. For these locations, an entropy model employing a harmonic approximation may give a better result. The mixed case where some fluid molecules are stationary and others are not can present a problem for these estimators. These may occur when water molecules become 'frozen' into the binding pocket of a protein. In this situation, the frozen water molecules should be excluded from the general calculation and their entropies calculated separately and combined to give a final figure.

Another limitation is for systems which express symmetry and are effectively frozen. A water molecule is invariant under relabelling of hydrogen atoms (ignoring para- and ortho-water), but a frozen molecule will not be able to explore this degeneracy upon sampling and a manual degeneracy factor should be included into the entropy calculation. This factor can be increasingly large for highly symmetric molecules, such as cucurbit[7]uril used in guest-host studies [35]. This arises from the rotational degeneracy factor which divides the rotational partition function which can be calculated from a group theoretical appraisal of the molecule in question [36]. This mechanism leads to residual entropy, for example in the ice models of Pauling and Lieb. The nearest neighbours implementation of IFST will not be able to detect such entropic terms.

2.6 Grid Cell Theory

There are other methods which give the entropic and enthalpic components of a free energy change and it is worth briefly comparing the alternatives. A notable example is grid cell theory [37–40] which has been suggested as an alternative to IFST and GIST methods [14]. GCT handles the entropy calculations using harmonic approximations and a generalised Pauling residual entropy model which avoids the need for a truncated series expansion as used in IFST. It is argued that this method would capture the higher order entropy terms implicitly but may suffer different uncertainties to IFST leaving the two methods comparable in terms of accuracy of prediction [14].

Symbol	Definition
U_{elec}	Electrostatic energy between two charges.
U_{LJ}	Lennard-Jones repulsion dispersion energy between any two atoms.
U_{bond}	Harmonic bond energy between two bonded atoms.
U_{angle}	Harmonic angle energy between three atoms bonded in a line.
U_{UB}	Urey-Bradly harmonic energy between two unconnected atoms in an angle.
U_{dihedral}	Energy associated with four atoms bonded in a line based on dihedral angle defined by planes containing atoms (1 and 2) and (3 and 4).
U_{improper}	Energy associated with four atoms bonded with a single branch.
C_{Coulomb}	electrostatic constant which is approximately 332.0636 kcal·Å/(mol·e^2)
q_k	Number of electron charges on atom k, usually fractional.
η	Dielectric constant, usually 1 (explicit solvent) or 80 (implicit solvent).
ε_k	L-J well depth parameter on atom k.
r_{min}	The average of two atoms L-J radius parameters.
k_X	Force constant for the type of interaction X.
$\phi_0, \theta_0, r_{\text{eq}}, r_{\text{ub}}$	Equilibrium angles or distances for harmonic terms.
$\text{ang}(x_1, x_2, x_3)$	Angle between atoms 1,2 and 3, where 1 is bonded to 2 and 2 bonded to 3.
$\text{dih}(x_1, \cdots, x_4)$	Dihedral angle between the plane with atoms 1 and 2 and the plane with atoms 3 and 4.
$\text{imp}(x_1, \cdots, x_4)$	Improper angle between the three atoms.

Table 2.1 Description of MD parameters and force field terms.

Chapter 3

On the Accuracy of One and Two Particle Solvation Entropies

This Chapter is mostly based on the publication "On the Accuracy of One and Two Particle Solvation Entropies" [41].

3.1 Motivation and Overview

The ability to estimate the free energy difference between two defined states is a useful tool in computational chemistry. Direct applications are predicting the free energy of a solvation process, whether it be testing a small molecule in a solvent to see if the two are likely to be miscible [33], or a larger system, for example a peptide, protein [42], or protein-ligand complex [43]. The latter has a direct implication to in-silico drug design. Being able to quantitatively measure such a change then allows the relative comparison of the binding strength ligands for a given protein [44]. For the protein-ligand complex if the free energy of the bound state is less than the unbound state then the equilibrium will favour the bound state.

There exist a number of methods of estimating changes in free energy with computational simulations. These include Free Energy Perturbation (FEP) type methods [15] and Thermodynamic Integration (TI)[25]. Both of these methods rely on a well defined path from the reference state to the new state, but are very generally applicable and work with different levels of theory for the description of the physical system [45]. Another method which has seen growing success is Inhomogeneous Fluid Solvation Theory (IFST) [20, 21, 34], which does not need a well defined path between the reference state and the new state, however IFST can only calculate changes in free energy in the context of a solvation process. IFST performs this by using a direct computation of the change in entropy due to solvation and combining this with a direct energy measurement [46, 47] to give a free energy. There are numerous examples of calculating and estimating such a change in free energy [48–53].

This work attempts to expand on one method of measuring the solvation entropy change directly, namely the Mutual Information Expansion (MIE) [54], and uses a k-Nearest Neighbours (KNN) [55, 56] estimator to evaluate an approximation to the change in solvation entropy. This work is different to the majority of previous work, in that we truncate the MIE at a higher order, including an additional term. This additional term represents correlations between two solvent molecules in the presence of a solute. Such terms have been measured before in the context of water in protein binding pockets [57] using Grid Inhomogeneous Solvation Theory (GIST) [35, 58]. We seek to find quantitatively the change in free energy associated with the extra information, and whether it is necessary to include this term in the MIE. The system under study in this work is fundamental and simple, a Lennard-Jones neon solvent with a fixed Lennard-Jones atom in the centre of the simulation cell which represents the solute in a solvation process. The changes in free energy will be compared to an equivalent FEP simulation. The comparison of FEP and IFST in this work is independent of the force field used.

First we will discuss the theory used in all calculations, then review the computational methods used and the simulation parameters. The results will then be discussed and analysed.

3.2 Theory

Our calculations of the solvation free energy associated with fixing a Lennard Jones atom at the centre of a neon simulation box (see Fig. 3.1) are based on the following procedure:

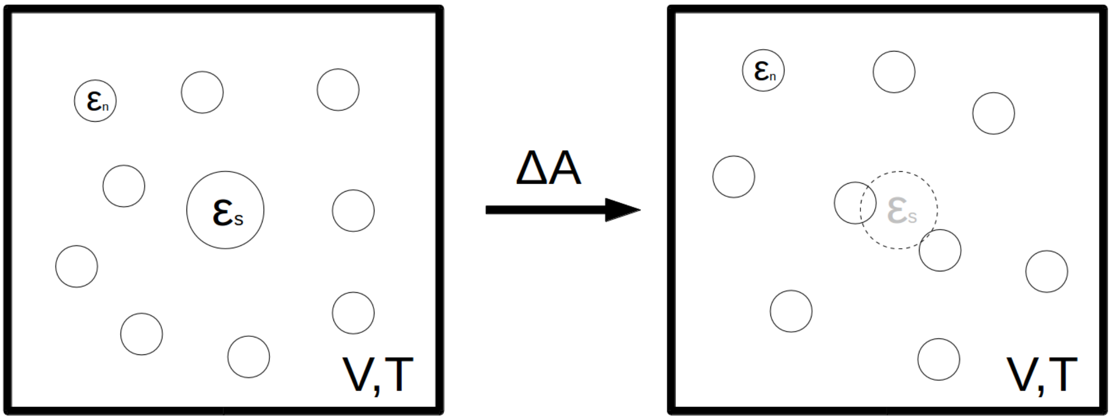


Fig. 3.1 Schematic of the canonical ensemble simulation for both IFST and FEP. N solvent neons with VDW parameter ε_n surround the solute with VDW parameter ε_s, (left). The solute interactions are turned off, while the volume and temperature remain constant, (right). Both methods in this study measure the change in Helmholtz free energy, ΔA between these two states.

Step 1

A change in free energy in the canonical ensemble is given by a change in the Helmholtz free energy

$$\Delta A = \Delta U - T\Delta S \tag{3.1}$$

where ΔU is the change of internal energy in the system, T is the temperature of the system and ΔS is the change in entropy of the system. The changes here are the difference between the solute-liquid system and the bulk liquid system.

Step 2

By using a molecular dynamics (MD) simulation with a parametrised force field it is possible to estimate ΔU as

$$\Delta U = \bar{U}_2 - \bar{U}_1 \tag{3.2}$$

where $\bar{U}_1$ is the equilibrium expectation energy of system of N neon atoms in a periodic cell of volume V at temperature T, and $\bar{U}_2$ is the equilibrium expectation energy of a system of N neon atoms in the presence of a fixed Lennard-Jones atom at volume and temperature V and T.

Step 3

It is possible to write the total change in solvation entropy, ΔS in equation 3.1, as an expansion over correlations of one body, two bodies, three bodies an so on, as discussed by Baranyai and Evans [59] for the homogeneous fluid and by Lazaridis [20, 21] for the inhomogeneous fluid.

For the system of N ideal atoms with coordinates $\{\mathbf{r}\}_N = \{\mathbf{r}_1, \cdots, \mathbf{r}_N\}$, and momenta $\{\mathbf{p}\}_N = \{\mathbf{p}_1, \cdots, \mathbf{p}_N\}$ we have the N body distribution in positions and momenta

$$f_N = f_N(\{\mathbf{r}\}_N, \{\mathbf{p}\}_N) = g_N(\{\mathbf{r}\}_N) \prod_{k=1}^{N} f_k(\mathbf{p}_k) \tag{3.3}$$

The total entropy of a bulk fluid is given by

$$S_{\text{liquid}} = -\frac{Rh^{3N}}{N!} \int f_N \ln f_N \, d\{\mathbf{r}\}_N d\{\mathbf{p}\}_N \tag{3.4}$$

with R the gas constant, h the Planck constant. Then the separability of the momentum can be exploited to give

$$S_{\text{liquid}} = S_{\text{momentum}} + S_{\text{configuration}}, \tag{3.5}$$

which written explicitly is

$$S_{\text{liquid}} = \underbrace{-\frac{NR}{\rho} \int f_1(\mathbf{p}_1) \ln f_1(\mathbf{p}_1) \, d\mathbf{p}_1}_{\text{Momentum}} \underbrace{- \frac{R\rho^N}{N!} \int g_N(\{\mathbf{r}\}_N) \ln g_N(\{\mathbf{r}\}_N) \, d\{\mathbf{r}\}_N}_{\text{Configuration}}. \tag{3.6}$$

The momentum terms are the same as those of an ideal gas where the one-body distribution of momenta is given by

$$f_1(\mathbf{p}) = \rho(2\pi mkT)^{-3/2} \exp\left(\frac{\mathbf{p}^2}{2mkT}\right). \tag{3.7}$$

where k is the Boltzmann constant, ρ is the number density of the equivalent ideal gas and m is the mass of the atom. Then

$$S_{\text{momentum}} = -\frac{NR}{\rho} \int f_1(\mathbf{p}_1) \ln f_1(\mathbf{p}_1) \, d\mathbf{p}_1 = \frac{3NR}{2} - NR \ln(\rho\lambda^3) \tag{3.8}$$

with λ the thermal wavelength of an atom in the liquid. For a vector of random variables $\mathbf{X} = (X_1, \cdots, X_n)$ we may write [18]

$$H(\mathbf{X}) = H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \cdots + H(X_n|X_1, \cdots, X_{n-1}) \tag{3.9}$$

where H is an information entropy, and the notation $H(X|Y)$ is the conditional entropy of X given Y. It should be stressed that expansions of this form are usually only useful where correlations between random variables decay exponentially with separation [24]. This is expected to be the case for the systems used in this study because there are no charge interactions. In a similar fashion we may write

$$\underbrace{-\frac{R\rho^N}{N!}\int g_N(\{\mathbf{r}\}_N)\ln g_N(\{\mathbf{r}\}_N)\; d\{\mathbf{r}\}_N}_{\text{Configuration}} = S_2 - I_3 + I_4 - \cdots \tag{3.10}$$

giving

$$S_{\text{liquid}} = S_{\text{momentum}} + S_{2\text{ liquid}} - I_{3\text{ liquid}} + I_{4\text{ liquid}} - \cdots \tag{3.11}$$

This expansion was performed for a homogeneous fluid by Wallace, and is only exact if the entire system is integrated over in the NVT ensemble, i.e. it is non-local [29, 59]. An ensemble-invariant and local form of the expansion is discussed by Baranyai and Evans [59]. In the conditions of this study the distinguishing terms between the local and non-local forms cancel, so here the non-local form is used for simplicity. In this case the expansion has terms as follows

$$S_{\text{momentum}} = -\frac{NR}{\rho}\int f_N^{(1)}(\mathbf{p})\ln h^3 f_N^{(1)}(\mathbf{p})\; d\mathbf{p} \tag{3.12}$$

$$S_{2\text{ liquid}} = -\frac{R\rho^2}{2!}\iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)\ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)\; d\mathbf{r}_1 d\mathbf{r}_2 \tag{3.13}$$

$$I_{3\text{ liquid}} = \frac{R\rho^3}{3!}\iiint g_N^{(3)}(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)\ln\delta g_N^{(3)}(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)\; d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}_3 \tag{3.14}$$

where we can write the part of the three-body correlation function which cannot be expressed multiplicatively by its marginal distributions as

$$\delta g_N^{(3)}(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3) = \frac{g_N^{(3)}(\mathbf{r}_1,\mathbf{r}_2,\mathbf{r}_3)}{g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_3)g_N^{(2)}(\mathbf{r}_2,\mathbf{r}_3)} \tag{3.15}$$

where each of these N particle correlation functions for k bodies can be written in terms of the k body density as [60]

$$g_N(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_k) = \frac{1}{\rho^k} \rho_N(\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_k) \tag{3.16}$$

Finally we may write the excess entropy of the liquid as

$$S_{\text{excess liquid}} = S_{\text{liquid}} - S_{\text{ideal}} \tag{3.17}$$

we have

$$S_{\text{ideal}} = S_{\text{momentum}} + R = \frac{5NR}{2} - NR \ln(\rho\lambda^3) \tag{3.18}$$

Therefore,

$$S_{\text{excess liquid}} = S_{2\text{ liquid}} - I_{3\text{ liquid}} + I_{4\text{ liquid}} - \cdots - R, \tag{3.19}$$

where terms with an S denote an entropy, and terms with an I denote a mutual information term. The expansion with these terms is true for a homogeneous fluid. For the simulations used in this work we include the presence of the central solute [20]. The central solute in our calculations is spherically symmetric, our end goal is to generalise to any solute and solvent, so we must consider an inhomogenous system. This was analytically performed by Lazaridis [20, 21]. Equations 3.12 through 3.14 are the relevant terms for the liquid; for a system with a solute we may write

$$S_{1\text{ solute}} = -R\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s)\; d\mathbf{r}_1 \tag{3.20}$$

$$I_{2\text{ solute}} = \frac{R\rho^2}{2!} \iint g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) \ln \delta g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s)\; d\mathbf{r}_1 d\mathbf{r}_2 \tag{3.21}$$

$$I_{3\text{ solute}} = -\frac{R\rho^3}{3!} \iiint g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s) \ln \delta g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s)\; d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}_3 \tag{3.22}$$

where the $|s$ argument indicates the presence of a solute. We can write the part of the two and three body correlation function which cannot be expressed by its marginal distributions as

$$\delta g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) = \frac{g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s)}{g_N^{(1)}(\mathbf{r}_1|s) g_N^{(1)}(\mathbf{r}_2|s)} \tag{3.23}$$

$$\delta g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s) = \frac{g_N^{(3)}(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3|s) g_N^{(1)}(\mathbf{r}_1|s) g_N^{(1)}(\mathbf{r}_2|s) g_N^{(1)}(\mathbf{r}_3|s)}{g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_2|s) g_N^{(2)}(\mathbf{r}_1, \mathbf{r}_3|s) g_N^{(2)}(\mathbf{r}_2, \mathbf{r}_3|s)} \tag{3.24}$$

Some slight differences exist between the homogeneous and inhomogeneous formulations: $S_{2\,\text{liquid}}$ is not a mutual information term (denoted with an I rather than an S), as the marginal distributions vanish. We may then write

$$S_{\text{ solute}} = S_{\text{momentum}} + S_{1\text{ solute}} - I_{2\text{ solute}} + I_{3\text{ solute}} - \cdots \tag{3.25}$$

and subtracting the ideal gas terms, equation 3.18

$$S_{\text{excess solute}} = S_{1\text{ solute}} - I_{2\text{ solute}} + I_{3\text{ solute}} - \cdots - R \tag{3.26}$$

To find the change in entropy from the addition of the solute, we calculate the *excess entropy of solution,* $\Delta S_{\text{exc soln}}$. This is then the excess entropy of the solute system, equation 3.26 minus the excess entropy of the pure liquid, equation 3.19.

$$\Delta S_{\text{exc soln}} = S_{\text{excess solute}} - S_{\text{excess liquid}} \tag{3.27}$$

$$\Delta S_{\text{exc soln}} = [S_{1\text{ solute}} - I_{2\text{ solute}} + I_{3\text{ solute}} - \cdots] - [S_{2\text{ liquid}} - I_{3\text{ liquid}} + I_{4\text{ liquid}} - \cdots] \tag{3.28}$$

where we see the factors of $-R$ from $S_{\text{excess solute}}$ and $S_{\text{excess liquid}}$ cancel. It is at this point we choose a truncation of $\Delta S_{\text{exc soln}}$. The most severe truncation generates what we will call the *conditional one particle entropy* (C1PE)

$$\Delta S_{1|s} = S_{1\text{ solute}} \tag{3.29}$$

This is achieved by removing all integrals with a subscript greater than 1. If we include the two body terms then we have the *conditional two particle entropy* (C2PE)

$$\Delta S_{2|s} = S_{1\text{ solute}} - I_{2\text{ solute}} - S_{2\text{ liquid}} \tag{3.30}$$

It is $\Delta S_{2|s}$ that will be calculated in this work.

Step 4

Instead of directly integrating numerically the terms in equations 3.29 and 3.30, we can instead convert the integral into a sum which converges to the integral asymptotically in the limit of infinite data. We can then extrapolate toward that infinite limit by measuring the sum for finite quantities of data and fitting an appropriate extrapolating function. The estimators we use in this study are KNN estimators and are formulated as follows.

Inserting the explicit integral terms into equation 3.30 gives

$$\begin{aligned}\Delta S_{2|s} \quad &= -R\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s)\, d\mathbf{r}_1 \\ &-\tfrac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s) \ln \frac{g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s)}{g_N^{(1)}(\mathbf{r}_1|s) g_N^{(1)}(\mathbf{r}_2|s)}\, d\mathbf{r}_1 d\mathbf{r}_2 \\ &+\tfrac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \end{aligned} \tag{3.31}$$

we may separate out the denominators of the logarithm in the second term and then swap the coordinates $\mathbf{r}_1$ and $\mathbf{r}_2$ in one of those new integrals to give

$$\begin{aligned}\Delta S_{2|s} \quad &= -R\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s)\, d\mathbf{r}_1 \\ &-\tfrac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s) \ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s)\, d\mathbf{r}_1 d\mathbf{r}_2 \\ &+R\rho^2 \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s) \ln g_N^{(1)}(\mathbf{r}_1|s)\, d\mathbf{r}_1 d\mathbf{r}_2 \\ &+\tfrac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \end{aligned} \tag{3.32}$$

it is then possible to integrate over $\mathbf{r}_2$ in the third term giving

$$\begin{aligned}\Delta S_{2|s} \quad &= R(N-2)\rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s)\, d\mathbf{r}_1 \\ &-\tfrac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s) \ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s)\, d\mathbf{r}_1 d\mathbf{r}_2 \\ &+\tfrac{R\rho^2}{2} \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \end{aligned} \tag{3.33}$$

It is convenient to construct estimators to measure three distinct quantities

$$H_{1,s} = \rho \int g_N^{(1)}(\mathbf{r}_1|s) \ln g_N^{(1)}(\mathbf{r}_1|s)\, d\mathbf{r}_1 \tag{3.34}$$

$$H_{2,s} = \rho^2 \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s) \ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2|s)\, d\mathbf{r}_1 d\mathbf{r}_2 \tag{3.35}$$

$$H_{2,l} = \rho^2 \iint g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2) \ln g_N^{(2)}(\mathbf{r}_1,\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \tag{3.36}$$

giving an expression for the conditional one and two particle entropies in terms of these estimators

$$\Delta S_{1|s} = RH_{1,s} \tag{3.37}$$

$$\Delta S_{2|s} = R(N-2)H_{1,s} + \frac{R}{2}\left(H_{2,l} - H_{2,s}\right). \tag{3.38}$$

In general we may state the information entropy of a p.d.f over p random variables as a p dimensional integral

$$H[\rho] = -\int \rho(\{\mathbf{x}\}_p) \ln \rho(\{\mathbf{x}\}_p)\; d\{\mathbf{x}\}_p \tag{3.39}$$

which is the expectation

$$H[\rho] = \mathbb{E}[-\ln \rho(\{\mathbf{x}\}_p)] \tag{3.40}$$

However, in equations 3.34-3.36 we have correlation functions in the logarithm, rather than densities. We may then instead calculate

$$H[\rho] + \int \rho(\{\mathbf{x}\}_p) \ln \rho \; d\{\mathbf{x}\}_p \tag{3.41}$$

This quantity can then be approximated by a finite sum which is performed in the next section.

3.2.1 k-Nearest Neighbours (KNN) Estimators

To efficiently calculate the integrals in equations 3.29 and 3.30, we used the k-nearest neighbours method which has shown previous success in calculating first order solvation entropies [34, 48, 35], dihedral entropies in small molecules [61], and entropies of water in protein binding pockets [49, 57]. The general estimator for N_p p-dimensional objects across F frames of data is given by

$$H^{(k)} \cong \frac{1}{N_p F} \sum_{i=1}^{N} \sum_{j=1}^{F} \ln\left(\frac{N_p (F-1) \pi^{p/2} d_{ij,k}^{p}}{\Gamma\left(\frac{p}{2}+1\right)} \right) - \psi(k) \tag{3.42}$$

where $\Gamma(x)$ is the Euler gamma function, $\psi(k)$ is the Euler digamma function, and the symbol $\cong$ denotes an asymptotic equivalence in the limit of an infinite number of frames F. This estimator was initially worked on for the first nearest neighbour by Leonenko [55] and extended for the k^{th} nearest neighbour by Singh [56] and has been used by various studies [19, 61, 54, 34, 48, 49]. This estimator is a limiting case of the adaptive anisotropic elliptic kernel estimators [62, 63].

The estimator for an H_1 term, given F sufficiently uncorrelated MD frames containing N neon atoms is given by the expression

$$H_1^{(k)} \cong \frac{1}{NF}\sum_{i=1}^{N}\sum_{j=1}^{F}\ln\left(\frac{4N(F-1)\pi d_{ij,k}^3}{3V}\right) - \psi(k) \tag{3.43}$$

which is used in [48, 49]. The higher order expression is then

$$H_2^{(k)} \cong \frac{2}{N(N-1)F}\sum_{i_1=1}^{N}\sum_{i_2>i_1}^{N}\sum_{j=1}^{F}\ln\left(\frac{N(N-1)(F-1)\pi^3 d_{i_1 i_2 j,k}^6}{6V^2}\right) - \psi(k). \tag{3.44}$$

The nature of convergence with respect to frames F of equations 3.43 and 3.44 has previously been successfully fitted with a power law [48, 49, 54]

$$H_k(F) = a_k F^{b_k} + H_\infty \tag{3.45}$$

with a_k and b_k some constants for the k^{th} neighbour selected, and H_∞ the asymptotic value of the entropy. b_k is a negative constant, such that

$$\lim_{F\to\infty} H_k(F) = H_\infty \tag{3.46}$$

to extract this value a power law can be fitted to $H(F)$. Examples of this are shown in figure 3.2.

3.3 Simulation Details

FEP and IFST were used to calculate the free energy change associated with adding a fixed Lennard-Jones particle to the centre of a neon simulation box in the canonical ensemble. The IFST free energy with only $\Delta S_{1|s}$ and with both $\Delta S_{1|s}$ and $\Delta S_{2|s}$ were calculated. All of the simulation data used in this study came from the same force field and MD simulations were all carried out in NAMD [7]. Thus, the resulting free energies are directly comparable.

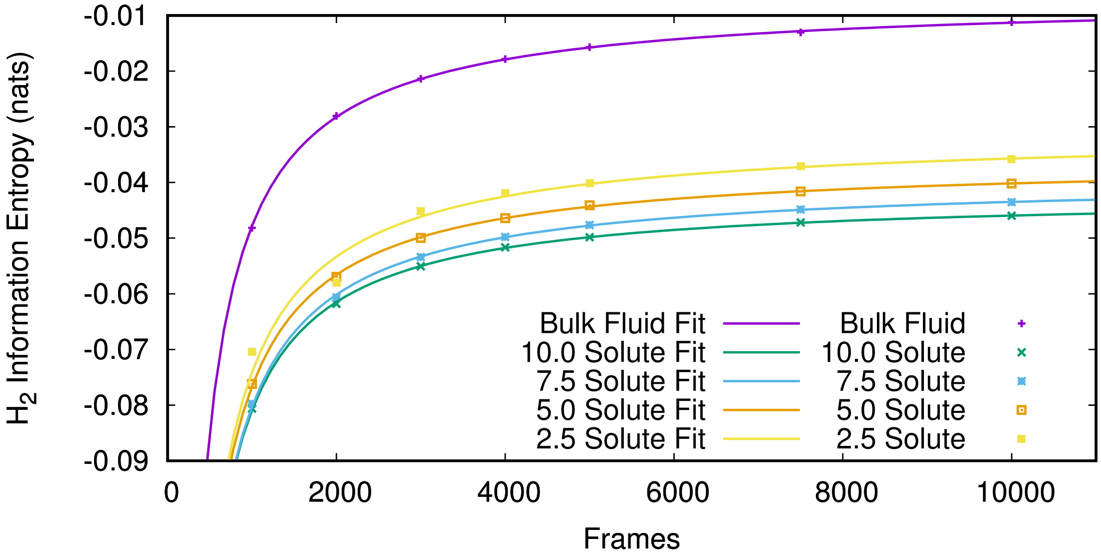


Fig. 3.2 A plot demonstrating how the the asymptotic values for the $H_{s,2}$ and $H_{l,2}$ estimators were extracted. A power law is fitted to the three data points for each solute. The extrapolated constant in the limit of infinite frames is used as according to equation 3.45.

3.3.1 Technical Details

The neon MD parameters were taken from the CHARMM27 force field [64]. There are no electrostatic interactions for this solvent-solute system, so all potential energy terms are in the form of a Lennard-Jones potential

$$V_{LJ}(\mathrm{r}_{ij}) = \sqrt{\varepsilon_i \varepsilon_j}\left(\left(\frac{R_i + R_j}{2|\mathrm{r}_{ij}|}\right)^{12} - 2\left(\frac{R_i + R_j}{2|\mathrm{r}_{ij}|}\right)^{6}\right) \tag{3.47}$$

where r_{ij} is the distance between atoms i and j, and R_k is the radius of minimum potential energy for atom k, and $R_k = 2^{1/6}\sigma_k$. 4 different solutes were made that had varying Lennard-Jones ε_s parameters as shown in table 3.1. All solutes had the same R_k parameter as the neon solvent. Each solute was initially solvated with 900 neon atoms in a cubic box of edge length 27.5Å. Where reduced units are given for reference they are calculated as temperature $T^* = k_B T/\varepsilon$, pressure $P^* = p\sigma^3/\varepsilon$, length $L^* = L/\sigma$ and number density $\rho^* = \rho\sigma^3$.

For the 12-6 LJ fluid, the triple point density of both the liquid ρ^*_{tl} and solid ρ^*_{ts} phase have been measured by previous studies to be in the ranges $\rho^*_{tl} = 0.818 - 0.864$ and $\rho^*_{ts} =$

Solute Name	ε_s [kcal/mol]	R_k [Å]	L	ρ [Å^{-3}]	T [K]
SOL1	0.215	3.06	27.3760	0.04382	25
SOL2	0.430	3.06	27.3710	0.04384	25
SOL3	0.645	3.06	27.3707	0.04384	25
SOL4	0.860	3.06	27.3685	0.04385	25
Solute Name	$\varepsilon_s/\varepsilon_n$	σ_s [Å]	L^*	ρ^*	T^*
SOL1	2.5	2.73	10.0420	0.8887	0.578
SOL2	5.0	2.73	10.0401	0.8892	0.578
SOL3	7.5	2.73	10.0400	0.8893	0.578
SOL4	10.0	2.73	10.0392	0.8895	0.578

Table 3.1 This table shows the force field parameters used for each of the solute atoms where ε_n is the value for bulk neon.

0.96 – 0.978, and the triple point has approximate temperature and pressure $T^* = 0.661$ and $p^* = 0.0018$ [65]. Thus the neon in the simulation was in the liquid phase.

3.3.2 IFST

IFST Equilibration

16 replicate equilibrations per solute type were performed in the NpT ensemble for 4 ns to find the equilibrium densities of the simulation cells with the each solute present. The resulting edge lengths are shown in table 3.1.

A 2 ns equilibration was then performed for each replicate in the NVT ensemble at $T = 25$ K ($T^* = 0.578$) and 1 atm ($p^* = 0.00344$) using Langevin temperature control. An MD timestep of 2 fs was used. Electrostatic interactions were turned off as they were not present in the force field. Van der Waals interactions were removed for separations over 10.5 Å and switching was used between 9.5 Å and 10.5 Å. The simulations were carried out with cubic periodic boundary conditions. This process was repeated for each of the 4 solutes and one bulk system with no solute.

IFST Production

The production simulations were carried out in the NVT ensemble for 60 ns at 25 K ($T^* = 0.578$) with the same MD parameters used in the NVT equilibration. NAMD produces a trajectory (.dcd) file, which is a set of system coordinate snapshots. A trajectory frame was saved every 500 fs such that neighbouring frames in the dcd file were not too strongly correlated; this is important for the KNN estimator when used later to extract entropies as commented in previous work by Huggins [48]. The energies $\Delta U = \bar{U}_{solute} - \bar{U}_{bulk}$ were calculated by taking the average energy across all 16 repeats of the 60 ns production simulation, these are shown in table 3.2 along with the standard deviation across 16 repeats, $\sigma_{\Delta U}$.

IFST Free Energy Calculations

To calculate the change in free energy for IFST simulations we need to calculate each component on the right hand side of

$$\Delta A_{1|s} = \Delta U - T\Delta S_{1|s}, \tag{3.48}$$

and of

$$\Delta A_{2|s} = \Delta U - T\Delta S_{2|s}. \tag{3.49}$$

ΔU was calculated by equation 3.2, therefore the quantities $\Delta S_{1|s}$ and $\Delta S_{2|s}$ must be calculated using equations 3.37 and 3.38. This then requires collecting a set of nearest neighbour distances and nearest pair distances from the respective MD data for the entropy estimators.

3.3.3 K-D Trees

Previous studies using IFST or GIST have used a grid of voxels, or cut-cell approach to find nearest neighbour distances [48, 49]. For the pair nearest neighbour distances required for equation 3.44 this method was not practical. In order to increase efficiency and allow the calculation of the pair terms, a K-Dimensional (K-D) Tree method was used. This is described in more detail in appendix B.

3.3.4 Free Energy Perturbation (FEP)

FEP Protocol

The FEP simulation measures the change in free energy of removing a fixed solute from a box of neon. This simulation is parametrised by a variable λ, such that when $\lambda = 0$, the solute is fully present, and when $\lambda = 1$, the solute is fully annihilated.

In this study, each forward and backward FEP simulation had $N = 64$ 'λ-windows'. Each window has a different value of lambda which is labelled *degree of annihilation* in the schedule displayed in Fig. 3.3. For the n^{th} window out of a total of N windows, the value of λ in that window was generated with the expressions

$$\lambda_f(n) = \frac{1}{N}\sum_{k=1}^{N}\left[1-\left(1-\frac{n}{N}\right)^{\max(k-N+n,1)}\right] \tag{3.50}$$

$$\lambda_b(n) = \frac{1}{N}\sum_{k=1}^{N}\left[1-\left(\frac{n}{N}\right)^{\max(k-n,1)}\right] \tag{3.51}$$

which satisfies $\lambda_f(0) = 0, \lambda_f(N) = 1, \lambda_b(0) = 1$ and $\lambda_b(N) = 0$ where $\lambda_f(n)$ and $\lambda_b(n)$ are the forward and backward schedules respectively. These curves were picked to sample the endpoints of the FEP simulation more heavily, to avoid the so-called 'end point catastrophe', which is where very weakly interacting atoms in the MD simulation can overlap and the divergent form of their interaction potential leads to very large energies and slow convergence [66]. The endpoints are when the solute is close to full annihilation in the forward simulation ($n \approx N$), or when the solute is just being created in the backward simulation ($n \approx 0$). A Van der Waals soft core potential parameter of 5.0 was used. One energy reading was stored every 100 steps which equates to every 0.2 ps.

FEP Equilibration

The ends of the NVT equilibrations for the IFST simulations were used as starting points for the FEP simulations. Each solute had 16 replicates. Each simulation underwent a further 0.2 ns equilibration at $T = 25$ K ($T^* = 0.578$) from this point to adjust to their individual λ values.

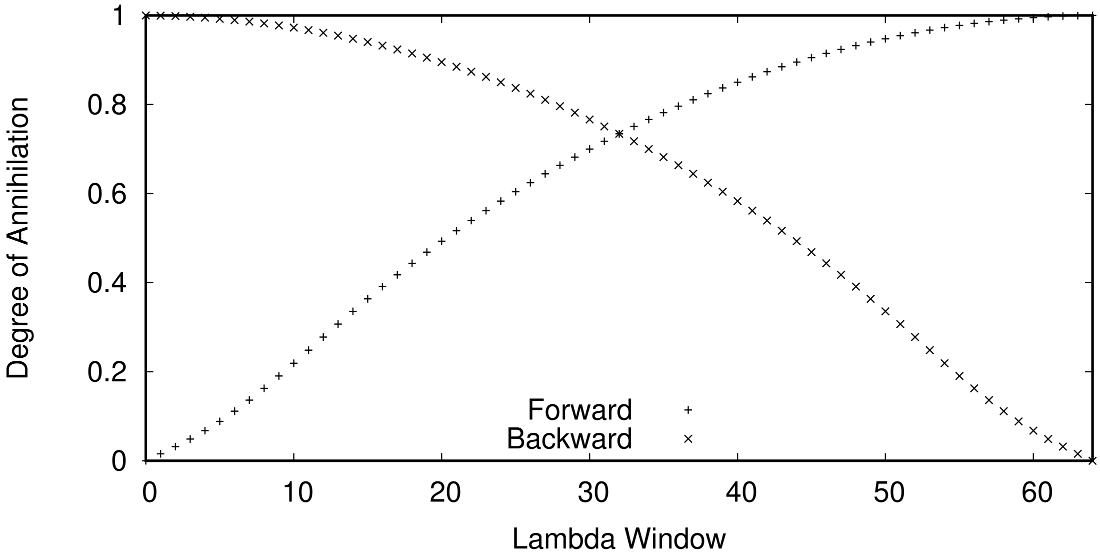


Fig. 3.3 The lambda schedule for the forward and backward FEP simulations across the 64 lambda windows as generated by equations 3.50 and 3.51.

FEP Simulation

For each λ-window, a 0.469 ns simulation was performed in the NVT ensemble at $T = 25$ K ($T^* = 0.578$). Then considering the 64 windows in both directions the overall simulation time was 60.1 ns, which is comparable to the 60 ns used for the IFST production run.

FEP Calculations

The final change in free energy from the FEP simulations is found by summing the changes in free energy between each pair of neighbouring λ windows. Then the forward change is

$$\Delta A_{f,FEP} = \sum_{n=0}^{63} \Delta A_{\lambda(n),\lambda(n+1)} \tag{3.52}$$

and the backward change is

$$\Delta A_{b,FEP} = \sum_{n=1}^{64} \Delta A_{\lambda(n-1),\lambda(n)} \tag{3.53}$$

to calculate the changes in free energy between the λ-windows, the Bennet Acceptance Ratio (BAR) method was used [16], which is included in the ParseFEP Plugin [17] for VMD

[67]. This is used for all FEP results for this thesis. The BAR estimator provides a calculated statistical error, these errors were less than 0.007 kcal/mol for all of the 64 simulations.

3.4 Results and Discussion

3.4.1 Results

Fig. 3.4 shows the solute-fluid radial distribution functions for all solutes and bulk neon.

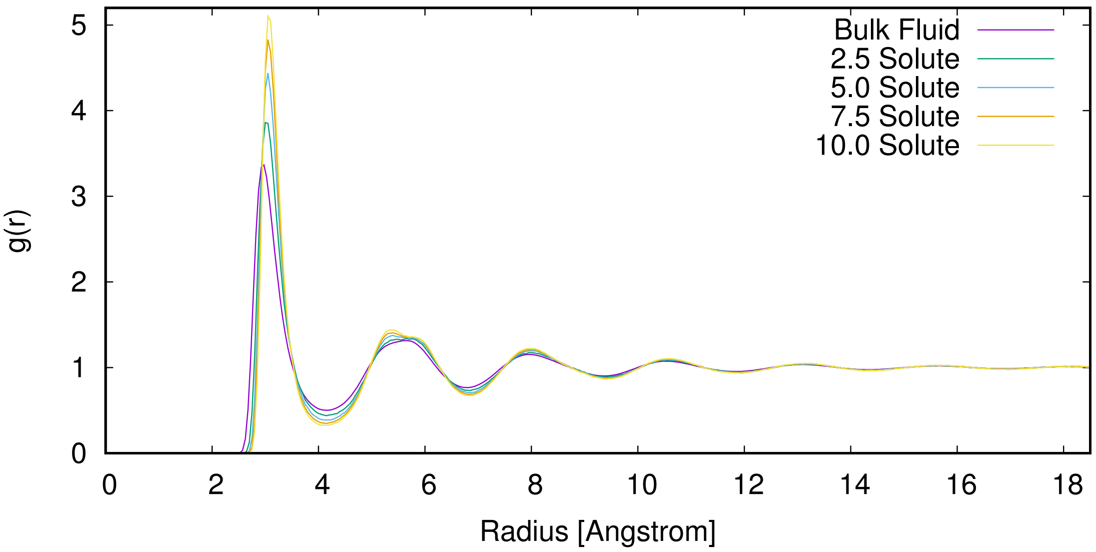


Fig. 3.4 The radial distribution functions for the solute to all fluid atoms which describes the relatively probability of finding a fluid particle at a radius (x-axis) from a solute fixed at the centre of the box. The fluid-fluid RDF is also plotted which is equivalent to a regular neon solute. For stronger central potentials the peaks get higher and the troughs get deeper along with a consistent distortion in the second solvation shell from 5 to 6 Å.

Fig. 3.5 shows a comparison of the free energy estimates from both levels of IFST results and from FEP.

Table 3.2 shows the comparison of the average free energies from FEP and IFST calculations, the IFST calculations show the conditional one particle and conditional two particle entropy values.

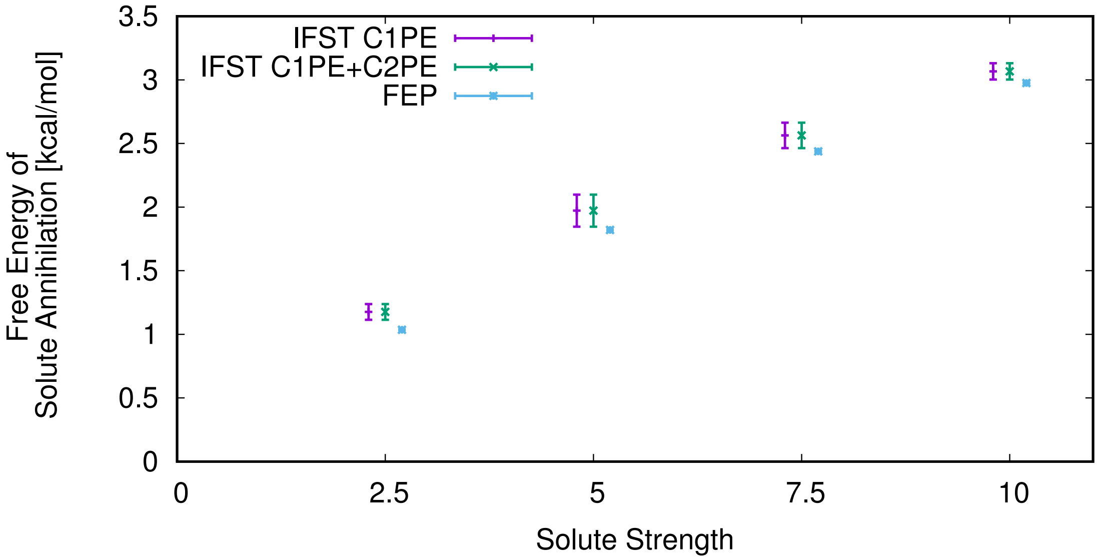


Fig. 3.5 The free energy estimates for the two IFST approximations and the FEP result at ε_s values of $2.5\varepsilon_n, 5.0\varepsilon_n, 7.5\varepsilon_n$ and $10.0\varepsilon_n$. Different levels of theory have been spaced for clarity. C1PE is the conditional one particle entropy. C2PE is the conditional two particle entropy correction. The error bars are the spread in the 16 repeats of each result.

	ΔA_{FEP}	ΔU	$\Delta S_{1\vert s}$	$\Delta S_{2\vert s}$	$\Delta A_{1\vert s}$	$\Delta A_{2\vert s}$
SOL1	1.036	-1.717	-0.541	$< 10^{-4}$	1.176	1.176
SOL2	1.820	-2.561	-0.589	$< 10^{-4}$	1.972	1.972
SOL3	2.438	-3.257	-0.694	$< 10^{-4}$	2.563	2.563
SOL4	2.975	-3.839	-0.773	$< 10^{-4}$	3.067	3.067

	σ_{FEP}	$\sigma_{\Delta U}$	$\sigma_{\Delta S_{1\vert s}}$	$\sigma_{\Delta S_{2\vert s}}$	$\sigma_{1\vert s}$	$\sigma_{2\vert s}$
SOL1	0.00495	0.0279	0.052	$< 10^{-4}$	0.062	0.062
SOL2	0.00382	0.0281	0.131	$< 10^{-4}$	0.126	0.126
SOL3	0.00585	0.0238	0.088	$< 10^{-4}$	0.100	0.100
SOL4	0.00624	0.0155	0.061	$< 10^{-4}$	0.064	0.064

Table 3.2 This table shows the average free energy results for each solute. $\sigma_{\Delta U}$ is the standard deviation of the 16 energy results for each solute. σ_{FEP} is the standard deviation across all 16 repeats of the FEP free energy change ΔA_{FEP}. $\sigma_{1|s}$ and $\sigma_{2|s}$ are the standard deviations across all 16 repeats of the IFST free energy changes $\Delta A_{1|s}$ and $\Delta A_{2|s}$ respectively. All units are in kcal/mol.

The IFST free energies in table 3.2 were calculated by equations 3.48 and 3.49. The entropy terms we calculated by extrapolating equation 3.45 for the conditional one particle entropies every 1000 frames from 1000 to 12000. For the conditional two particle entropies equation 3.45 was used with 16 repeats taken at intervals of $1000, 2000$ and 3000 frames. 2 repeats were also extended to $1000, 2000, 3000, 4000, 5000, 7500$ and 10000 frames which was a much more expensive calculation. Fig. 3.2 shows an example of the conditional two particle entropy extrapolation for all solutes. The error weighted extrapolation routine used was from the gnuplot software. When fitting it weighted data points by the spread of the 16 repeats of that point. For the conditional two particle entropy fitting, where no standard deviation was available due to lack of data it was estimated to be $10/F$ kcal/mol, where F is the number of frames for that data point. This brought the errors down smoothly for larger F values, but still left a reasonable amount of uncertainty in those data points.

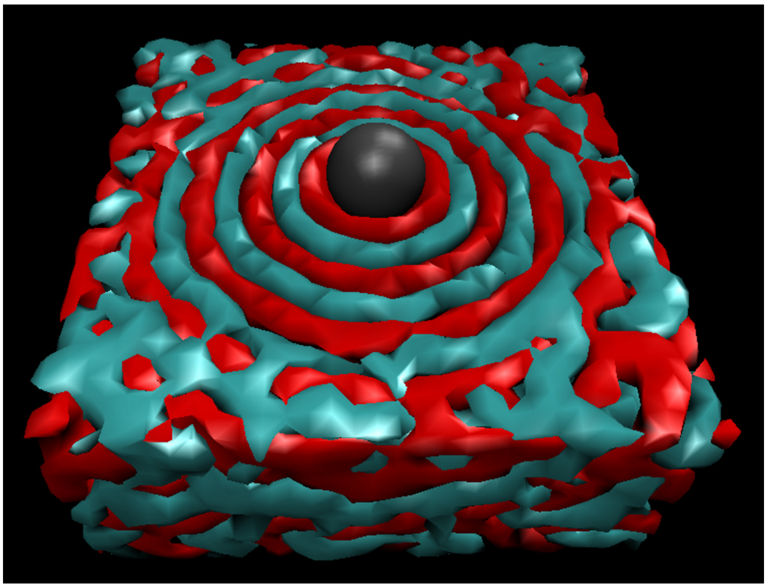


Fig. 3.6 A spatial model of the pair entropy distribution around the fixed neon solute. Clear ripples in the entropy can be seen with red indicating high values and blue indicating low values. The entropy follows the RDF as in figure 3.4

Figure 3.6 shows a decomposition of the pair entropy into the space around the solute neon atom (grey). This 3D histogram has voxels which each bin the individual contributions from atoms found within that voxel during the KNN summation. Only half of the cell is displayed so a cross section of the data can be seen. Red indicates regions of high entropy and blue regions of low entropy. The shape follows the RDF along a radial axis. For more complex solutes this symmetry would be broken and the field around the solute could be more complex.

3.4.2 Discussion

Fig. 3.4 shows the radial distribution functions (RDF) from the central solute to all solvent atoms for different solutes. The troughs become deeper and the peaks become higher for increased solute potential ε_s parameter. Also plotted is the RDF for the bulk solvent, which has the shallowest troughs and lowest peaks. There is a slight distortion in the second solvation shell which becomes more pronounced with higher ε_s. This is likely to be a gentle crowding effect as solvent molecules in the first solvation shell draw close to the solute, and the closest locations for second shell atoms correspond to the pits in the already tightly packed first solvation shell. The magnitude of the solvent-solvent correlation entropy was expected to be small in the system of monatomic species used in this study. It has been shown that in the Lennard-Jones system the primary source of solvent structure comes from packing[68]. This will correspond to an entropy associated with the volume exclusion of the central solute which is captured fully by the C1PE term. For more complex solvents such as water evidence already exists for solvent-solvent correlations [22, 69].

Fig. 3.5 shows the calculations of the free energy of the solute annihilation from all three methods of evaluation with the standard deviations of 16 repeats used as error bars. The average values of these estimates are displayed in table 3.2 along with the standard deviations of the 16 repeats. The average values for all methods are in good agreement. The IFST results including the conditional two particle corrections have the same averages and standard deviations as the first order IFST. This demonstrates that the C2PE has no clear contribution to the free energy this system. The FEP results have the lowest standard deviation of all results even though a comparable length of MD run was used. The increased uncertainty associated with the IFST results likely arise from fitting power laws to extract the asymptotic entropy. The standard deviation in the free energy is at least 3 times greater than that of the MD energies used, which indicates the added uncertainty is from the entropy. This could potentially be remedied by taking more data points during the extrapolation process at the expense of data processing time. Both the average IFST results are within 0.16 kcal/mol of the FEP results. These results indicate it is possible to reconstruct a fairly accurate measurement of the free energy change of this kind of process by only using the start and end points of an equivalent FEP path. The results indicate that the entropy term contributes relatively less to the free energy of solute annihilation for solutes with larger ε. It appears that the free energy was systematically overestimated by IFST methods against FEP methods. However, during the FEP calculation a softening parameter was used to help converge the results. This softening parameter may have been enough to change the Hamiltonians of the systems to create such a systematic difference.

3.5 Conclusions

The translational entropy associated with solute-solvent correlation and the solvent-solvent correlation can be used to evaluate a free energy of solvation in the IFST framework for a system of neon atoms solvating a fixed central Lennard-Jones potential. Both IFST estimates reproduce the value obtained from an equivalent FEP simulation to within around 0.16 kcal/mol and appear to consistently overestimate slightly. The IFST method has an advantage over FEP and can give a free energy estimate without having to define a path between the two systems. However, the accuracy associated with FEP estimates is greater, so it is a better tool for computation. We conclude that IFST in its current state is the better tool for physical interpretation as it avoids the non-physical lambda states utilized by FEP. We note that FEP is an already well developed method, and IFST may yet have future improvements to increase its optimisation. The conditional two particle entropy term did not contribute a noticeable change in free energy for this system for any of the strengths of solute potential tested.

The IFST framework can give the spatial contributions of configurational entropy when analysed with voxel based methods (as shown in fig. 3.6). This allows areas of interest around the solute to be highlighted. Although the two particle terms in the MIE do not appear to be significant in this simple system, they may be significant in a system with a liquid water solvent with hydrogen bonding and charges. The methods used in this work may be extended to such a system.

If solvent-solvent interactions became very strong in a particular solvation process it may be necessary to calculate the *conditional three particle entropy*

$$\Delta S_{3|s} = S_{1\,\mathrm{solute}} - I_{2\,\mathrm{solute}} + I_{3\,\mathrm{solute}} - S_{2\,\mathrm{liquid}} + I_{3\,\mathrm{liquid}} \tag{3.54}$$

The methodology used in this work could be extended to such a calculation. However, next order terms are likely to be expensive to evaluate.

Chapter 4

Atomic Contributions to Hydration and Binding using Free Energy Perturbation

This chapter relates to a general method called atom-wise free energy perturbation (AFEP), which extends a conventional molecular dynamics free energy perturbation (FEP) simulation to give the contribution to a free energy change from each atom. This work is based on the publication "Estimating Atomic Contributions to Hydration and Binding using Free Energy Perturbation" [70] which is a result of this work.

AFEP is derived from an expansion of the Zwanzig equation used in the exponential averaging method by proposing that the system total energy can be partitioned into contributions from each atom. A partitioning method is assumed and used to group terms in the expansion to correspond to individual atoms. AFEP is applied to six example free energy changes to demonstrate the effectiveness of the method; the hydration free energy of methane, methanol, methylamine, methanthiol and caffeine in water. AFEP highlights the atoms in the molecules that interact relatively (un)favourably with water. Finally AFEP is applied to the binding free energy of human immunodeficiency virus type 1 protease to lopinavir and AFEP reveals the contribution of each atom to the binding free energy, indicating candidate areas of the molecule to modify to produce a more strongly binding inhibitor. FEP gives a single value for the free energy change and is already a very useful method. AFEP gives a free energy change for each ‘part’ of the system being simulated, where part can mean individual atoms, chemical groups, amino acids or larger partitions depending on what the user is trying to measure. This method should find various applications in molecular dynamics studies of physical, chemical or biochemical phenomena, specifically in the field of computational drug discovery.

4.1 Introduction

Free energy methods refer to an existing set of methods for estimating free energy differences, for example traditional Free Energy Perturbation (FEP) summed using either exponential averaging (EXP) or the Bennett Acceptance Ratio (BAR) and thermodynamic integration (TI). These methods have become a cornerstone of accurate binding free energy calculations [71, 72] and many reviews have scrutinised the details of such methods, concluding that they are a promising addition to the set of tools used in the drug discovery industry [73–78]. Free energy methods are implemented in many commonly used biological simulation packages, for example NAMD[7, 67, 17], Desmond[79], GROMACS[80], BOSS[81], AMBER[82] and others. These methods are path based and rely on a set of intermediate states, named λ-windows, in which the interaction parameter λ is altered from 0 to 1. When $\lambda = 0$, the simulation represents the initial state of a physical system, this might be the unbound state of a protein and drug complex in a binding free energy calculation, or a simulation cell of water for a solvation free energy calculation. When $\lambda = 1$, the simulation represents the final state of a physical system. This might be the bound state for the protein-drug complex, or the solvated molecule for the solvation free energy calculation. Then a set of intermediate states is taken to gradually measure the free energy change as λ varies from 0 to 1. FEP can be used to extract hydration free energies [83, 34, 84], free energies associated with mutating one molecule into another, or one protein into another [85], and in the context of drug discovery, binding free energies of drug molecules to proteins [86, 44, 79].

The method described in this work is called atom-wise free energy perturbation (AFEP) and is an extension to the EXP method. It is applied to energy and trajectory data from a conventional molecular dynamics (MD) simulation with multiple lambda windows. AFEP estimates the contributions from each of the atoms in a system to a general free energy change, given that the total energy of the system is defined to be the sum of the atomic energies which are calculated from the MD trajectories. AFEP relies on the the same set of MD simulations as conventional FEP analysis does, running a simulation at each of the intermediate λ-values. It then calculates quantitatively the contributions to the free energy change from each 'part' of the system by approximating the decomposition of free energy in a simple and intuitive way. The definition of 'part' is flexible and could be single atoms or clusters of atoms grouped together (e.g. chemical groups, amino acids or the solvent) depending on the desired application.

4.1.1 Limitations of the Method

It is important to state clearly the limitations and assumptions of this method. Although the derivation of method is general for any free energy change that could be measured using the Zwanzig equation in the EXP method, that equation is already limited in its use cases:

1. If the free energy change is unstable, for example in a system with thermodynamic parameters near a phase change the algorithm may take a long time to converge.
2. If the free energy change is large the algorithm may take a long time to converge.
3. If the free energy change is ill-defined, for example moving from a static structure to a very flexible structure which is best described by an ensemble of states.

On top of these inherent limitations there are additional assumptions within the algorithm itself:

1. The algorithm assumes the energy can be written as a sum of energies from each part of the system. In the case of molecular dynamics this is naturally available because atoms are the smallest entities and they are fixed units.
2. In quantum based methods or probabilistic methods which contain interactions between fields and distributions the sum of energies breaks down. The total energy may be written as an integral or *functional* of a density, but the atom-wise derivation presented in this chapter will break down.
3. The energy terms for pairwise, triplet and quadruplet interactions are in this derivation shared evenly between interactions. This is a parsimonious assumption but not necessarily the best assumption.

Beyond these assumptions the derivation writes terms in the following way

$$\Delta F = \sum_i \Delta f_i, \tag{4.1}$$

where Δf_i were components of a free energy. Care must be taken when using expressions of this form. Questions arise about the rates of convergence of individual parts of the system and the correlation between different parts of the system.

The decomposition provided by AFEP cannot be *additive* between different systems as shown by Mark et al. [87]. Whereas enthalphy and energy terms are additive under the ensemble average

$$E = \langle H \rangle = \langle H_1 + H_2 \rangle = \langle H_1 \rangle + \langle H_2 \rangle = E_1 + E_2$$

the free energy can only be factored as

$$F = kT \log \left\langle e^{\frac{H}{kT}} \right\rangle = kT \log \left\langle e^{\frac{H_1}{kT}} e^{\frac{H_2}{kT}} \right\rangle \rightarrow kT \log \left(\left\langle e^{\frac{H_1}{kT}} \right\rangle \left\langle e^{\frac{H_2}{kT}} \right\rangle \right) = F_1 + F_2$$

if the two terms $e^{\frac{H_1}{kT}}$ and $e^{\frac{H_2}{kT}}$ are uncorrelated [87]. This means a decomposition of free energy in this way will only give sensible results for parts of the system which are detached.

Similar chemical motifs cannot be expected to have the same free energy contribution in different molecules. However, the results can be interpreted *within* a given molecule and used as an empirical tool to highlight atoms which contribute relatively favourably or unfavourably to a free energy change. The formulae in this chapter will be constructed over a set of system components which may be freely split or joined up to the limiting unit of one atoms on the smallest scale and the largest unit of the whole system on the largest scale. The formulae are correct exactly for the largest scale. The results may be come progressively harder to interpret at small scales due to the statements for Mark et al. [87]. It is conceivable that at some intermediate length scale the results have a balance between interpretability and decomposition.

In conclusion care must be taken on the kinds of system this method is applied to. For highly flexible molecules and peptides, the decomposition is unlikely to give a meaningful description of the system. But for smaller, static molecules and for larger partitions it is worth interpreting the results.

4.1.2 System Backgrounds

In this work the atom-wise decomposition is calculated for two types of free energy change; a number of hydration free energies and one binding free energy. The extension to FEP discussed in this study is called atom-wise free energy perturbation (AFEP) because it processes the results of an FEP simulation to give the atom-wise (per atom) breakdown of the free energies. Many extensions have previously been developed for free energy methods [88, 71, 72], and methods for statistically optimizing the results have been developed [17] including BAR [16, 89] which is used in this study to calculate total free energy changes. To the authors' knowledge, none of these methods offer the atom-wise distribution results that AFEP provides.

In general, free energy methods output a single number, a free energy change, that can help predict *whether* a drug will bind, and this has the potential to improve the computational drug design workflow [79]. With collaborative development there is hope that free energy methods will become an important part of the drug design and discovery process [90]. With AFEP one can attempt to determine *why* a drug binds and which parts play which role in the binding. In this chapter we give the mathematical derivation of the AFEP method in full. The results are shown for AFEP directly applied to simulations of methane, methanol, methylamine, methanthiol and caffeine solvated in water to measure a hydration free energy, and a system of human immunodeficiency virus type 1 protease (HIV-1pr) with bound and unbound lopinavir drug molecules to estimate the atom-wise distribution of the binding free energy. The four methane-like molecules are chosen as similar simple examples of common chemical side groups.

Caffeine is a small molecule that is renowned for blocking the adenosine receptor in the human body. Although quite small compared to some medicinal molecules, it is an example of a molecule that might be considered in the field of drug discovery. Caffeine dissolves in water at room temperature, therefore it is then expected to have a negative free energy of hydration. Caffeine was chosen as it is a small and familiar molecule that mixes polar and non-polar groups and can be used to test the predictions of the AFEP method against chemical intuition.

HIV-1pr is one of many constituent proteins in HIV-1. This particular protein is responsible for cutting and cleaving parts of the host, after which the virus will go on to reproduce inside the host using the host's cell based machinery [91]. Specifically the virus uses HIV-1pr to cleave Gag and Gag-Pol, two proteins that are essential for the virus to hijack to synthesize a functional and intact viral particle. If HIV-1pr is successfully blocked with a strongly binding inhibitor, that is, an inhibitor with a highly negative binding free energy, the cutting process will blocked. This will make it impossible for the virus to reproduce, and the infection can subsequently be eliminated by the human body. Lopinavir is a widely used inhibitor of HIV-1pr that is often combined with ritonavir, another drug

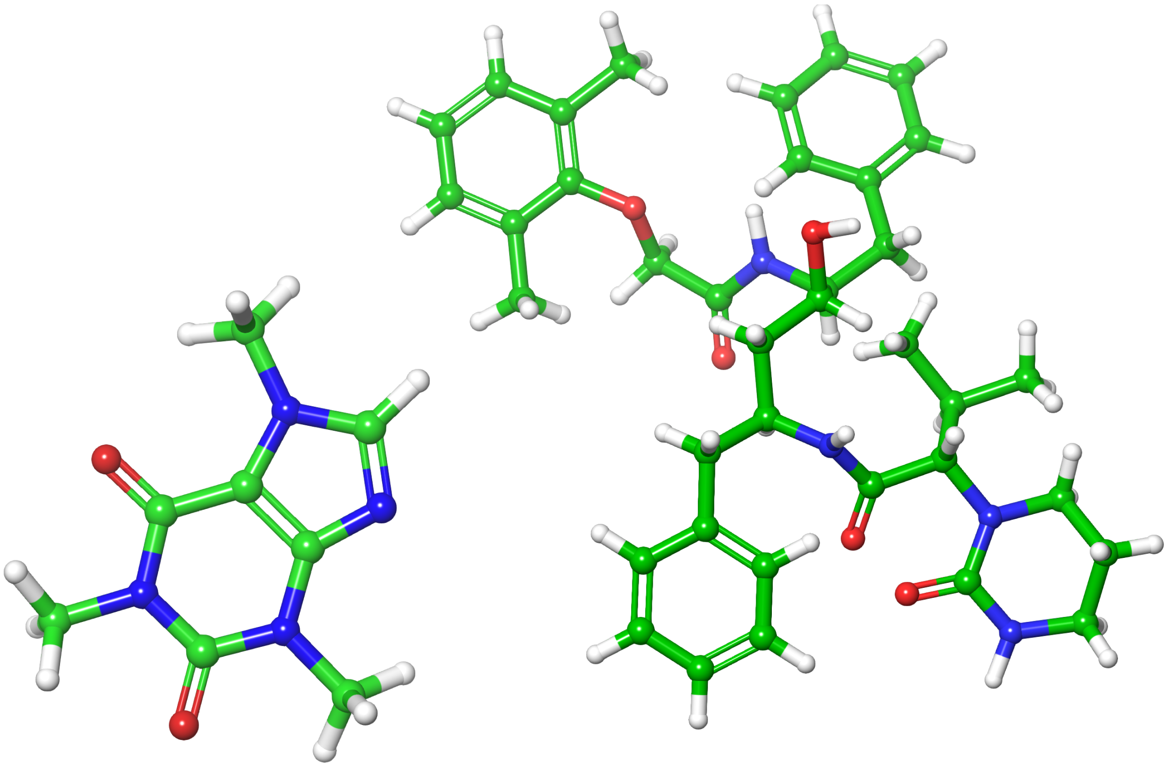


Fig. 4.1 The two molecules investigated in this study. Left: The molecule caffeine. Right: The molecule lopinavir, an inhibitor of HIV-1pr.

molecule [92]. This application was picked because lopinavir binds strongly to the well studied HIV-1pr structure [93].

4.2 Theory

AFEP is a methodology that splits the total free energy change of the system into a sum of atom-wise contributions. The method is based on the Zwanzig equation [15] used in the EXP free energy method to calculate the Helmholtz free energy difference ΔF_{AB} between two thermodynamic states A and B

$$\Delta F_{AB} = F_B - F_A = -\beta^{-1}\log\left\langle e^{-\beta(U_B-U_A)}\right\rangle_A, \tag{4.2}$$

with $\beta = (k_B T)^{-1}$, where k_B is Boltzmann's constant, U_X the total energy of system X, and $\langle\cdot\rangle_A$ represents a state average over system A. From here-on, all state averages are

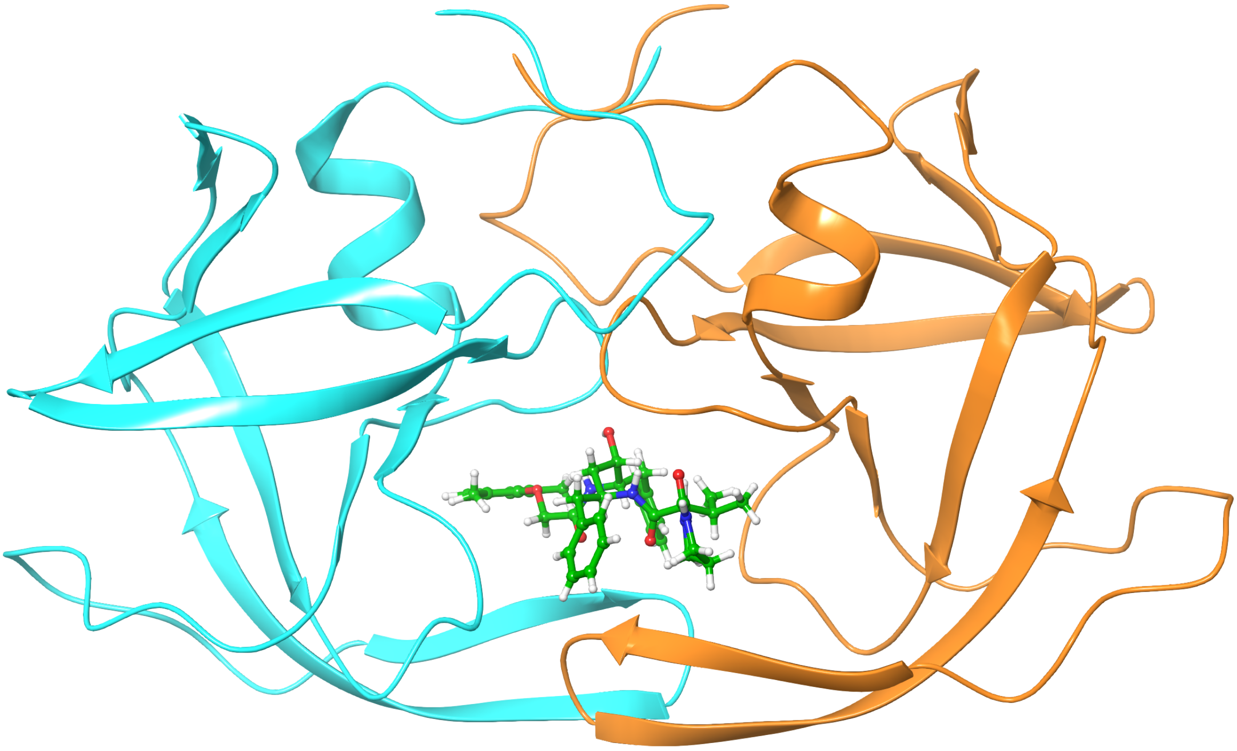


Fig. 4.2 HIV-1pr with lopinavir in the bound state. Lopinavir has been coloured with green carbon atoms. The water and ions in the simulation cell are not shown.

performed over system A without loss of generality, and we will omit the subscript A from equations. Equation 4.12 is a statement that for any free energy differences between systems only the difference in system *energy* is needed to calculate the free energy difference. In practice the free energy difference must be reasonably small because of limited sampling overlap from the start and end states, which would lead to poor convergence. The Zwanzig equation is the starting point for the EXP method and this equation would normally be applied between each pair of subsequent intermediate λ-values to help smoothly measure a larger system change in smaller steps.

For the AFEP method we define the expansion of the difference in total system energy, ΔU, in terms of the N constituent atoms in those systems

$$\Delta U = U_B - U_A = \sum_{k=1}^{N} \Delta u_k, \tag{4.3}$$

$$\Delta u_k = u_{Bk} - u_{Ak}, \tag{4.4}$$

where u_{Xk} is the potential energy associated with atom k in system X. The exact representation of u_{Xk} depends on the force field used to simulate the system and an expression is given later in the text. The goal of AFEP is to write the free energy change as

$$\Delta F_{AB} = \sum_{k=1}^{N} \Delta \mathcal{F}_{AB}(k), \tag{4.5}$$

where $\Delta\mathcal{F}(k)$ is the contribution to the free energy from atom k, and the sum ranges over all of the N atoms in the system. It is worth stating that the free energy cannot usually be separated into a simple sum in this way, because the entropic contributions are formed from correlations of the multi-body density of the system which are not necessarily separable over atoms [41]. The results should not be over interpreted, nor expected to sum together to conserve free energy in an additive way [87]. The special stylization of $\Delta\mathcal{F}(k)$ then represents that this is a 'contribution' from an atom k, but not a true free energy per se. The goal of this work is to derive a mathematical expression for such a free energy contribution, and assess if such an estimate has any meaningful predictive power as an empirical tool.

During the derivation numerous series are used. It should be stressed that these are formal power series of infinite order. As such there is no loss of terms or approximation and the formal power series can be manipulated as a representation of the functional relationship between variable. In much the way that a generating function is used in combinatorics, even though the formal power series may diverge for arguments outside the domain of convergence during manipulations, the coefficients act as a one to one mapping of analytic functions. This is the same as the Mellin transform of a function being unique and invertible if the strip of holomorphy is retained for the inverse transform [94]. With this in mind one should not worry about any approximations from the series expansions being used.

To achieve this atom-wise decomposition we perform the following steps:

- Replace ΔU in the Zwanzig equation (Equation 4.12) with the sum of energies (Equation 4.3). These energies serve as a means of distinguishing which contribution belongs to which atom.

- Expand the exponential term as a Taylor series to infinite order.

- Perform a multinomial expansion on each power term of this Taylor series.

- Observe that the resulting series is a sum of products of differences in atom-wise energy Δu_k. The Δu_k in the products are raised to integer exponents.

- Weight the terms by their exponents and group them into contributions towards each atom in the system. For example a term that looks like $\Delta u_1^2 \Delta u_2 \Delta u_3$ half belongs to atom 1 and one quarter to atoms 2 and 3.

- Write the logarithm in the Zwanzig equation as a Taylor series.

- Perform a multinomial expansion on each term in the Taylor series of the logarithm.

- Observe that this series is also a sum of products of differences in the previous terms. These previous terms are also raised to integer powers.

- Apply the same grouping technique to all products of terms to get individual contributions for atoms

- Find the closed form for these individual groupings that corresponds to the $\Delta\mathcal{F}_{AB}(k)$ in Equation 4.5.

The resulting closed form expression for $\Delta\mathcal{F}_{AB}(k)$ is given by

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1} \frac{\left\langle \frac{\Delta u_a}{\Delta U}\left(e^{-\beta\Delta U} - 1\right)\right\rangle}{\left\langle e^{-\beta\Delta U} - 1\right\rangle} \log\left(1 + \left\langle e^{-\beta\Delta U} - 1\right\rangle\right), \tag{4.6}$$

and by inspection of Equations 4.12 and 4.6, we have

$$-\beta^{-1} \log\left(1 + \left\langle e^{-\beta\Delta U} - 1\right\rangle\right) = -\beta^{-1} \log\left(\left\langle e^{-\beta\Delta U}\right\rangle\right) = \Delta F_{AB}, \tag{4.7}$$

so in effect we have defined a set of weights w_k such that

$$\Delta\mathcal{F}_{AB}(a) = w_a \Delta F_{AB}, \tag{4.8}$$

by equation 4.5 then

$$\sum_{k=1}^{N} w_k = 1, \tag{4.9}$$

and the weight for atom k is defined by

$$w_k = \frac{\left\langle \frac{\Delta u_k}{\Delta U}\left(e^{-\beta\Delta U} - 1\right)\right\rangle}{\left\langle e^{-\beta\Delta U} - 1\right\rangle}. \tag{4.10}$$

The calculation of these weights will allow the calculation of the free energy contribution for each atom in the system. Because the weights are independent of the value of the total free energy change ΔF_{AB} any method desired can be used to calculate ΔF_{AB}, for example a standard BAR FEP simulation [16], which is relatively easy to carry out using modern methods [17]. All that is needed are the energies associated with each atom in the system, u_{Xa}, which will come from the force field used to simulate the system. The full AFEP expression for the estimate of the free energy contribution from atom a is then given by

$$\Delta\mathcal{F}_{AB}(a) = \frac{\left\langle \frac{\Delta u_a}{\Delta U}\left(e^{-\beta\Delta U} - 1\right)\right\rangle}{\left\langle e^{-\beta\Delta U} - 1\right\rangle}\Delta F_{AB} \tag{4.11}$$

One could define other partitioning schemes that give $\Delta\mathcal{F}_{AB}(a) = f(\Delta U, \Delta u_a)\Delta F_{AB}$, for some weight function $f(\Delta U, \Delta u_a)$. A simple example that sums to 1 for all a is $f(\Delta U, \Delta u_a) = \langle \Delta u_a / \Delta U\rangle$. This weight function only considers the energy of each atom and would struggle to predict entropic effects. It is also not rooted in a formal derivation like the AFEP weights are.

4.2.1 Detailed Derivation to Reach Expanded Form of AFEP Contributions

We now derive in full mathematical detail the AFEP expression show in equation 4.11. The reader can skip to section 4.3 to avoid the details. We start from the Zwanzig equation for the Helmholtz free energy difference ΔF_{AB} between two thermodynamic states A and B

$$\Delta F_{AB} = F_B - F_A = -\beta^{-1}\log\left\langle e^{-\beta(U_B - U_A)}\right\rangle_A, \tag{4.12}$$

The ensemble average in equation 4.12 can be written

$$\Delta F_{AB} = -\beta^{-1}\log\left[\frac{1}{Q_A}\int e^{-\beta(U_B(\vec{q}) - U_A(\vec{q}))}e^{-\beta U_A(\vec{q})}\, d\vec{q}\right] \tag{4.13}$$

where Q_A is the partition function of system A

$$Q_A = \int e^{-\beta U_A(\vec{q})}\, d\vec{q} \tag{4.14}$$

and $\vec{q}$ is a $3N$ dimensional vector, where there are N atoms in the system. We want to express this in terms of individual atoms, so we expand the potential energy in terms of these atoms

$$U_A(\vec{q}) = \sum_{k=1}^{N} u_{Ak}(q_k) \tag{4.15}$$

$$U_B(\vec{q}) = \sum_{k=1}^{N} u_{Bk}(q_k) \tag{4.16}$$

$$U_B(\vec{q}) - U_A(\vec{q}) = \sum_{k=1}^{N} \Delta u_k(q_k) \tag{4.17}$$

$$\Delta u_k = u_{Bk}(q_k) - u_{Ak}(q_k) \tag{4.18}$$

where the u_{Xk} is the potential energy function of atom k in system X. We still want to retain the reference to system A so we only expand the difference term

$$\Delta F_{AB} = -\beta^{-1} \log \left[\frac{1}{Q_A} \int e^{-\beta(\sum_{k=1}^{N} u_{Bk}(q_k) - u_{Ak}(q_k))} e^{-\beta U_A(\vec{q})} \, d\vec{q} \right] \tag{4.19}$$

Our approach is to expand the whole exponentiated sum in the equation

$$\Delta F_{AB} = -\beta^{-1} \log \left[\frac{1}{Q_A} \int e^{-\beta \sum_{k=1}^{N} \Delta u_k} e^{-\beta U_A(\vec{q})} \, d\vec{q} \right] \tag{4.20}$$

we can use the absolutely convergent Maclaurin Series for $e^{-\alpha x}$ which gives

$$e^{-\beta \sum_{k=1}^{N} \Delta u_k} = 1 - \beta \sum_{k=1}^{N} \Delta u_k + \frac{\beta^2}{2!} \left(\sum_{k=1}^{N} \Delta u_k \right)^2 - \frac{\beta^3}{3!} \left(\sum_{k=1}^{N} \Delta u_k \right)^3 + \cdots \tag{4.21}$$

which when written in summation form is

$$e^{-\beta \sum_{k=1}^{N} \Delta u_k} = \sum_{m=0}^{\infty} \frac{(-1)^m \beta^m}{m!} \left(\sum_{k=1}^{N} \Delta u_k \right)^m \tag{4.22}$$

The multinomial expansion identity is

$$(x_1 + x_2 + \cdots + x_N)^m = \sum_{k_1+k_2+\cdots+k_N=m} \frac{m!}{k_1! k_2! \cdots k_N!} \prod_{t=1}^{N} x_t^{k_t} \tag{4.23}$$

where the sum is over all sets of indices $k_t \in [0, m]$ that sum to m. We can then write

$$e^{-\beta \sum_{k=1}^{N} \Delta u_k} = \sum_{m=0}^{\infty} \frac{(-1)^m \beta^m}{m!} \sum_{k_1+k_2+\cdots+k_N=m} \frac{m!}{k_1! k_2! \cdots k_N!} \prod_{t=1}^{N} \Delta u_t^{k_t} \tag{4.24}$$

which can be reduced to

$$e^{-\beta\Sigma_{k=1}^{N}\Delta u_k} = \sum_{m=0}^{\infty}\left(\sum_{k_1+k_2+\cdots+k_N=m}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_N!}\prod_{t=1}^{N}\Delta u_t^{k_t}\right) \tag{4.25}$$

This is the point at which we implement the grouping scheme mentioned earlier. We introduce a weighting to each term such that terms with higher powers of the energy difference associated with atom a have a larger share in the term. It is clear that all of the summation terms in equation 4.25 are products of the Δu_t raised to different powers k_t with different coefficients. Supposing we are only interested in some atom with index a, we can try to separate all the terms that contain atom a from the other terms and sum them together. We can call this sum $J(a)$. $J(a)$ will contain cross terms because the free energy is not directly separable in this manner; that is, terms which also contain the Δu_t for atoms other than a. To deal with these cross terms we can partition (or weight) each term according to atoms that contribute to them.

Here are two examples of the weighting scheme: A term that looks like $2\Delta u_1\Delta u_2$, is half 'owned' by atom 1 and half by atom 2, so the $J(1)$ and $J(2)$ terms will each contain a contribution of $\Delta u_1\Delta u_2$. A term that looks like $24\Delta u_1^2\Delta u_2\Delta u_3$, is 2/4 'owned' by atom 1, because of the power 2, and 1/4 each by atoms 2 and 3, so the $J(1)$ term would have $12\Delta u_1^2\Delta u_2\Delta u_3$, and $J(2)$ and $J(3)$ each get $6\Delta u_1^2\Delta u_2\Delta u_3$. When written mathematically we prepare the exponential for the separation by removing the constant term of 1

$$e^{-\beta\Sigma_{k=1}^{N}\Delta u_k} = \sum_{m=0}^{\infty}\left(\sum_{k_1+k_2+\cdots+k_N=m}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_N!}\prod_{t=1}^{N}\Delta u_t^{k_t}\right) \tag{4.26}$$

$$e^{-\beta\Sigma_{k=1}^{N}\Delta u_k} = 1+\sum_{m=1}^{\infty}\left(\sum_{k_1+k_2+\cdots+k_N=m}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_N!}\prod_{t=1}^{N}\Delta u_t^{k_t}\right) \tag{4.27}$$

this will become useful later for expanding the logarithm in the Zwanzig equation. Meanwhile we can apply the grouping rule above by introducing a factor of k_a/m and summing over all a giving

$$e^{-\beta\Sigma_{k=1}^{N}\Delta u_k} = 1+\sum_{a=1}^{N}\sum_{m=1}^{\infty}\left(\sum_{k_1+k_2+\cdots+k_N=m}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_N!}\frac{k_a}{m}\prod_{t=1}^{N}\Delta u_t^{k_t}\right) \tag{4.28}$$

and then define the $J(a)$ function as

$$J(a) := \sum_{m=1}^{\infty}\left(\sum_{k_1+k_2+\cdots+k_N=m}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_N!}\frac{k_a}{m}\prod_{t=1}^{N}\Delta u_t^{k_t}\right) \tag{4.29}$$

which gives

$$e^{-\beta\sum_{k=1}^{N}\Delta u_k} = 1 + \sum_{k=1}^{N} J(k). \tag{4.30}$$

We can now derive the closed form of $J(a)$. We will prove that $J(a)$ as defined in equation 4.29 can also be written as

$$J(a) = \frac{\Delta u_a}{\sum_{k=1}^{N}\Delta u_k}\left(e^{-\beta\sum_{k=1}^{N}\Delta u_k} - 1\right). \tag{4.31}$$

We can remove the m from inside the multinomial sum, and separate out the factorial of k_a, and the Δu_a term in the product to give

$$J(a) = \sum_{m=1}^{\infty}\frac{1}{m}\left(\sum_{k_1+k_2+\cdots+k_N=m}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_a!\cdots k_N!}k_a\Delta u_a^{k_a}\prod_{t=1,t\neq a}^{N}\Delta u_t^{k_t}\right). \tag{4.32}$$

Now we use the principle of distinguished element on k_a; if $k_a = 0$ the whole term is zero and otherwise we know k_a is in the range $1,\cdots,m$, due to the conditions on the sum. If we take each case that k_a is in the range 1 to m and fully separate all terms out the inner part of the above sum in equation 4.32 becomes

$$\left(\sum_{k_1+k_2+\cdots+k_N=m}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_a!\cdots k_N!}k_a\Delta u_a^{k_a}\prod_{t=1}^{N}\Delta u_t^{k_t}\right) =$$
$$\sum_{k_a=1}^{m}\frac{k_a\Delta u_a^{k_a}}{k_a!}\left(\sum_{k_1+\cdots+k_{a-1}+k_{a+1}+\cdots+k_N=m-k_a}\frac{(-1)^m\beta^m}{k_1!k_2!\cdots k_{a-1}!k_{a+1}!\cdots k_N!}\prod_{t=1,t\neq a}^{N}\Delta u_t^{k_t}\right) \tag{4.33}$$

We can divide the top and bottom by $(m-k_a)!$ and pull out the other m-dependent terms

$$= \sum_{k_a=1}^{m}\frac{(-1)^m\beta^m k_a\Delta u_a^{k_a}}{(m-k_a)!k_a!}\left(\sum_{k_1+\cdots+k_{a-1}+k_{a+1}+\cdots+k_N=m-k_a}\frac{(m-k_a)!}{k_1!k_2!\cdots k_{a-1}!k_{a+1}!\cdots k_N!}\prod_{t=1,t\neq a}^{N}\Delta u_t^{k_t}\right) \tag{4.34}$$

We can simply relabel the index $k_a = s$

$$= \sum_{s=1}^{m}\frac{(-1)^m\beta^m s\Delta u_a^{s}}{(m-s)!s!}\left(\sum_{k_1+\cdots+k_{a-1}+k_{a+1}+\cdots+k_N=m-s}\frac{(m-s)!}{k_1!k_2!\cdots k_{a-1}!k_{a+1}!\cdots k_N!}\prod_{t=1,t\neq a}^{N}\Delta u_t^{k_t}\right) \tag{4.35}$$

We then compare and substitute this to the bracket expanded with the multinomial theorem

$$\left(\sum_{k=1,k\neq a}^{N}\Delta u_k\right)^{m-s} = \sum_{m-s=\sum_{i=1,i\neq a}^{N}k_i}\frac{(m-s)!}{\prod_{i=1,i\neq a}^{N}k_i!}\prod_{t=1,t\neq a}^{N}x_t^{k_t} \tag{4.36}$$

So we can write a new form for $J(a)$ as

$$J(a) = \sum_{m=1}^{\infty} \frac{(-1)^m \beta^m}{m} \sum_{s=1}^{m} \frac{1}{(m-s)!} \frac{s\Delta u_a^s}{s!} \left(\sum_{k=1,k\neq a}^{N} \Delta u_k \right)^{m-s} \tag{4.37}$$

For convenience label the sum of the energy difference terms X

$$X = \sum_{k=1,k\neq a}^{N} \Delta u_k \tag{4.38}$$

then

$$J(a) = \sum_{m=1}^{\infty} \frac{(-1)^m \beta^m}{m} \sum_{s=1}^{m} \frac{1}{(m-s)!} \frac{s\Delta u_a^s}{s!} X^{m-s} \tag{4.39}$$

Also, as $s > 0$, we can take out a factor of Δu_a

$$J(a) = \sum_{m=1}^{\infty} \frac{(-1)^m \beta^m \Delta u_a}{m} \sum_{s=1}^{m} \frac{X^{m-s}}{(m-s)!} \frac{\Delta u_a^{s-1}}{(s-1)!} \tag{4.40}$$

We have that

$$\sum_{s=1}^{m} \frac{X^{m-s}}{(m-s)!} \frac{\Delta u_a^{s-1}}{(s-1)!} = \frac{(X+\Delta u_a)^{m-1}}{(m-1)!} \tag{4.41}$$

which can be proved by a relabelling of indices from the binomial expansion

$$(a+b)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} a^{n-k} b^k \tag{4.42}$$

under the mappings

$$a \to X, \tag{4.43}$$

$$b \to \Delta u_a, \tag{4.44}$$

$$n \to m-1, \tag{4.45}$$

$$k \to s-1. \tag{4.46}$$

Using this identity gives

$$J(a) = \Delta u_a \sum_{m=1}^{\infty} \frac{(-1)^m \beta^m}{m} \frac{(X + \Delta u_a)^{m-1}}{(m-1)!} \tag{4.47}$$

$$J(a) = \Delta u_a \sum_{m=1}^{\infty} \frac{(-1)^m \beta^m}{m!} (X + \Delta u_a)^{m-1} \tag{4.48}$$

$$J(a) = \Delta u_a \sum_{m=1}^{\infty} \frac{(-1)^m \beta^m}{m!} \frac{(X + \Delta u_a)^m}{(X + \Delta u_a)} \tag{4.49}$$

$$J(a) = \frac{\Delta u_a}{(X + \Delta u_a)} \sum_{m=1}^{\infty} \frac{(-1)^m \beta^m}{m!} (X + \Delta u_a)^m \tag{4.50}$$

$$J(a) = \frac{\Delta u_a}{(X + \Delta u_a)} (e^{-\beta(X+\Delta u_a)} - 1) \tag{4.51}$$

Finally we have that

$$X + \Delta u_a = \sum_{k=1}^{N} \Delta u_k \tag{4.52}$$

$$J(a) = \frac{\Delta u_a}{\sum_{k=1}^{N} \Delta u_k} \left(e^{-\beta \sum_{k=1}^{N} \Delta u_k} - 1 \right) \tag{4.53}$$

which was the desired result.

From the Zwanzig equation, we have

$$\Delta F_{AB} = -\beta^{-1} \log \left[\frac{1}{Q_A} \int e^{-\beta \sum_{k=1}^{N} \Delta u_k} e^{-\beta U_A(\vec{q})} \, d\vec{q} \right] \tag{4.54}$$

and from the definition of $J(a)$ we showed earlier that

$$e^{-\beta \sum_{k=1}^{N} \Delta u_k} = 1 + \sum_{k=1}^{N} J(k) \tag{4.55}$$

We can use this in the free energy equation

$$-\beta \Delta F_{AB} = \log \left[\frac{1}{Q_A} \int \left(1 + \sum_{k=1}^{N} J(k) \right) e^{-\beta U_A(\vec{q})} \, d\vec{q} \right] \tag{4.56}$$

which due to linearity is

$$-\beta \Delta F_{AB} = \log \left[1 + \sum_{k=1}^{N} \frac{1}{Q_A} \int J(k) e^{-\beta U_A(\vec{q})} \, d\vec{q} \right] \tag{4.57}$$

which is then the sum of the ensemble averages of the terms by definition

$$-\beta\Delta F_{AB} = \log\left[1 + \sum_{k=1}^{N} \langle J(k)\rangle\right] \tag{4.58}$$

There is not much that can be done with the logarithm of a sum, but as the 1 was conveniently left there from previous steps we can use another Taylor expansion on the logarithm this time to give a formal series:

$$-\beta\Delta F_{AB} = \sum_{k=1}^{N} \langle J(k)\rangle - \frac{1}{2}\left(\sum_{k=1}^{N} \langle J(k)\rangle\right)^2 + \frac{1}{3}\left(\sum_{k=1}^{N} \langle J(k)\rangle\right)^3 - \cdots \tag{4.59}$$

or written in summation notation

$$-\beta\Delta F_{AB} = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}\left(\sum_{k=1}^{N} \langle J(k)\rangle\right)^m \tag{4.60}$$

we can reuse the multinomial expansion rules for the powers of the sums of $J(a)$ as used in section 4.2.1, and get

$$\Delta F_{AB}(a) = -\beta^{-1} \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \sum_{k_1+k_2+\cdots+k_N=m} \frac{m!}{k_1!k_2!\cdots k_N!} \prod_{t=1}^{N} \langle J(t)\rangle^{k_t} \tag{4.61}$$

Similarly to the methods used in 4.2.1 we can define contributions

$$\Delta F_{AB} = \sum_{a=1}^{N} \Delta\mathcal{F}_{AB}(a) \tag{4.62}$$

where using the regrouping trick by including a factor of k_a/m gives the definition of the atom-wise free energy contributions

$$\Delta\mathcal{F}_{AB}(a) := -\beta^{-1} \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \sum_{k_1+k_2+\cdots+k_N=m} \frac{m!}{k_1!k_2!\cdots k_N!} \frac{k_a}{m} \prod_{t=1}^{N} \langle J(t)\rangle^{k_t} \tag{4.63}$$

at this stage we have an atom-wise description of ΔF_{AB}. We now find a tractable closed form for the $\Delta\mathcal{F}_{AB}(a)$. We can then prove that equation 4.63 can also be written as

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1} \frac{\langle J(a)\rangle}{\sum_{k=1}^{N} \langle J(k)\rangle} \log\left(1 + \sum_{k=1}^{N} \langle J(k)\rangle\right). \tag{4.64}$$

As in the previous section, we will separate out all m-dependent terms and all k_a terms, and divide top and bottom by $(m-k_a)!$. For compactness the condition in the multinomial sum can be rewritten in summation notation

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1} \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{m!}{m} \sum_{k_a=1}^{m} \frac{k_a \langle J(a)\rangle^{k_a}}{k_a!(m-k_a)!} \sum_{m-k_a=\sum_{i=1,i\neq a}^{N} k_i} \frac{(m-k_a)!}{\prod_{i=1,i\neq a}^{N} k_i!} \prod_{t=1,t\neq a}^{N} \langle J(t)\rangle^{k_t}. \tag{4.65}$$

Similarly to the previous expansion, labelling $k_a = s$ we may use the multinomial expansion identity

$$\left(\sum_{k=1,k\neq a}^{N} \langle J(k)\rangle\right)^{m-s} = \sum_{m-s=\sum_{i=1,i\neq a}^{N} k_i} \frac{(m-s)!}{\prod_{i=1,i\neq a}^{N} k_i!} \prod_{t=1,t\neq a}^{N} \langle J(t)\rangle^{k_t} \tag{4.66}$$

to give

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1} \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{m!}{m} \sum_{s=1}^{m} \frac{\langle J(a)\rangle^{s}}{(s-1)!(m-s)!} \left(\sum_{k=1,k\neq a}^{N} \langle J(k)\rangle\right)^{m-s} \tag{4.67}$$

for convenience label the sum of terms as a separate variable

$$Y = \sum_{k=1,k\neq a}^{N} \langle J(k)\rangle \tag{4.68}$$

which gives

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1} \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{m!}{m} \sum_{s=1}^{m} \frac{\langle J(a)\rangle^{s}\, Y^{m-s}}{(s-1)!(m-s)!}. \tag{4.69}$$

As the summation index $s > 0$ we can remove a factor of $\langle J(a)\rangle$,

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1} \langle J(a)\rangle \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{m!}{m} \sum_{s=1}^{m} \frac{\langle J(a)\rangle^{s-1}\, Y^{m-s}}{(s-1)!(m-s)!} \tag{4.70}$$

and then use the identity

$$\sum_{s=1}^{m} \frac{Y^{m-s}}{(m-s)!} \frac{\langle J(a)\rangle^{s-1}}{(s-1)!} = \frac{(Y+\langle J(a)\rangle)^{m-1}}{(m-1)!} \tag{4.71}$$

which again can be obtained from the binomial expansion

$$(a+b)^n = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} a^{n-k} b^k \tag{4.72}$$

under the mappings

$$a \to Y, \tag{4.73}$$

$$b \to \langle J(a)\rangle\,, \tag{4.74}$$

$$n \to m-1, \tag{4.75}$$

$$k \to s-1 \tag{4.76}$$

This gives us

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1}\langle J(a)\rangle \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}\frac{m!}{m}\frac{(Y+\langle J(a)\rangle)^{m-1}}{(m-1)!} \tag{4.77}$$

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1}\langle J(a)\rangle \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}\frac{m!}{m!}\frac{(Y+\langle J(a)\rangle)^{m}}{(Y+\langle J(a)\rangle)} \tag{4.78}$$

$$\Delta\mathcal{F}_{AB}(a) = -\frac{\beta^{-1}\langle J(a)\rangle}{(Y+\langle J(a)\rangle)} \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}(Y+\langle J(a)\rangle)^{m} \tag{4.79}$$

$$\Delta\mathcal{F}_{AB}(a) = -\frac{\beta^{-1}\langle J(a)\rangle}{(Y+\langle J(a)\rangle)}\log(1+Y+\langle J(a)\rangle) \tag{4.80}$$

$$\tag{4.81}$$

Then we have that

$$Y+\langle J(a)\rangle = \sum_{k=1}^{N}\langle J(k)\rangle \tag{4.82}$$

giving

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1}\frac{\langle J(a)\rangle}{\sum_{k=1}^{N}\langle J(k)\rangle}\log\left(1+\sum_{k=1}^{N}\langle J(k)\rangle\right) \tag{4.83}$$

which was the intermediate formula to be proved. From here the aim is to express $\Delta\mathcal{F}_{AB}(a)$ as a function of the Δu_k. Combining this and the closed form of $J(a)$

$$J(a) = \frac{\Delta u_a}{\sum_{k=1}^{N}\Delta u_k}\left(e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right) \tag{4.84}$$

as well as the definition

$$e^{-\beta\sum_{k=1}^{N}\Delta u_k} = 1+\sum_{k=1}^{N}J(k) \tag{4.85}$$

gives the final form with no occurrence of $J(a)$:

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1}\frac{\left\langle \frac{\Delta u_a}{\sum_{k=1}^{N}\Delta u_k}\left(e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right)\right\rangle}{\left\langle e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right\rangle}\log\left(1+\left\langle e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right\rangle\right), \tag{4.86}$$

this is equal to

$$\Delta\mathcal{F}_{AB}(a) = -\beta^{-1}\frac{\left\langle \frac{\Delta u_a}{\sum_{k=1}^{N}\Delta u_k}\left(e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right)\right\rangle}{\left\langle e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right\rangle}\log\left(\left\langle e^{-\beta\sum_{k=1}^{N}\Delta u_k}\right\rangle\right), \tag{4.87}$$

which, using the Zwanzig equation can be written

$$\Delta\mathcal{F}_{AB}(a) = \frac{\left\langle \frac{\Delta u_a}{\sum_{k=1}^{N}\Delta u_k}\left(e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right)\right\rangle}{\left\langle e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right\rangle}\Delta F_{AB}. \tag{4.88}$$

This is the full atom-wise FEP expression. Summing over all a then gives

$$\sum_{a=1}^{N}\Delta\mathcal{F}_{AB}(a) = \Delta F_{AB} = \sum_{a=1}^{N}\frac{\left\langle \frac{\Delta u_a}{\sum_{k=1}^{N}\Delta u_k}\left(e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right)\right\rangle}{\left\langle e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right\rangle}\Delta F_{AB}, \tag{4.89}$$

so dividing through by the free energy gives a set of weights

$$\sum_{a=1}^{N}\frac{\left\langle \frac{\Delta u_a}{\sum_{k=1}^{N}\Delta u_k}\left(e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right)\right\rangle}{\left\langle e^{-\beta\sum_{k=1}^{N}\Delta u_k}-1\right\rangle} = 1 \tag{4.90}$$

then the correct form of the weights is simply

$$w_a = \frac{\left\langle \frac{\Delta u_a}{\Delta U}\left(e^{-\beta\Delta U}-1\right)\right\rangle}{\left\langle e^{-\beta\Delta U}-1\right\rangle} \tag{4.91}$$

and the free energy change ΔF_{AB} can be calculated using standard techniques.

4.3 Simulation Details

Full information about the MD simulations can be found in appendix B. All AFEP results for all systems were taken from 1800, 0.2 ps frames. The first 200 frames were treated as equilibration frames for each lambda window.

In the MD simulations, the total system energy comes from electrostatic, Lennard-Jones, bonded, Urey-Bradley, angle, dihedral and improper terms in the force field, which is a set of parameters for each type of atom in the system. The CHARMM36 force field was used in these simulations [95]. This includes the parametrisation of each amino acid in the protein and the explicit water molecules around the protein in the simulation. Fixed bonds and bond angles were used, meaning certain terms are not included in the energy calculations. To get the atom-wise energies u_{Xa} needed to construct the AFEP weights (Equation 4.10) from the MD simulation we must define the connection between the energy and the force field being used. This is given by

$$u_{Xa} = \frac{1}{2}(u_{\text{electrostatic}} + u_{\text{Lennard-Jones}} + u_{\text{bonded}} + u_{\text{Urey-Bradley}}) + \frac{1}{3}u_{\text{angle}} + \frac{1}{4}(u_{\text{dihedral}} + u_{\text{improper}}) \tag{4.92}$$

where each term is the sum of all appropriate interactions containing atom a. This energy is a shared contribution from all contributions to the energy contained in Hamiltonian X (the Hamiltonian will change as the FEP variable, λ, changes). For simulations with fixed bond lengths and fixed angles, which are common in simulations of biological molecules, we have $u_{\text{bonded}} = u_{\text{angle}} = 0$. Equation 4.92 is a statement that for all possible types of interaction, if a number of atoms are involved in an interaction, the energy of that interaction is shared equally between all of the atoms involved. This is an approximation and variations to this scheme could potentially be made. We believe sharing the terms equally is the least biased scheme. This work does not include intra-molecular terms between the nearest and next nearest bonded neighbors (1-2 and 1-3 terms).

4.3.1 Methane-like Molecules and Caffeine Simulations Details

Caffeine and the four methane-like molecules were simulated with explicit TIP3P [96] water with a lambda schedule that starts with the molecule completely non-interacting with the water ($\lambda = 0$) and ends with full interaction between the molecule and the water ($\lambda = 1$). The resulting free energy change is then the free energy of hydration for the molecule. For the methane-like molecules these free energies are show in table 4.2 along with experimental values [97] [98]. There is experimental data for the caffeine free energy measurement taken from the FreeSolv database [83] and a similar computational measurement was performed by Mobley et al. [83]. These results are presented in table 4.4, which compare experimental and computational results. The force field parameters for all molecules were generated with the ParamChem CGenFF software [99, 100]. This software picks the best atom types and charges to represent a given molecule according to its bonding and connectivity similarity with a test set of molecules. The original chemical structure and topology files for caffeine were taken from the ZINC database entry ZINC1084 [101]. The systems were first equilibrated under NVT conditions and then NpT conditions to find a suitable density and subsequently run under NpT conditions for the main trajectory data collection. 32 λ-windows were used and an MD trajectory was collected for each value of λ. Further details of the MD simulations are given in appendix B.

4.3.2 HIV-1pr with Lopinavir Simulations

The structures used for the simulation of lopinavir in the binding pose of HIV-1pr are from The Protein Data Bank, reference 2Q5K [102, 103]. For convenience of processing the entire free energy calculation was performed in one cell. In this cell there are two copies of the drug lopinavir. One is in solution and is fully interacting at the beginning of the simulation, and the other is restrained in the natural binding pose in the protein from the PDB data [102] and is non-interacting at the beginning of the simulation. The copy in the solvent was placed 42 Å from the protein which was deemed to be suitably far to prevent the majority of interactions with the protein. The simulations had three stages of lambda schedule, as is common for such binding free energy calculations [86]. The first stage is to turn on the Lennard-Jones interactions of lopinavir in the binding site; this creates a cavity in the protein and the interactions are turned on very slowly at first. This is to prevent the so called ‘end point catastrophe’ in an FEP simulation [17]. The second stage is to turn the Coulombic charge-based interactions of the solvent-based copy of lopinavir off, and to turn the charge-based interactions of the binding site ligand on. The third stage is to turn the Lennard-Jones interactions of the solvent copy of lopinavir off. This leaves the end point of the simulation representing the fully interacting and bound inhibitor in the protein and the copy of the inhibitor in the solution fully uninteracting. To keep the inhibitor in the binding site throughout the simulation, a tethering force is used between three atoms in the protein and one atom in the drug molecule. These restraints can be seen in figure 4.3. There is a free energy term that must be considered from this tethering which can be decomposed into two parts, the unnatural energy associated with the unphysical tethering force and the unnatural entropic term associated with prohibiting the ligand from accessing the full volume of the simulation cell. The first contribution is measured using a thermodynamic integration (TI) [13, 25] of the system by varying the strength of each degree of freedom in the tethering forces. The strength of these contributions can then be calculated and corrected for. The latter contribution is analytically calculated using a derived equation, the expression for this correction is given by Equation 4.93. Further details of the MD simulation of HIV-1pr with lopinavir are given in appendix B.

4.3.3 Restraints Correction Factor

A three term restraint across four atoms is used to hold lopinavir in the binding pose in the binding site of HIV-1pr. Three atoms from the binding site and one atom in the ligand are connected with a dihedral term across all four atoms, an angular term across two of the protein atoms and the ligand atom and a separation term from one of the atoms in the protein to the atom in the ligand. This allows the ligand to adopt different orientational poses during the simulation, but keeps it fixed to the protein such that it does not wander through the partially interacting protein when the ligand and protein are only weakly interacting. An analytic correction term ΔA_r can be derived to compensate for the restricted environment the ligand resides in, as there is an entropic cost of the ligand not being free to wander around the simulation cell. The strength of the restraints is chosen to be suitably weak as not to affect the AFEP results.

Following the method used in the paper by Boresch et al. [86] who analytically calculated this entropic cost for a six-point restraint of the ligand in the binding site of the protein, a similar expression was derived for a three-point restraint. This gives the free energy correction term associated with constraining the ligand when interactions are turned off as

$$\Delta A_r = -k_B T \log \left[\frac{8\pi^2 V \sqrt{K_r K_\theta K_\phi}}{r_{aA,0}^2 \sin\theta_{A,0} (2\pi k_B T)^{3/2}} \right], \tag{4.93}$$

where V is the standard system volume for 1 molar concentration, K_r is the strength constant of the distance constraint, K_θ is the strength constant of the angular constraint and K_ϕ is the strength constant of the dihedral constraint. $r_{aA,0}$ is the equilibrium distance between the drug atom and the protein atom. $\theta_{A,0}$ is the equilibrium angle between the atoms with the angular constraint. In addition to this entropic term, there is the energetic cost of the restraints themselves. This should be made as small as necessary to not perturb

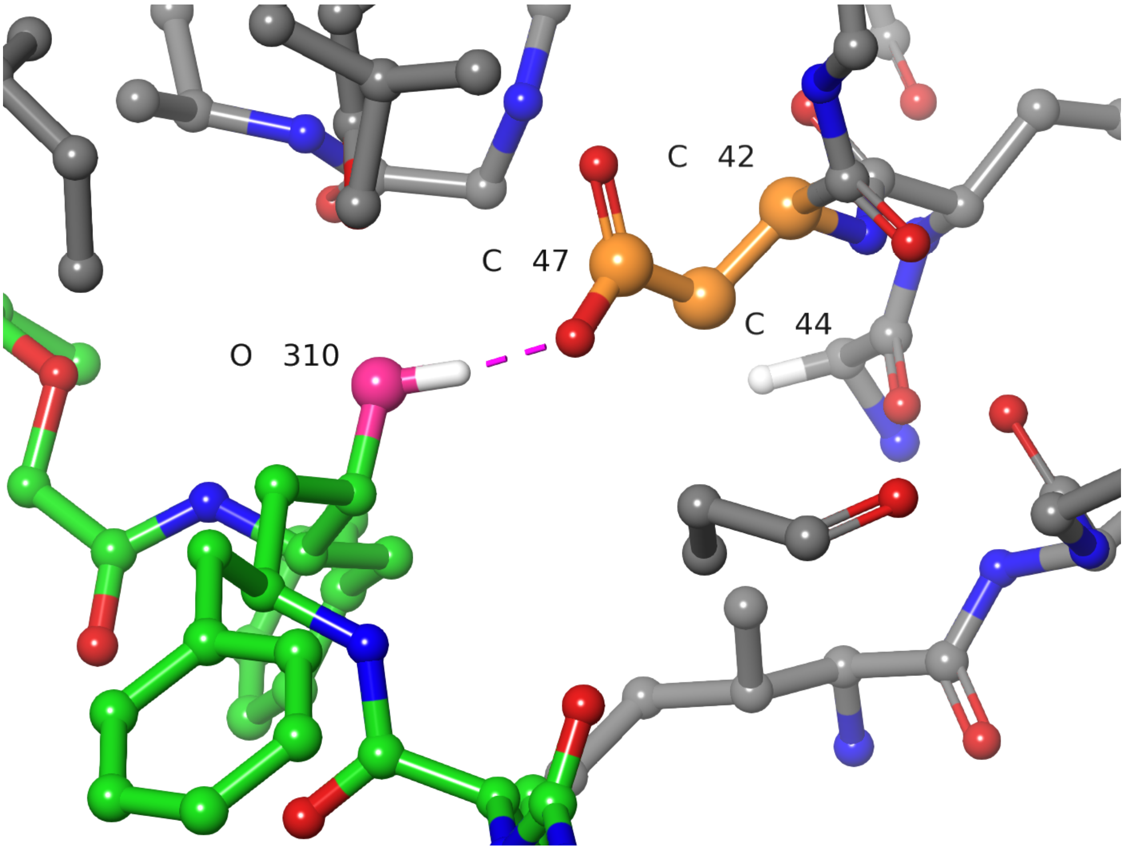


Fig. 4.3 The tethering of lopinavir (green carbons) to the protein (gray carbons). There is a strong hydrogen bond between H and O310 (dashed pink). Four atoms were chosen to apply the restraints, O310 the ligand (pink oxygen) and C42, C47 and C44 in HIV-1pr (orange carbons). There is one dihedral restraint through all four, one angular restraint through O310, C47 and C44 and one distance restraint between O310 and C47.

the system unnaturally when interactions are at the normal level. The restraint must also be strong enough to hold the ligand in the natural binding pose when interactions with the rest of the system are switched off. A set of TI measurements were taken out to ensure the contribution from the restraints when interactions were turned on was close to 0 kcal/mol, which would indicate that it is not interfering with the dynamics of the fully interacting system. The total free energy associated with the restraint was found to be 0.6113 kcal/mol. This value is relatively small and is deemed to be acceptable for the purposes of this simulation.

Figure 4.3 shows a small section of the system to demonstrate the restraints used. Table 4.1 shows the values of the constants used in the simulation, which are then used to calculate the correction factor (Equation 4.93).

Parameter	Value	Units
K_r	0.5	kcal/mol/Å^2
K_θ	0.005	kcal/mol/rad^2
K_ϕ	0.005	kcal/mol
V	1660.5	Å^3
$r_{aA,0}$	3.318	Å
$\theta_{A,0}$	2.07	rad
T	298	K

Table 4.1 A table of the input parameters for the calculation in equation 4.93. The equilibrium parameters are found from the natural binding pose in experimental data.

Molecule	Experimental	ΔF BAR FEP ± Statistical Error
Methane	2.00 [97]	2.45 ± 0.03
Methanol	-5.10[98]	-4.49 ± 0.05
Methylamine	-4.57[98]	-3.41 ± 0.05
Methanethiol	-1.20[97]	-0.06 ± 0.04

Table 4.2 FEP calculations for methane, methanol, methylamine and methanethiol compared to experimental results. Units for all energies are kcal/mol. References for experimental values are given.

4.4 Results

4.4.1 Methane, Methanol, Methylamine and Methanethiol

FEP calculations were performed for methane, methanol, methylamine and methanethiol to demonstrate the AFEP method for similar small molecules with different side chains. Table 4.2 shows the total hydration free energy change for these four molecules. The calculated values agree with the experimental values reasonably well. Table 4.3 shows that atom-wise contributions for the molecules as calculated using AFEP.

4.4.2 Caffeine

A free energy calculation was performed for caffeine solvated in water, AFEP was then applied to the trajectory information to produce an atom-wise breakdown of the hydration free energy. Table 4.4 shows the total free energy calculated from a standard FEP simulation compared with experiment and a similar computation.

Methane	ΔF_{AFEP}	Methanol	ΔF_{AFEP}	Methylamine	ΔF_{AFEP}	Methanethiol	ΔF_{AFEP}
C	2.21	C	-0.07	C	-0.04	C	-0.44
H1	-0.26	O	-4.79	H1	0.34	H1	0.29
H2	-0.24	H1	0.38	H2	0.34	H2	0.35
H3	-0.25	H2	0.60	H3	0.40	H3	0.32
H4	-0.23	H3	0.52	N	-4.30	S	-0.82
-	-	H4 (O)	1.11	H4 (N)	1.05	H4 (S)	0.28
-	-	-	-	H5 (N)	0.51	-	-

Table 4.3 AFEP free energy contributions for each atom in methane, methanol, methylamine and methanethiol. The units of all free energy changes are kcal/mol.

Molecule	Mobley et al. [83]	Experimental [83]	ΔF BAR FEP
Caffeine	-17.62 ± 0.04	-12.64 ± 0.74	-18.53 ± 0.17

Table 4.4 Total free energy values computed by Mobley et al [83], experimentally measured values and values computed with a BAR FEP simulation all in units of kcal/mol. The uncertainty quoted in the BAR FEP table is the statistical error from the BAR routine also in units of kcal/mol.

Figure 4.4 shows the atom-wise breakdown of free energies associated with each atom in the caffeine molecule in units of kcal/mol. All of the water molecules in the simulation have been partitioned into one group because they are indistinguishable. For the caffeine simulation the free energy contribution from all the water molecules sums to exactly $\Delta F/2$. This arises because the atom-wise energy is defined to be shared evenly across interactions in equation 4.92. When partitioned in this way the free energy contribution per water molecule is less meaningful and specific water molecules of chemical interest, ideally, should not be partitioned into the bulk. Contributions from water molecules within certain shells from the solute could be analysed by partitioning them carefully according to their distance from the solute.

Some of the atoms in a caffeine molecule are effectively in the same chemical environment. For example H1, H2 and H3 (as labeled in figure 4.4) are in the same environment because C1 could rotate freely about its bond with N1. In the limit of perfect sampling we would expect these three hydrogen atoms to have the same free energy contribution. In practice it may take a long time in an MD simulation to observe the conformational changes required to sample this. This problem is common in many forms of MD. Obvious symmetries can be input manually by grouping atoms together; however, the user should be careful not to insert fictitious symmetries. In the case of the MD simulations performed in this study, fixed bonds and angles prevent the rotation, and the weights for these atoms are not

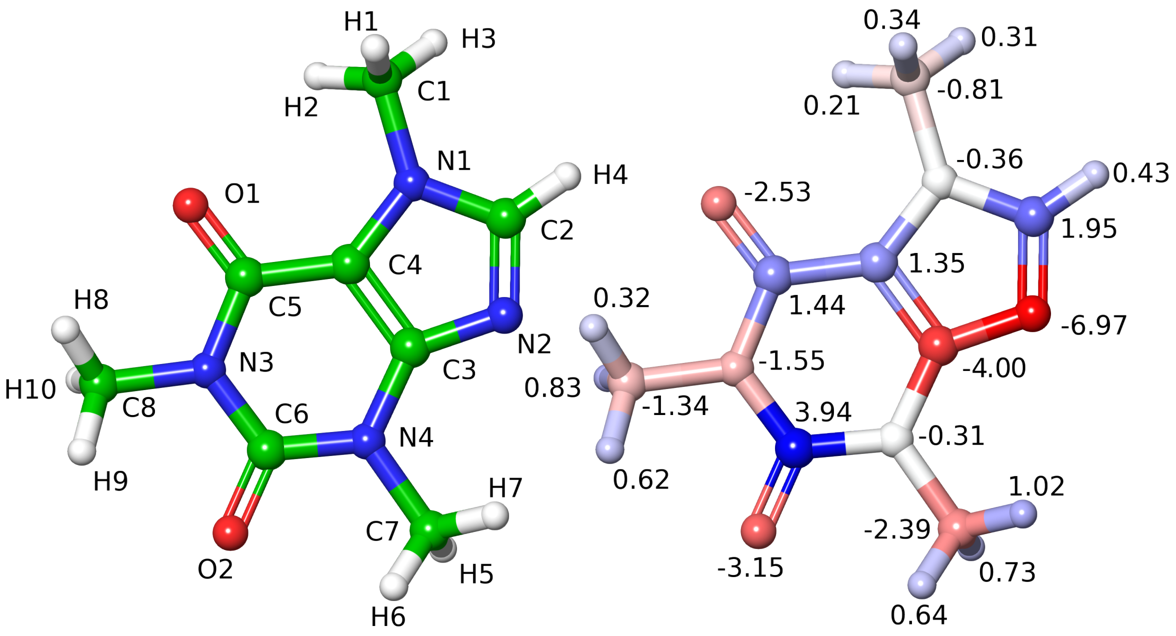


Fig. 4.4 The hydration free energy contributions from each atom in the caffeine molecule in kcal/mol. In the right hand side image, blue, white and red atoms have unfavorable, neutral and favorable contributions respectively.

expected to be the same. A deeper analysis of the atomwise free energy values shown in figure 4.4 is given in the analysis section.

Figure 4.5 shows the convergence of some of the atom-wise weights associated with atoms in the caffeine molecule as the number of considered trajectory frames is increased. The convergence is steady and fairly rapid, only requiring around 1000 frames to get a reasonable estimate for the contribution from each atom. The values in figure 4.4 are calculated from 1800 MD trajectory frames and are well converged based on the data presented in figure 4.5.

4.4.3 HIV-1pr with Lopinavir

Table 4.5 shows the five terms that together make up the total system free energy, the total free energy and an experimental comparison. Three terms are from the different stages of simulation, and the remaining two are correction terms arising from the restraints used to control the simulations. Figure 4.6 (located at the end of this chapter) shows the results of an AFEP calculation on the MD simulation trajectories. Each contribution is given by $w_{\text{bound}}\Delta F_{\text{bound}} + w_{\text{swap}}\Delta F_{\text{swap}} + w_{\text{unbound}}\Delta F_{\text{unbound}}$. In each case the weights are

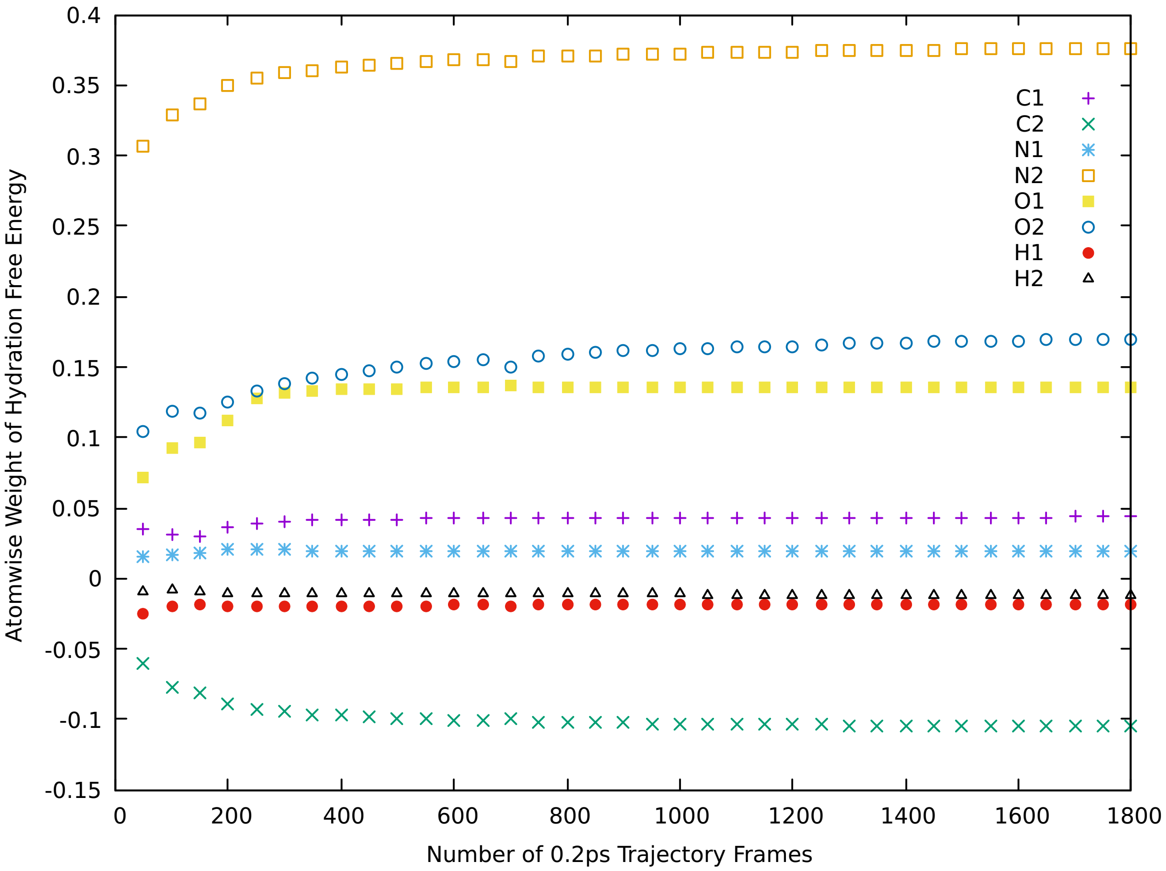


Fig. 4.5 The convergence of the atom-wise weights with considered trajectory frames for some of the atoms in a caffeine molecule solvated in water. To increase the readability of the plot not all atoms are shown. Points are taken at every 50 frames from 50 to 1800.

given by the difference between the atom-wise weights from the copy of the inhibitor in the binding site and the inhibitor in the solvent. The 12 most contributing and 12 least contributing sites across the lopinavir molecule are labeled. Red atoms contribute strongly to the binding. Blue sites are those that disfavor the binding. White sites are neutral. Some of the red atoms are points of symmetry, for example C35, H6, C15, C6 and H2 could all be reflected on the molecular structure by rotations of the side groups. This again implies perfect sampling of all conformational degrees of freedom has not occurred, however, in this case the binding site will restrict the ligand and prohibit these rotations. AFEP proves useful in determining that this asymmetry is present in the results.

Stage	Method	ΔF [kcal/mol]	Statistical Error [kcal/mol]
Turning on bound molecule non-bonded	BAR FEP	-21.44	0.57*
Turning off unbound molecule non-bonded	BAR FEP	-5.92	0.28*
Switching Coulombic interactions	BAR FEP	-1.77	0.12*
Analytic correction term	Calculated	5.92	-
TI Restraints (Energetic)	TI	-0.61	0.02 †
Total binding free energy	All	-23.82	0.65
Experimental binding free energy‡ [104][105]		-15.7	-

Table 4.5 The five contributions towards the total binding free energy and their total compared with an experimental value. The errors are statistical errors and do not take into account systematic errors from simulation parameters. ∗ Statistical errors from the BAR FEP routine. † Half the spread of the forward and backward TI calculations. ‡ Sourced from BindingDB [105].

4.5 Analysis

4.5.1 Methane-like Molecules

The free energy calculations for four methane-like molecules presented in tables 4.2 and 4.3 show a reasonable agreement with experimental free energy values. For methane the AFEP decomposition predicts that each of the hydrogen atoms is weakly favourable and of similar magnitude which could be interpreted as arising from interactions with the surrounding water. In methanol the oxygen atom is very favourable and the central carbon has become slightly polarised by the addition of the oxygen. The least favourable atom is the hydrogen in the hydroxyl group, H4, which may be due to hydrogen atoms from surrounding water molecules which are bonding with O crowding closely to H4 and offsetting the contributions made from hydrogen bonds from water to H4. If H4 and O are combined then the contribution from the OH group is still very favourable. A similar solvent crowding effect may explain the positive contributions from H4 and H5 in methylamine. Again the central carbon is slightly polarised and the nitrogen atom is very favourable, presumably from hydrogen bonding with the surrounding water. For methanethiol, the net free energy change is small, but each of the contributions is of similar magnitude to methanol and methylamine. The sulphur atom is relatively favourable. The free energies between different molecules cannot be easily interpreted, which is expected according to Mark et al. [87] due to the non-additivity of free energy partitions. For example, if methane and methanol are compared, the free energy of mutating H to OH cannot be found by using the total free energy and by splitting methane into CH_3 + H. This is because the CH_3 partitions of methane and methanol are not equivalent, partly due to the polarisation of the central carbon in methanol. In all cases, the results potentially make more sense when chemical groups are summed over.

4.5.2 Caffeine Total Free Energy

Table 4.4 shows the total hydration free energy for experimental and simulated caffeine. There is a mismatch between the total free energy for computational simulations and experimental measurements. This is probably because the force field doesn't contain full information about the molecular interaction and is a common problem with classical molecular dynamics simulations and is not specific to the AFEP methodology. The agreement between the two computational methods is within 1 kcal/mol. This is a reasonable tolerance for demonstrating the simulation was run correctly and the results can be used to explore the AFEP method. The simulation by Mobley et al. [83] was performed in the GROMACS molecular dynamics software [6] and has slight differences in the input and parameter files. It is possible that convergence of the total free energy change could be improved by adding more λ-windows to the BAR calculation. However, showing an accurate result for the total free energy change is not the goal of this work. Moreover, one of the goals is to show that AFEP produces a valid decomposition independent of the accuracy of the input free energy change.

4.5.3 Caffeine Atom-wise Free Energy

For caffeine the atom-wise weights given by Equation 4.10 were well-converged. This can be seen in figure 4.5, where after around 1000 considered trajectory frames the individual contributions to the free energy do not change significantly. The atom-wise values shown in figure 4.6 correspond to 1800 input trajectory frames. Although the results appear well-converged, we can tell that perfect sampling has not occurred. Chemically speaking, H1, H2 and H3 should have exactly the same contribution by rotational symmetry about the C1-N1 axis. However, the free energy contribution from H2 is somewhat different to H1 and H3. This is because fixed atoms were used in the simulation; if fixed atoms were not used then such a difference would have indicated a sampling problem. This kind of sampling problem is not particular to the AFEP methodology and is common in MD simulations. However, for AFEP there is the advantage that prior symmetry information of this kind could be input manually using the atom partitioning scheme. The researcher may still have to be careful before inputting such degeneracies. For example, if the energy of a conformational barrier is particularly high, much greater than the average thermal energy in the system, $k_B T$, then one may lose information by partitioning the atoms together. In this case AFEP could even be a way of diagnosing such conformational barriers, by checking sets of atoms that should have the same contribution by symmetry for distinctly different contributions.

If the AFEP contributions are summed over methyl groups, two out of three cancel to almost zero net contribution. The signs of contributions are consistent across the molecule, for example all three methyl groups show positive hydrogens attached to a negative carbon attached to a negative nitrogen. N2 is the best contributor to the solubility of the molecule. This makes sense as it is the most solvent exposed part of the molecule and the N2 molecule is assigned a large partial charge in the force field parameters. According to the AFEP simulation C6 is actively resisting the solvation of the molecule. Both of the oxygen atoms have quite large contributions to the hydration free energy. This makes sense as they will create hydrogen bonds with the surrounding water. H4 and C2 are both unfavorably contributing to the solvation. This may suggest that replacing H4 with an OH or NH_2 group would increase the solvation free energy of the molecule.

4.5.4 Lopinvair Total Binding Free Energy

The total free energy change of the HIV-1pr and lopinavir complex was calculated from five distinct contributions: There are the contributions from the three lambda stages, turning the VDW/repulsion interactions on in the binding site, swapping the charge interactions from the solvent molecule to the bound molecule and and turning the VDW interactions off in the solvent, as well as the analytic correction for the entropic cost of restraining a ligand in free space to the binding site (Equation 4.93), and the free energy associated with the restraints measured using thermodynamic integration. The sum of these five contributions gave a final value of -22.59 kcal/mol. The experimental result for this binding is -15.2 kcal/mol [106] as found using the BindingDB [105].

4.5.5 Lopinvair Atom-Wise Free Energy

The AFEP contributions were calculated for atoms in the lopinavir molecule. The unfavorably contributing sites are coloured blue and the favourable red in figure 4.6. Both the oxygen atoms and the nearest carbons are positively contributing for O1 and O4. This is in contrast to the caffeine molecule, where oxygen atoms are paired with unfavourable carbons. There are a few examples of bonded pairs with one favourable and one unfavourable atom. Examples are HN4 and N4, HN2 and N2, HN3 and N3 and HO4 and O4. The latter, pairs the strongest positive and negative contribution. All pairs result in significant cancellation if summed over. There are some obvious side groups on the molecule that have few strongly binding atoms and are quite neutral to the binding process. One example is the ring containing H6, which could potentially be replaced with a similar-sized side group that might display a stronger binding.

The protein atoms in the binding site can also be considered and also have atom-wise contributions. Visualisation of the interactions between the drug and the binding site are hard due to the complicated 3-dimensional nature of the problem. In a drug discovery context only the drug molecule can be altered and not the protein so these binding site atoms are not shown in this work. However, in investigations looking into the mechanism of specific protein interactions, this additional information is expected to be useful.

4.6 Conclusions

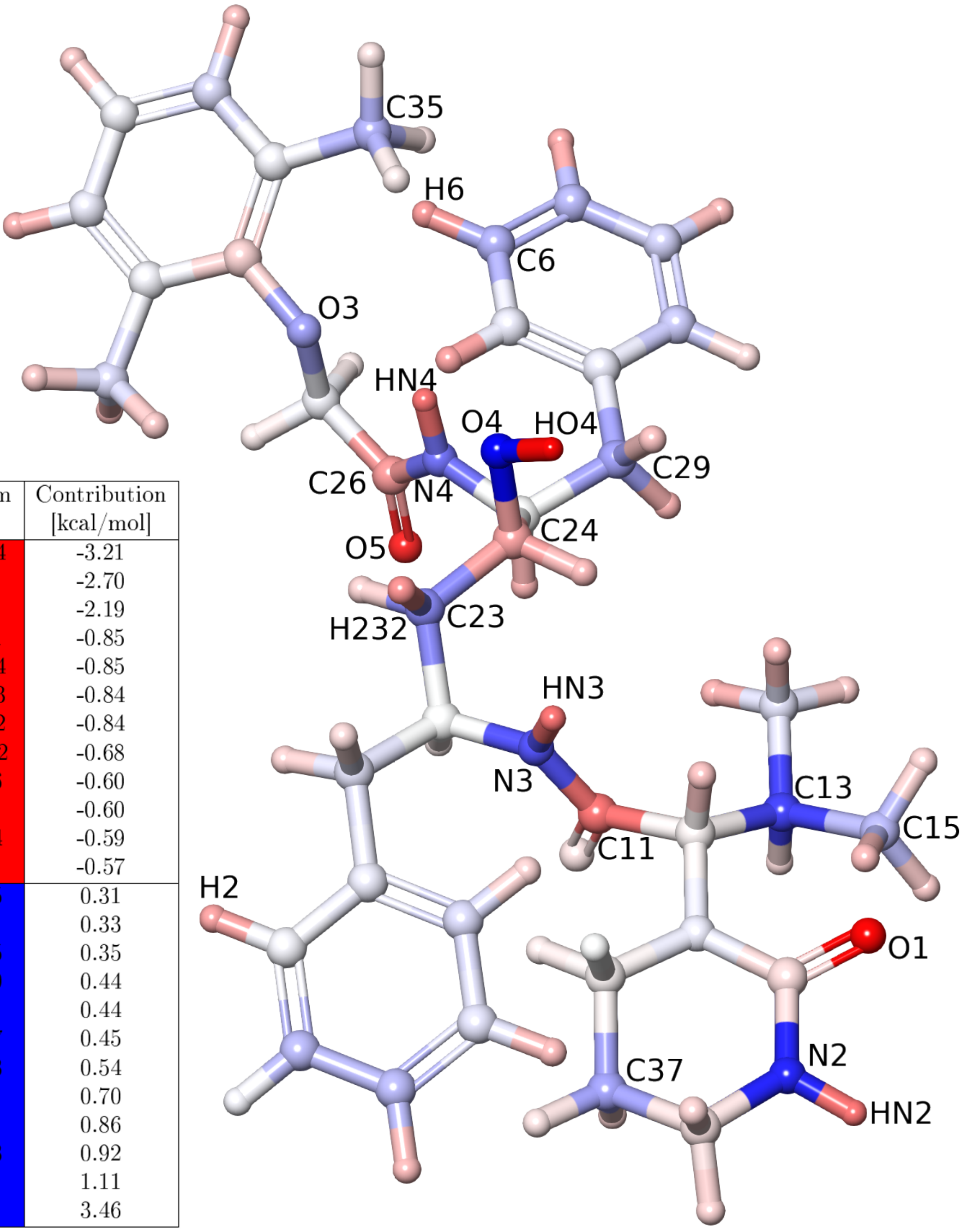
The Atom-wise Free Energy Perturbation (AFEP) method described in this article provides a detailed breakdown of a free energy change across partitions of atoms in a molecular dynamics simulation. The partitions can be selected by the user to capture information at an appropriate length scale. The results shown in this work indicate that a full atom-wise decomposition may be too detailed for certain features, and groups such as amines and hydroxyls should probably be summed over. The main equations used were directly derived from the Zwanzig equation as used in the exponential averaging method (EXP) for free energy changes at fixed volume. During this derivation it was defined that the system energy is decomposable as a sum over the atoms in the system. Two further assumptions were made in the derivation: The first is that when decomposing the system energy into individual contributions associated to atoms, the energy is shared equally between groups of atoms interacting with a given force field term; the second is that a simple repartitioning scheme can be used during the multinomial expansion stages of the derivation to give a meaningful estimate of the free energy contribution from each atom. Because of the nature of the last approximation, the AFEP method cannot produce an additive decomposition of the free energy, and comparisons between the partitions of *different molecules* should be made very carefully. However, AFEP can be used as an empirical tool for studying the internal free energy differences of a given molecule.

The AFEP weights used in the decomposition appear to converge relatively quickly. The system may be partitioned as the user chooses under the method because these weights can be linearly combined. This means interchangeable components of the system, namely the fluid molecules that are in an indistinguishable environment, can be combined together. Atoms that are symmetric by rotation of chemical groups can also be partitioned together in this manner, as could entire amino acids if a large protein was being studied.

The potential applications of this information are numerous. Only six specific examples were covered in this work, but AFEP could be applied to simulations relating to various branches of physics, materials science, chemistry and biochemistry. AFEP was applied to a calculation of the hydration free energy of methane, methanol, methylamine, methanethiol and caffeine. AFEP highlighted which atoms appear to interact the most with the surrounding water molecules. AFEP makes sensible decompositions for the four methane-like molecules and indicates a particular hydrogen atom in caffeine (H4) is relatively weakly interacting with the surrounding water. This type of analysis could be a useful tool to predict, for example, which molecule similar to caffeine dissolves into water more readily.

AFEP was also used to produce a free energy breakdown of the binding free energy change for a protein-ligand complex, HIV-1pr with lopinavir. The binding free energy was carried out using three simulation stages, and AFEP was applied to each stage and the results were summed along with two correction factors for the restraints used in the MD simulations. AFEP highlighted some sections of lopinavir that are neutral to the binding and may find use as a tool to suggest improvements to a given ligand to increase its binding strength. AFEP also showed which components of the lopinavir molecule are the strongest binding and are likely to be essential in the binding process.

This technology will have applications in the field of computational drug discovery and may assist in developing ligands for other disease target proteins in less time with reduced cost. There is a great scope for further biological and chemical uses away from the field of drug discovery, and the methods used are general and unassuming about the underlying physical system. The AFEP method can used to analyse the free energy contributions of atoms in most physical and chemical systems from appropriate molecular dynamics simulations arranged in lambda windows.



Atom	Contribution [kcal/mol]
HO4	-3.21
O1	-2.70
O5	-2.19
C11	-0.85
HN4	-0.85
HN3	-0.84
HN2	-0.84
H232	-0.68
C26	-0.60
H6	-0.60
C24	-0.59
H2	-0.57
C15	0.31
C6	0.33
C35	0.35
C29	0.44
O3	0.44
C37	0.45
C23	0.54
N4	0.70
N3	0.86
C13	0.92
N2	1.11
O4	3.46

Fig. 4.6 A representation of the binding free energy contributions from the lopinavir molecule in the binding site of the protein (not shown). The binding pose has been altered to display the atoms clearly. Blue atoms (C15-O4) are the least favorably contributing. Red atoms (HO4-H2) are strongly favorable with white weakly favorable or neutral. The atom-wise free energy contributions are shown for 12 unfavorable contributors and 12 favorable contributors as labeled in the figure.

Chapter 5

Hydration Free Energy Decomposition:

Prediction of GABARAP Interaction with the GABA type A Receptor

This chapter is mostly based on the publication "Prediction of GABARAP Interaction with the GABA type A Receptor" [107] which was accepted for publication in Aug 2018 by Proteins: Structure, Function, and Bioinformatics.

5.1 Introduction

This chapter uses a different kind of free energy decomposition to those considered previously and applies it to two proteins which bind together. The method uses IFST as described in Chapter 3, and calculates the free energy associated with removing water molecules from locations called hydration sites (HS) on the surface of the protein. HS are points which water molecules frequent on the surface of a protein, and are identified by 'clustering' the molecules across frames from an MD simulation. By calculating the free energy of hydration of each HS using IFST, information is gleaned about the nature of water in the local protein environment. If a water molecule is easy to displace (i.e. highly displaceable) then the HS associated with that water molecule is relatively hydrophobic. If the HS is not displaceable, then it is hydrophilic.

By analysing the distribution of hydration free energy across the protein surface, one can see which locations have the most displaceable water. If a suitable algorithm is run across the HS, the most displaceable connected set of HS can be found. If this set was large enough, this would be a volume of water in which water molecules could easily be pushed aside, which is a precursor for ligandability. One can hypothesise that for suitably hydrophobic interaction mechanisms, this ‘best site‘ is also the most ligandable site, as it is the easiest for a ligand to access. This line of thought was carried out by Vukovič and Huggins and applied to the bromodomains family of proteins [108]. In that work they developed a combinatoric search algorithm which finds the best connected clusters of hydration sites on the protein surface. This shows great promise for detecting hydrophobic ligand-protein interaction sites.

5.1.1 Motivation

This chapter covers a new application of this technology to the problem of protein-protein binding. Hydrophobicity is also an important factor in protein-protein binding. An interesting system to apply the above methodology to is the binding of the GABA$_A$-receptor and the GABA$_A$-receptor associated protein (GABARAP). It is not known exactly how the two proteins bind and the decomposition of hydration free energy across the binding surface may shed some light on this. GABARAP binding to the GABA$_A$-receptor was chosen as a system to study for a number of reasons:

1. There is experimental evidence that the two proteins bind. The evidence indicates GABARAP binds to a specific helix in the intracellular domain of the GABA$_A$-receptor, namely the $\gamma 2$ subunit helix. In addition, GABARAP and GABARAP-like proteins have been studied binding to similar segments of α-helix in different proteins.

2. Information exists about the structures of each component. The structure of GABARAP is known and the amino acid sequence of the $\gamma 2$ helix is known which allows a molecular model of the bound conformation to be built. The experimental structure of the intracellular domain of the GABA$_A$-receptor is not known.

3. The GABA$_A$-receptor is responsible for the majority of fast neuronal inhibition in the central nervous system (CNS) and the target for an important group of compounds: the benzodiazepines, which are anxiolytics, hypnotics and anticonvulsants. The structure and function of this important protein are still not fully understood and any project relating to it will help complete that wider understanding.

4. Interesting research by Fan et al. has shown that a drug called metformin, which is commonly used to treat type 2 diabetes, has an impact on the expression of GABA$_A$ receptors in rats [109]. The authors make a strong case that this change in protein

expression is due to the expression of GABARAP which has a hypothesised role in transporting receptors to the cell membrane. Metformin shows no evidence of tolerance or dependence to the therapeutic effects on the CNS.

5. A structural understanding how GABARAP acts in the chain responsible for shipping GABA receptors to their anatomical focus will help build a structural understanding of the proponents of tolerance and dependence in this important class of medicines.

6. Studies have shown that GABARAP binding modulates ion passage conductivity. Having a proposed dock of the two proteins will help explore further theoretical avenues to measure ion passage conductivity using computer simulation.

An overview of the key properties of the $\mathrm{GABA_A}$-receptor and GABARAP is given below. In the following sections, peptides will be denoted by strings of single letters each indicating amino acids. For example RTGAWRHGRIHIRIAKMD is a peptide starting with arginine (R) and ending with aspartic acid (D).

5.2 System Background

5.2.1 The $\mathrm{GABA_A}$-receptor

$\mathrm{GABA_A}$-receptors are important proteins in neurology as they control inhibition of neurons in mammals. They interact with anaesthetics, where different molecules bind at various sites on the receptor. They have roughly cylindrical symmetry about a central ion channel in which 5 long subunits lie in a pentagonal shape when viewed from above (see Figure 5.1). There are three significant sections, the extracellular domain, the transmembrane domain and the intracellular domain. The intracellular domain is hardest to resolve experimentally and is truncated in most structures. In the model of the $\mathrm{GABA_A}$-R used in this work, the intracellular domain consists of 5 helices, one per subunit. The extracellular domain is the entrance point for ions to access the ion channel, it is also the place where two GABA molecules can bind and other molecules such as benzodiazepines bind. The transmembrane domain is the site of most of the mechanical function of the protein, the narrowest point in the ion channel is here along with many layers of α helices which together control the gating of the protein.

There are many types of naturally occurring subunit and combinations of these create different variants of the protein. The subunits are denoted by Greek letters $(\alpha,\beta,\gamma,\rho,\delta,\epsilon,\theta,\pi)$ of which there are $(6,3,3,3,1,1,1,1)$ types respectively, not all combinations form naturally, for example ρ subunits only join to other ρ subunits. In this work, the most commonly occurring type in the body was used, which is the $(\alpha 1)_2(\beta_2)_2\gamma 2$ type. The subunits in this variant are arranged in the pattern $\alpha\beta\alpha\beta\gamma$ such that they alternate in type with no two similar subunits next to each other.

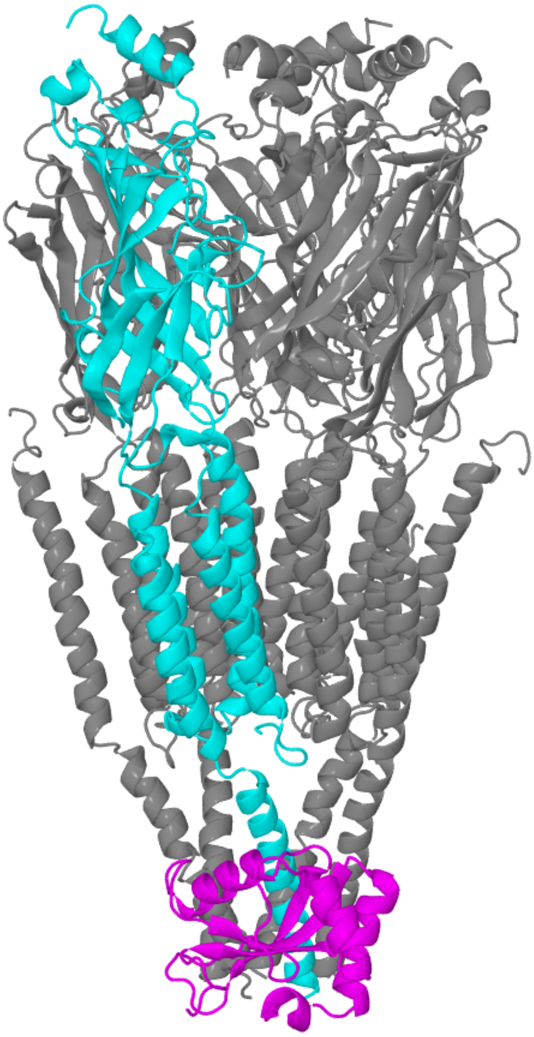


Fig. 5.1 Diagram showing a model of the GABA$_A$ receptor and a proposed docking pose of the GABARAP, 1KOT model 1 dock 17d. The GABA$_A$ receptor is modelled using 2BG9 as the template, and so only part of the intracellular domain is modelled. The γ2-subunit is shown in cyan, and the rest of the receptor shown in grey; GABARAP is shown in magenta.

5.2.2 GABARAP

GABARAP, was first described by Wang et al. [110]. It is a protein of 117 amino acids and has a relative molecular mass of 13900 (Daltons). These authors also determined that GABARAP interacted with amino acids 394–411 of the intracellular domain of the γ2-subunit of the GABA$_\mathrm{A}$ receptor. If this sequence was shortened from either end to either 399–411 or 389–402, then the interaction was no longer observed. These authors also reported that GABARAP 36–117 and GABARAP 1–68 both interacted with the γ2-subunit of the GABA$_\mathrm{A}$ receptor during a Glutathione-S-transferase (GST) pull-down assay, indicating that the interaction domain spanned GABARAP amino acids 36–68. In a subsequent paper, Nymann-Andersen et al. [111] concluded that the octadecapeptide (18-mer) RTGAWRHGRIHIRIAKMD from the GABA$_\mathrm{A}$ receptor γ2-subunit was necessary and sufficient for interacting with the GABARAP, but the interaction, as determined by the GST pull-down assay, was not as strong as that given by the tricosapeptide (23-mer) CFEDCRT-GAWRHGRIHIRIAKMD. This molecule gave the highest level of activity in the assay.

Knight et al. [112] examined the nuclear magnetic resonance (NMR) shift of the GABARAP cross-peaks when the octadecapeptide RT..MD was present. They noticed that the NMR signals from GABARAP amino acids Val 31, Arg 40, Asp 45, Lys 46, Leu 50, Val 51, Leu 55, Thr 56, Phe 60, Ile 64, Arg 65 and Glu 101 were significantly changed, with Lys 46, Val 51, Phe 60 and Ile 64 displaying changes of the order of 1 linewidth. These authors also estimated the dissociation constant, K_d, of the octadecapeptide RT..MD from GABARAP to be higher than 0.2 mM, so the measured binding was weak.

Coyle et al. [113] measured intrinsic tryptophan fluorescence (ITF) to study the binding between GABARAP and the γ2-subunit of the GABA$_\mathrm{A}$ receptor. They used native GABARAP, GABARAP with the first 10 amino acids truncated (denoted ΔN10) and GABARAP with the first 27 amino acids truncated (denoted ΔN27). They found that the dissociation constant between the octadecapeptide RT..MD and native GABARAP was $1.29 \pm 0.09\,\mu$M, between the octadecapeptide and ΔN10 was $1.17 \pm 0.06\,\mu$M, and between the octadecapeptide and ΔN27 was $6.10 \pm 0.29\,\mu$M. The dissociation constant between native GABARAP and the tridecapeptide (13-mer) RTGAWRHGRIHIR was $3.33 \pm 0.34\,\mu$M, and between native GABARAP and the undecapeptide (11-mer) GAWRHGRIHIR was $5.52 \pm 0.52\,\mu$M. These dissociation constants are much smaller than that determined from NMR by Knight et al. [112], and it is still unclear where the source of the large discrepancy lies [114].

The function of GABARAP is most probably twofold: anchoring the GABA$_A$ receptor to the cytoskeleton, and modulating the function of the receptor. Amino acids near the N-terminus of GABARAP could bind to tubulin [113], whilst the amino acids nearer the C-terminus bind to the GABA$_A$ receptor [111]. Chen et al. [115] showed that GABARAP caused GABA$_A$ receptor clustering, and clustered receptors exhibited lower affinity for GABA (EC$_{50}$ increased from $5.74 \pm 1.4\,\mu$M to $20.27 \pm 3.8\,\mu$M), and they desensitised less quickly (the desensitisation time constant τ increased from 1 s to 2 s). Everitt et al. [116] performed electrophysiology experiments and showed that GABARAP promotes the clustering of GABA$_A$ receptors, and increases the conductance of the GABA$_A$ receptor from below 40 pS to above 50 pS.

Despite all these studies on the interaction between the GABA$_A$ receptor and GABARAP, we still do not know the structural details of this interaction. Weiergräber et al. [114] co-crystallised GABARAP with the K1-peptide (sequence DATYTWEHLAWP) and determined the structure to 1.3-Å resolution. They used this data and previous published data to infer the interaction between GABARAP and the GABA$_A$ receptor.

In this chapter the experimental structures of GABARAP and a modelled structure of the intracellular domain of the GABA$_A$ receptor were used to performed docking simulations. Independently, IFST [20, 21] was used to calculate the free energy of displacing all reasonable clusters of water containing 7–18 molecules from the surface of the intracellular domain of the GABA$_A$ receptor, and from the surface of experimental structures of GABARAP. This information was applied to validate the docking interaction between the GABA$_A$ receptor and GABARAP, in the context of surface hydration following the methods of Vukovič et al. [108].

5.3 Clustering and IFST Calculations on Hydration Sites

5.3.1 Clustering

To cluster water into hydration sites an algorithm is used that takes an MD trajectory as input and outputs a structure of the protein surrounded by hydration sites. These sites also have an occupancy and density associated with them, a quantity ranging from 0 to 1 which describes how often the site was occupied across the trajectory analysed. The input trajectory will contain a few thousand frames of uncorrelated MD snapshots. If the

MD frames are sampled with short time intervals between them, the data will be skewed and will not represent the average behaviour of water correctly. The algorithm uses a grid to find the location of the densest patch of water, i.e. the patch with the most water molecules within a cut-off radius on the grid. Then a hydration site will be placed at that location, this process is repeated but the location of the water at the hydration site is taken into account, such that a new site will not be assigned in previously defined locations. The hydration sites will not be placed too closely to each other, i.e. a distance around 2.4 Å that the user can select and control.

5.3.2 IFST Calculations

The clustering step outputs a protein surrounded by hydration sites. This information is then used for the IFST calculation step. Each hydration site has an IFST calculation performed on it for which the previous MD trajectory is also used as input. If the protein moves too much, then the location of the hydration site will not make sense across the trajectory. To control this restraints are used during the MD simulation of the protein, this would usually be in the form of harmonic restraints on backbone atoms, or in very severe cases fixing atoms around exceptionally dynamic side groups. The IFST calculation is equivalent to solvating a water molecule at the location of the hydration site. As in chapter 3, the hydration free energy can be calculated by dividing it into an energetic and entropic part. The entropic part is calculated using a nearest neighbours estimator [49]. The calculation also takes an approximation of the next order entropy integrals into account for both translational and orientational entropies. The physical quantities computed for the hydration sites are

1. ΔG of hydration for the entire site
2. occupancy
3. density
4. ΔS total
5. ΔH total
6. ΔS solute-water translational
7. ΔS solute-water orientational

8. ΔS water-water translational

9. ΔS water-water orientational

along with others which are not considered such as the protein enthalpy and the binding enthalpy of the site. Hence, not only is the free energy then decomposed into the entropy and enthalpy, but the entropy is further decomposed into first and second order terms for translational and orientational types.

5.3.3 Finding the Best Cluster

Once the hydration free energies of the hydration sites have been calculated, the search algorithm can begin. The search algorithm has multiple modes of operation and was first used on HS by Vukovič and Huggins [108]. A common mode of application is to find the connected cluster with the 'best' net free energy. 'Best' means that it is the most easily displaced connected cluster of k hydration sites across the whole protein. k can be changed and the definition of whether two HS are connected depends on a distance cut-off which is usually set to around 4.8 Å, to represent the hydrogen bond length between water molecules. As k gets larger the algorithm takes longer to complete due to the complexity of selecting larger clusters. A naive algorithm that tries all combinations of clusters with k HS from n total HS around the protein will have to try $\binom{n}{k}$ selections. For a protein with 1000 HS and a cluster of 18 HS, this number has 38 digits. In addition to this, each combination will require a routine that checks if the HS in the combination are connected. A more efficient strategy is used in which all connected combinations of HS are generated first, this vastly reduces the number of combinations that need to be checked and removes the process of checking if the combination is connected. Once this preprocessing stage is made, the algorithm to find the best site is reduced to a simple loop over sites. This compiles very efficiently allowing even billions of combinations to be checked in a reasonable time-scale. For most proteins, and clusters of up to size 18, the algorithm only needs to run for a few days to locate all of the connected clusters.

Once this process is finished either the best cluster or best few clusters will be output and analysis can begin on those cluster sites.

Limitations of the Algorithm

It should be noted that taking the sum of the free energy values for the cluster as a ’cluster free energy’ is an approximation. This again is due to the inseparability of free energies where correlation exists. For the clusters of hydration sites, their free energies are almost certainly correlated because the existence of a hydration site implies a water molecule which will likely hydrogen bond with neighbouring water molecules. This being said, the free energy sum of a cluster is an indicator of displacability. The algorithm is already combinatorially demanding, and it is unlikely that a simple better approach exists with a competitive run time.

5.4 Methods for Simulation

5.4.1 Molecular coordinates

In this chapter, we used the coordinates of a GABA$_A$ receptor model from the work of Mokrab et al. [117]. This model used the nicotinic acetylcholine receptor (nAChR) structure from the work of Unwin [118] as a template. Unwin resolved the five intracellular helices on the model (PDB: 2BG9). The model used is the only model of the GABA$_A$ receptor that includes part of the intracellular domain which will be required for the binding of GABARAP. The subunit composition of the receptor in this work is $(\alpha 1)_2(\beta 2)_2\gamma 2$ which is the most common composition in the body.

There exist five stand-alone structures of GABARAP, and their PDB codes are 1GNU, 1KJT, 1KOT, 1KLV and 1KM7. 1GNU and 1KJT come from X-ray crystallography experiments, and 1GNU was chosen because of its higher experimental resolution of 1.75 Å. 1KOT, 1KLV and 1KM7 all come from NMR experiments; 1KM7 contains only one conformer, whilst residues 1–17 in 1KLV could not be located and so 1KOT was chosen, which has fifteen conformers. We thus used two structures of GABARAP. One is an NMR solution structure, PDB code 1KOT [119], and the other is an X-ray crystallography structure, PDB code 1GNU [112].

5.4.2 Docking

The fifteen slightly different conformations in the NMR structure 1KOT are labelled 1KOT model 1 to 1KOT model 15. The X-ray structure 1GNU contains only one coordinate set, but Ser 16, Ser 53 and Arg 65 have been resolved with two alternative conformations, each with experimental occupancy 1/2. This makes eight structures from the 1GNU coordinate set, each with slightly different conformations called 1GNU-(aaa,aab,$\cdots$,bbb) depending on whether the A-form or the B-form from the PDB was chosen.

The 23 structures, fifteen from NMR experiments, and eight from X-ray crystallography experiments, were used as the ligand in a docking calculation using SwarmDock [120, 121]. This docking method allows for flexibility in the molecules by using normal mode analysis [122], and the program is available to be used on a public server[1]. For the receptor (i.e. the helix in the intra-cellular domain), we used the modelled coordinates of the tricosapeptide C^{420}FEDCRTGAWRHGRIHIRIAKMD^{442} from the γ2-subunit of the GABA$_A$ receptor; this is the section from Cys 420 to Asp 442. Experiments by Nymann-Andersen et al. [111] showed that this tricosapeptide gave full binding to GABARAP. An attempt to dock GABARAP to the complete GABA$_A$ receptor was made, but this was rejected by SwarmDock as the GABA$_A$ receptor contained too many atoms (14900 non-hydrogen atoms). Therefore we used only part of the γ2-subunit in the docking. In this work, we did not specify the interface amino acids on the proteins and only used the 'blind' docking mode. A maximum of five normal modes were allowed for each molecule.

SwarmDock produced 468 docks for each GABARAP conformation. The output consisted of 10764 coordinates of different conformations of GABARAP and the tricosapeptide from the GABA$_A$ receptor. The coordinates of the latter were slightly different from the original tricosapeptide coordinates, as the SwarmDock flexible docking had changed the structure of both the receptor and the ligand. A least-squares fit was used to superimpose the SwarmDock structure of the receptor onto the original tricosapeptide coordinates; the translation vector and rotation matrix used were noted. The same vector and matrix were subsequently used to move GABARAP to a model of the complete GABA$_A$ receptor whose γ2 tricosapeptide position was coincident with that of the tricosapeptide used in the docking. Steric clashes were then tested for between GABARAP and the GABA$_A$ receptor. If two atoms, one from each protein, were found to be within 1 Å of each other, that dock was rejected.

[1] As of the time of writing this article, the SwarmDock server resides on https://bmm.crick.ac.uk/ svc-bmm-swarmdock/index.html

The results filtered for steric clashes were then subjected to a further filtering process using the following criteria:

1. At the interface, the GABARAP amino acids Lys 46, Val 51 and Phe 60 were all present.

2. At the interface, at least one of the GABA_A receptor amino acids Arg 425, Thr 426, Gly 427, Ala 428 or Trp 429 was present.

3. At the interface, at least one of the GABA_A receptor amino acids Arg 433, Ile 434, His 435, Ile 436, Arg 437, Ile 438, Ala 439, Lys 440, Met 441 or Asp 442 was present.

Criterion 1 was applied to locate docking positions consistent with NMR experiments [112]. In this paper, Ile 64 was also identified as an important interface amino acid, but its position means that we were unable to obtain any docking poses with Ile 64 at the interface. Criteria 2 and 3 were applied to extract docks consistent with the yeast two-hybrid assay [110]. 161 docks were selected after these procedures.

We undertook further filters to select the optimal docks from these 161 docks: we examined the distribution of these 161 docks according to the following seven criteria:

4. The SwarmDock energy score should be in the more favourable half of the energy score distribution.

5. The number of ligand amino acids with at least one atomic contact to the receptor amino acids Arg 425 to Trp 429 and Arg 433 to Asp 442 should be in the higher half of the corresponding distribution.

6. The number of ligand amino acids with at least one atomic contact to the ‘cytoplasmic’ receptor amino acids Arg 425 to Trp 429 should be in the higher half of the corresponding distribution.

7. The number of ligand amino acids with at least one atomic contact to the ‘membrane’ receptor amino acids Arg 433 to Asp 442 should be in the higher half of the corresponding distribution.

8. The number of receptor amino acids with at least one atomic contact to any ligand amino acid should be in the higher half of the corresponding distribution.

9. The number of atomic contacts from the ligand to any of the receptor amino acids Arg 425 to Trp 429 and Arg 433 to Asp 442 should be in the higher half of the corresponding distribution.

10. The number of atomic contacts from the receptor to any ligand amino acids should be in the higher half of the corresponding distribution.

In the above criteria, a contact was defined as an atom which was less than the sum of the van der Waals radii of the two atoms + 20% [120, 121]. A dock was selected from these 161 configurations if all of these additional seven criteria were met.

These seven additional criteria were chosen to ensure that the best ligand structure should have a competitive energy score such that the structure is stable (criterion 4), maintain an overall high contact to the receptor (criterion 5) to multiple sites which are distributed between the upper (criterion 6) and lower (criterion 7) portions of the receptor sequence. The best structures must also reciprocate contact across many sites on the ligand (criterion 8) and the strength of all contacts should be a close and strong as possible on the receptor (criterion 9) and ligand (criterion 10).

5.4.3 Free energy change calculations

The molecules were prepared using the CHARMM-GUI freely available on the web [123][2]. The molecular dynamics package NAMD2 [7] was used in this work. A full description of the MD protocol used to simulate the proteins is given in Appendix B.

The MD trajectory for the GABA$_A$ receptor was processed as described by Vukovič et al. [108]. First, hydration sites as defined by Haider and Huggins [124] were created on all surface regions of the GABA$_A$ receptor. The hydration sites represented time averaged water molecules and were assigned positions, densities and occupancies [125, 48]. Hydration sites with a radius of 1.2 Å were picked starting from the densest patch of water in order of decreasing density and no sites were picked within 2.4 Å of an already existing site. Next, an IFST calculation for the free energy was carried out for each of the hydration sites according to IFST described in Vukovič et al. [108]. IFST had previously been used on water molecules around proteins where the proteins are involved in binding small ligands [126–128] and in protein-protein interactions [129]. All 10000 snapshots of the protein

[2] As of the time of writing this article, the address of the CHARMM-GUI is http://charmm-gui.org

sampled at 0.5 ps intervals were used to calculate the free energy difference associated with hydrating each site with a single water molecule. These free energy differences were mostly negative because solvation was favourable.

At this stage some hydration sites were removed to improve the efficiency of the combinatorial algorithm. Hydration sites inside the ion channel of the GABA_A receptor were removed; the ion channel was aligned to the z-axis, the positions of all protein atoms were converted to cylindrical coordinates with a height z, and a radius and angle in the xy-plane. The cylindrical mid-plane of the protein atoms as a function of height and averaged over angle was found by fitting a quadratic polynomial to the protein atom data. Hydration sites on the inside of this mid-plane were removed. Hydration sites with coordinate $z > -48$ Å were also removed as this region was close to the lipid bilayer in the full GABA_A receptor model.

Then a combinatoric search scheme was employed to search for up to the best 1000 clusters containing from 7 to 18 hydration sites within an energy of 12.5 kJ/mol of the best cluster. The search was run three times with these parameters, the first time searching for 'near' clusters with hydration sites at most 3.1 Å away from non-hydrogen atoms and 3.6 Å away from hydrophobic non-hydrogen atoms, the second time searching for 'regular' clusters with hydration sites at most 3.6 Å away from non-hydrogen atoms and 4.1 Å away from hydrophobic non-hydrogen atoms, as originally performed by Vukovič et al. [108]. The third search was for 'far' clusters with hydration sites at most 4.1 Å away for non-hydrogen atoms and 4.5 Å away for hydrophobic non-hydrogen atoms. These three ranges were selected to investigate how the hydration patches changed upon variation of the hydration site cut-off distance from the protein *i.e.*, the degree to which bulk-like distal waters are included in hydration patches.

The method used by Vukovič et al. [108] finds sites with high ligandability for drug molecules binding to a protein. Advances in the combinatoric search method allow clusters of the size of a drug molecule to be found. These authors conclude that, for a small peptide, clusters of up to 30 hydration sites may need to be considered. Finding clusters with volumes commensurate with the ligand in this case is computationally infeasible, especially as GABARAP is much larger than a small peptide. As the free energy change of displacing hydration sites relative to bulk water atoms tends to zero at distances as small as 7 Å–8 Å from the surface [108], one could instead search for a *clustering of clusters* with the most favourable displacement free energy scores used to estimate candidate regions for larger objects to bind, namely proteins. This method was employed for the GABA_A

receptor. The set of hydration sites within the best 1000 clusters for each size of 7 to 18 hydration sites were filtered, and turned into hydration patch data for all three classes of clusters, ‘near’, ‘regular’ and ‘far’. For GABARAP, multiple ‘regular’ passes were made of the hydration patch combinatoric search, and after each iteration, the hydration sites associated with patches identified previously were removed. There were 5 passes for the 1KOT file and 4 passes on the 1GNU file, after which no more sites could be found. The first-pass sites take the least energy to displace and hence are the most displaceable and the fifth-pass ones are the least displaceable.

5.5 Results

5.5.1 Docking

SwarmDock produced 10764 docks, and 161 docks were selected according to the first three criteria described in the previous section. Using the seven additional criteria, we identified eleven docks, two of them from 1GNU and nine from 1KOT. The configurations of these docks are shown in figures 5.1 and 5.4 and they show a high degree of similarity between all eleven docks. The root-mean-square deviation of Cα-atoms between all eleven docks was calculated and the values are shown in table 5.1. The largest deviation in the structure comparisons was 2.57 Å, between 1GNU-bbb dock 41d and 1KOT model 1 dock 17d.

In table 5.2, we list the contacts between amino acids pairs, one from each protein. Some of these contacts have few contact atoms and are only observed in one docked pair. Other contacts have many contact atoms, and are found in all eleven docked pairs. In this table, we only list contact pairs where there are more than 10 contact atoms, and where they are observed in at least nine out of the eleven docked poses.

These contacts can be roughly grouped into five groups and their contact positions are shown in figure 5.2. We also display the two contact faces individually in figure 5.3. Experimental NMR research showed that GABARAP Lys 46, Val 51, Phe 60 and Ile 64 exhibited large shifts in their NMR spectrum on binding to the octadecapeptide R^{425}TGAWRHGRIHIRIAKMD^{442} [112]. Yeast assays [110] and fluorescence titration experiments [113] showed that, in the tricosapeptide C^{420}FEDCRTGAWRHGRIHIRIAKMD^{442}, the amino acids RTGAW and GRIHIRIAKMD at both ends were of particular importance. Our docking results show that

Dock num	2	3	4	5	6	7	8	9	10	11
1	0.67	2.56	2.36	2.36	2.42	2.35	2.35	2.35	2.35	2.35
2		2.57	2.40	2.40	2.47	2.39	2.40	2.39	2.38	2.56
3			1.69	1.62	1.65	1.60	1.58	1.68	1.60	1.60
4				1.55	1.54	1.53	1.51	1.56	1.53	1.54
5					0.34	0.17	0.19	0.18	0.11	0.08
6						0.41	0.41	0.41	0.38	0.38
7							0.09	0.17	0.14	0.14
8								0.22	0.15	0.16
9									0.20	0.19
10										0.06

Table 5.1 This table shows the root-mean-square deviation between these eleven structures in Å. Column 1 and row 11 have been omitted due to redundancy. The eleven chosen docks were: (1) 1GNU-aaa dock 28b, (2) 1GNU-bbb dock 41d, (3) 1KOT model 1 dock 17d, (4) 1KOT model 11 dock 29d, (5) 1KOT model 15 dock 39b, (6) 1KOT model 15 dock 40c, (7) 1KOT model 15 dock 40d, (8) 1KOT model 15 dock 41a, (9) 1KOT model 15 dock 42c, (10) 1KOT model 15 dock 54a, (11) 1KOT model 15 dock 54d.

GABARAP Lys 46 is in contact with Asp 423 of the γ2-subunit of the GABA$_A$ receptor in all eleven docks, but we are unable to observe large contacts between GABARAP Val 51, Phe 60 and Ile 64. However, there are large contact areas in the neighbouring amino acids: γ2-subunit Cys 424 and Ala 428 both make contact with GABARAP Leu 50 in all eleven docks, γ2-subunit Ile 438 makes contact with GABARAP Gln 59 in all eleven docks, and γ2-subunit Ile 434 makes contact with GABARAP Leu 63 in all eleven docks. In addition, γ2-subunit His 431 makes contact with GABARAP Leu 63 in ten out of eleven docks, and γ2-subunit His 435 makes contact with GABARAP Gln 59 in ten out of eleven docks.

$GABA_A$-R amino acid	GABARAP amino acid	freq. of occurrence
Asp 423	Lys 46	11/11
Cys 424	Leu 50	11/11
Ala 428	Leu 50	11/11
Ala 428	Arg 28	11/11
Arg 430	Arg 67	11/11
Ile 434	Leu 63	11/11
Ile 438	Gln 59	11/11
His 431	Leu 63	10/11
His 435	Gln 59	10/11
Cys 420	Lys 48	9/11
His 431	Tyr 49	9/11

Table 5.2 Table showing contact pairs between the receptor and ligand, and the frequency of finding that contact pair.

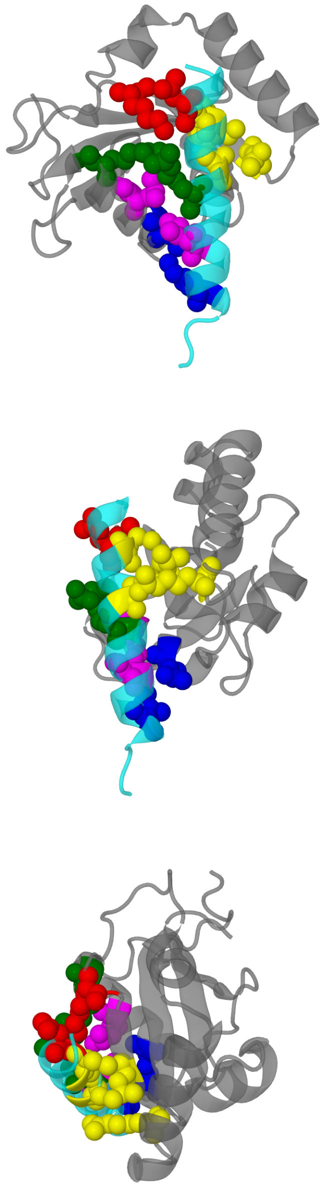


Fig. 5.2 Diagrams comparing the key contact amino acid pairs between the intracellular helix of the γ2-subunit of the GABA$_A$ receptor and GABARAP. The intracellular helix is shown in cyan, whilst GABARAP is shown in grey. The contact amino acid pairs are divided into five groups, each group colour coded in the following manner: (1) red – γ2-subunit Asp 423, GABARAP Lys 46 (2) yellow – γ2-subunit Cys 424 and Ala 428, and GABARAP Arg 28 and Leu 50 (3) green – γ2-subunit Cys 430 and GABARAP Arg 67 (4) magenta – γ2-subunit Ile 434 and GABARAP Leu 63 (5) blue – γ2-subunit Ile 438 and GABARAP Gln 59. The view in the top panel is from the ion channel towards the outside of the protein, that in the middle panel is from the side of the intracellular helix, and that in the lower panel is from the membrane towards the cytoplasm. These three viewing directions are roughly orthogonal to each other.

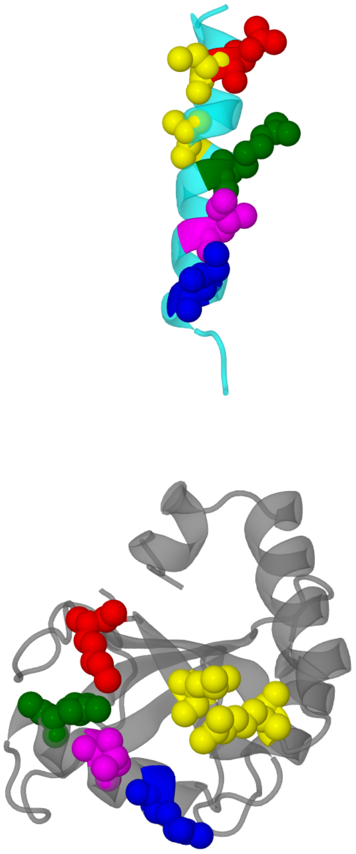


Fig. 5.3 Diagrams comparing the key contact amino acid pairs between the intracellular helix of the γ2-subunit of the GABA$_A$ receptor and GABARAP. The intracellular helix is shown in cyan, whilst GABARAP is shown in grey. The contact amino acid pairs are divided into five groups, each group colour coded in the following manner: (1) red – γ2-subunit Asp 423, GABARAP Lys 46 (2) yellow – γ2-subunit Cys 424 and Ala 428, and GABARAP Arg 28 and Leu 50 (3) green – γ2-subunit Cys 430 and GABARAP Arg 67 (4) magenta – γ2-subunit Ile 434 and GABARAP Leu 63 (5) blue – γ2-subunit Ile 438 and GABARAP Gln 59. The top panel shows the amino acids on the intracellular helix, and the bottom panel shows the amino acids on GABARAP.

5.5.2 Hydration of the GABA$_A$ receptor intracellular domain

The top panel of figure 5.5 shows the most displaceable 'close' hydration sites near the intracellular domain of the γ2-subunit of the GABA$_A$ receptor. It can be seen that there is a clustering of hydration sites on the γ2-subunit as well as hydration sites on the adjacent β2-subunit. The middle panel shows the most displaceable 'regular' hydration sites near the intracellular domain of the γ2-subunit of the GABA$_A$ receptor. There is a similar clustering of hydration sites on the γ2-subunit as well as hydration sites on the adjacent β2-subunit including an additional higher patch. The bottom panel shows the most displaceable 'far' hydration sites near the intracellular domain of the γ2-subunit of the GABA$_A$ receptor. The clustering of hydration sites on the subunits is similar to the 'regular' case.

Figure 5.6 compares the location of the hydration sites with the location of the predicted SwarmDock poses. The GABARAP positions are very close to the red and orange hydration sites. It can be seen that there is considerable agreement between the predicted docked poses of GABARAP, and the identified hydration sites. This could form the interface between the γ2-subunit of the GABA$_A$ receptor and GABARAP.

The three classes of hydration site clustering, 'close', 'regular' and 'far' all show a set of most displaceable clusters: those primarily situated on the γ2-subunit (red), those between the γ2 and β2-subunits (orange) and those on the lower, cytoplasmic portion of the β2-subunit (yellow). In addition to this, a patch was found on the β2-subunit (green) in the 'regular' and 'far' classes. As can be seen in table 5.3, the red patch on the γ2-subunit is the easiest to displace on average across all classes.

The amino acids within 5 Å of the red patch, in order of highest degree of contact to lowest degree of contact (name followed by frequency), are listed in table 5.4. Experiments have shown the tricosapeptide C^{420}FEDCRTGAW**RH**GR**IH**IR**I**AK**M**D^{442} is required for full interaction, and all of these amino acids are found near the hydration sites (the amino acids shown in bold are of greater importance in the interaction). For example, Met 441 is not found in the 'close' binding but has increasing impact as distance from the protein is increased. This amino acid may help influence GABARAP binding at far distances drawing the two proteins together. Arg 430 is more contacted at close distances; this may help GABARAP settle the bind once it is close.

Patch	Mean	Median	Std Dev	Number
Close (red)	−36.8	−39.0	5.0	244
Close (orange)	−40.5	−41.4	2.1	113
Close (yellow)	−40.8	−41.6	1.8	238
Regular (red)	−31.4	−38.8	13.6	246
Regular (orange)	−41.2	−41.5	0.8	76
Regular (yellow)	−40.1	−41.1	2.5	308
Regular (green)	−40.8	−41.2	1.0	119
Far (red)	−35.0	−40.6	11.1	252
Far (orange)	−40.5	−41.1	1.5	70
Far (yellow)	−38.0	−40.2	5.6	365
Far (green)	−39.4	−41.3	4.1	294

Table 5.3 Table of displacement statistics for clusters of GABA_A receptor hydration sites featuring in the top set, i.e. those which are most hydrophobic. The units of all statistics are in kJ/mol except the number of patches combined to make the patch. The patches are those displayed in figure 5.5.

5.5.3 GABARAP hydration

Table 5.5 and figure 5.8 show the location of the main hydration patches on the surface of GABARAP. It is useful to divide these patches into two: those with known binding proteins and those without. We define two kinds of hydration sites, 'overlapping' sites where the hydration patch is directly over the binding face of the protein, and 'surrounding' sites where the hydration patch is near the binding face of the protein. Note that these GABARAP hydration sites use the same colour codes used the GABA_A receptor hydration sites, this does not imply a connection.

Table 5.6 shows the hydration patches involved in GABARAP binding to known proteins, and the patches probably involved in GABARAP oligomerisation. The GABA_A receptor γ2-subunit binds GABARAP with site 33 (orange) as the overlapping site, and sites 11 (red) and 12 (purple) as the surrounding sites [111–113]. Calreticulin probably binds to two hydrophobic pockets [130]; for hydrophobic pocket 1, the overlapping site is site 32, and the surrounding site is site 33. For hydrophobic pocket 2, the overlapping site is site 33, and the surrounding site is site 42. The key GABARAP amino acids involved are Ile 21, Tyr 25, Ile 32, Lys 46, Lys 48, Tyr 49, Leu 50, Phe 60 and Leu 63 (PDB dataset 3DOW). The ALFY dodecapeptide [131] binds to GABARAP overlapping sites 32 and 33, and surrounding site 11 (PDB dataset 3WIM). The KBTBD6 undecapeptide [132] binds to GABARAP overlapping sites 11, 32 and 33, and surrounding sites 12 and 41 (PDB dataset

Name	Close	Regular	Far
Cys 420	8	0	0
Asp 423	4	0	0
Cys 424	8	1	1
Gly 427	5	5	6
Ala 428	9	8	8
Arg 430	19	13	13
His 431	17	17	17
Gly 432	1	1	1
Ile 434	17	16	16
His 435	16	17	17
Ile 436	9	9	9
Arg 437	1	1	13
Ile 438	17	17	18
Ala 439	0	5	5
Lys 440	0	0	2
Met 441	0	15	17
Ser 443	0	6	2

Table 5.4 The frequency of occurrence of amino acids within 5 Å of the most displaceable (red) water patch. The amino acids shown in bold are of particular importance in binding GABARAP to the GABA$_A$ receptor.

4XC2). The K1 dodecapeptide [114] binds to GABARAP overlapping sites 32, 33, 41 and 42 and surrounding site 11 (PDB dataset 3D32). From the data from Coyle et al. [113], we also suggest that site 43 is involved in GABARAP dimerisation. Lastly, the key tubulin-binding amino acids in GABARAP are residues 10–22. Tubulin binds GABARAP with sites 13, 31 and 32 as the overlapping sites, and site 11 as the surrounding site.

There are a large number of hydration sites not involved in the binding of these three proteins. However, when we examine the crystallographic datasets, we find that these sites are involved in dimerisation or trimerisation. It is still unknown how GABARAP dimerises in the cell, so it is uncertain if these crystallographic oligomers represent the natural state of oligomerisation. Table 5.6 also shows the sites involved in GABARAP-GABARAP interfaces ('self-interaction'). Note that Coyle et al. [113] suggested a dimerisation face for GABARAP, but since no related PDB dataset has been reported, we have deduced the overlapping site from figure 1 of the paper by Coyle et al. [113]. Moreover, the dimerisation suggested involves the N-terminal amino acids 'swinging out' to produce the 'open' form of GABARAP; this 'open' form of GABARAP normally exists in a dimer form, and also simultaneously binds tubulin and the GABA$_A$ receptor. We have access only to structural

Name	Colour	GABARAP Nearby Residues (≤ 3Å)	Mean	N_{hs}	K	G
11	red	D45 E8 E17 H9 K13 K47 K48 Y5	6.1	24	1	1
12	purple	A36 A39 R67 D43 D45 E34 G42 I41 L44 K35 K2 K47 F3 Y115 Y5 V4	8.8	23	1	1(2)
13	green	R14 D102 E100 H99 L105 K6 F104 F11 Y106	11.3	10	1	(1)3
21	blue	N82 I84 L117 K2 K38 M1 P37 S113 Y115 V114	6.8	25	2	1(4)
22	cyan	A75 R40 D111 D74 E112 G116 L117 S110 V114	6.7	25	2	1
31	yellow	R15 R22 E101 E12 E19 K13 K23 F103 F11 P10 16	4.5	42	3	(1)34
32	pink	D27 E17 I21 K13 K20 K24 K48 P26 Y25	4.5	31	3	(1)234
33	orange	R67 D45 L50 L63 K46 K66 Y49	5.5	16	3	2(4)
41	tan	R28 D27 P52	5.4	8	4	2(4)
42	d. grey	L63 K66 F62	3.8	10	4	24
43	silver	Q93 E97	3.8	13	4	23
51	mauve	E73	3.1	9	5	-

Table 5.5 A guide to the locations of the hydration patches on GABARAP. Residues within 3 Å are listed. The mean displacement free energy (in kJ/mol) of hydration sites at that site and the number of hydration sites in the patch. The numbers under the sections K (1KOT) and G (1GNU) indicate which pass of the hydration site search these regions are highlighted. The sites are displayed in figure 5.8.

data of the 'closed' form of GABARAP so the overlapping site identity is less certain than other sites.

Some of the PDB files analysed were analogues of the GABARAP protein. Further analysis was made for these similar proteins, but some residues have changed. The clusters associated with these are shown in table 5.7. The proteins examined were 5LXH, 5LXI, which are described as GABARAP-L1 ATG4B LIR Complex, 5DPS which is described as the crystal structure of PLEKHM1 LIR-fused human GABARAP$_2$-117, and 4CO7 which is the crystal structure of human GATE-16. GABARAP-L1 is a GABARAP 'like' protein and has some smaller mutations. GATE-16 is another ubiquitin like protein which is moderately different to GABARAP.

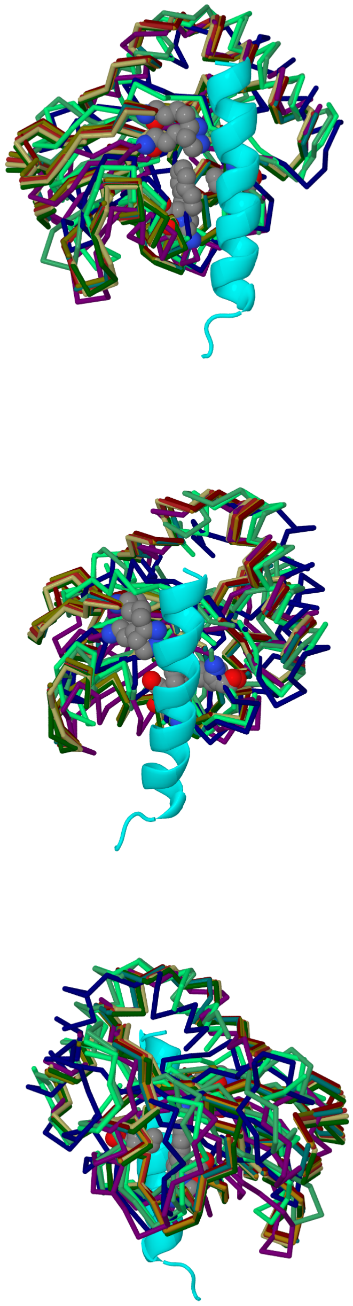


Fig. 5.4 Diagrams showing the eleven proposed docks; they were selected from the SwarmDock results, according to criteria from experiments. The three panels show alternative views of the docking. A section of the γ2-subunit is shown in cyan, and the eleven docked poses of GABARAP shown in different colours. GABARAP amino acids Lys 46, Val 51 and Phe 60 are highlighted in space-filling models coloured according to atom identity. The extracellular space is towards the upper part of the diagram. In the top and middle panels, the angle of view is from the ion channel towards the outside of the receptor. In the bottom panel, the angle of view is from outside the receptor towards the ion channel.

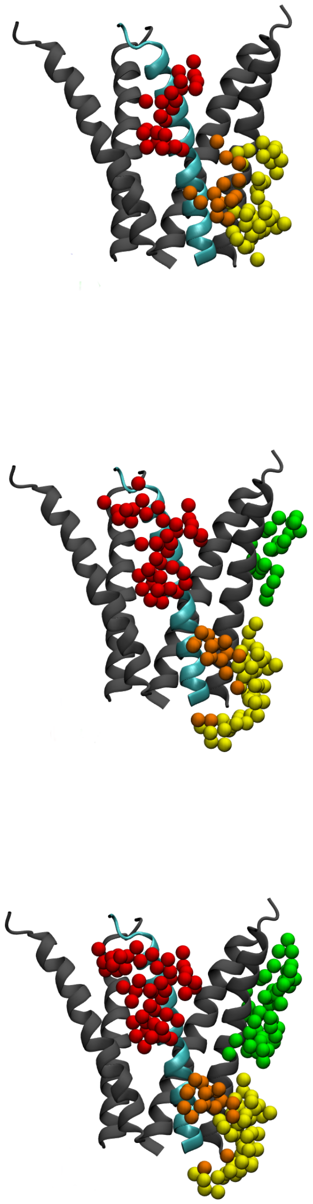


Fig. 5.5 Diagram showing a model of the intracellular helices of the GABA$_A$ receptor; the γ2-subunit is shown in cyan. In the top panel, the hydration sites from the best ‘close’ clusters of sizes 7–18 as red, orange and yellow spheres. In the middle panel, the ‘regular clusters are shown, while in the bottom panel, the ‘far’ clusters are shown. The hydration sites are shown in colour as described in table 5.3.

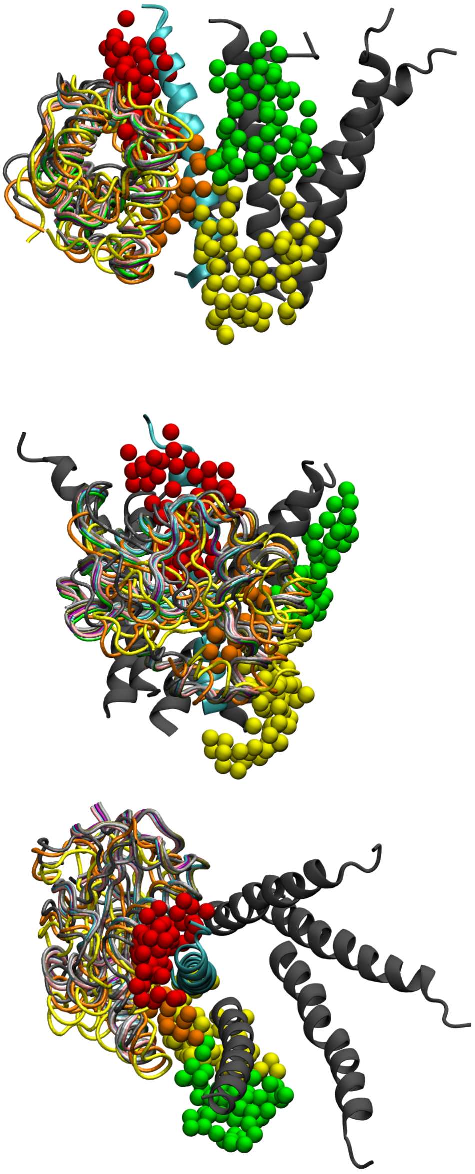


Fig. 5.6 Diagrams comparing the overlaid main chains of predicted docking positions of GABARAP (multiple colours), and all the hydration sites (red, orange, yellow, green) identified in this work from ‘close’, ‘regular’ and ‘far’ searches. The γ2-subunit is shown in cyan. The hydration sites are shown in colour as described in table 5.3.

GABARAP binding another protein		
Protein	Overlapping Sites	Surrounding Sites
GABA$_A$-R γ2-subunit	11 32 33 41 42	
Calreticulin hp-1 (3DOW)	32 (11)	33
Calreticulin hp-2 (3DOW)	33	42 (12)
ALFY peptide (3WIM)	32 33 (41) (42)	11 (12)
KBTBD6 (4XC2)	11 32 33	12 41 (13) (31) (42)
K1 (3D32)	32 33 41 42 (12)	11 (13) (31)
Tubulin ([130])	13 31 32	11
GABARAP ‘self-interaction’		
PDB Chain(s)	Overlapping Sites	Surrounding Sites
dimerisation ([113])	43	
4XC2 AC	42	(33)
4XC2 AD-BC	21 22	(12)
4XC2 CA	[other, weak]	
4XC2 CB	32 (33)	(31)
4XC2 DA	32 33	11 12 41
4XC2 AH-BG	21 22	12
4XC2 CE	[other]	(42)
3D32 AB	[other] 22 51	21
3D32 AD	[other] 21 22 51	
3D32 BA	32 41	31 (11) (33)
3D32 BD	[other] 32 33 42	11 12 41
3D32 BC	21 22	

Table 5.6 Dictionary of hydration patches used for protein-protein interactions from PDB files related to GABARAP. The first half of the table lists the interaction between GABARAP and another protein, with the relevant PDB dataset or relevant publication shown in parenthesis. The second half of the table lists the interaction between GABARAP molecules (‘self-interaction’) in any oligomer; the relevant PDB dataset or relevant publication is listed with the chains involved. Parentheses () around a site number means it is partial. [other] indicates that lots of the amino acids are not near a hydration patch.

GABARAP binding another protein		
PDB Chain(s)	Overlapping Sites	Surrounding Sites
5LXH AE-BF-CG	11 32 33 (41) (42)	(12) (13)
5LXH AB-BA	12 33 (42)	11 (22)
5LXH AC	[other] 42	(21) (22) (33) (51)
5LXH CA	21 22	12
5LXH BC	11 31 (32) (13)	
5LXH CB	[other] 43 (21)	13 31
5LXI BE	11 32 33 41 (42)	12 (13) (31)
5LXI BC	13 31	(11)
5LXI DC	11 32 33 41	13 31 (12) (42)
5LXI BD	13 31	(11) (32)
5LXI DB	11 12 33	(31) (32)
5DPS AB	22 51	21 (12)
5DPS AC	[other]	21 22
5DPS BA	[other]	21 22
5DPS CA	22 51	12
4CO7 AB	13 31	
4CO7 BA	[other]	

Table 5.7 Dictionary of hydration patches used for protein-protein interactions from PDB files related to GABARAP. The table lists the interaction between GABARAP like proteins ('self-interaction') in any oligomer; the relevant PDB dataset or relevant publication is listed with the chains involved. Parentheses () around a site number means it is partial. [other] indicates that lots of the amino acids are not near a hydration patch.

5.5.4 Summary

Using SwarmDock and subsequent filtering based on available experimental evidence, we have identified 11 docked poses of GABARAP. These docked positions are all very similar, and they are all in contact with highly-displaceable GABA_A receptor hydration sites. We note that the GABA_A receptor amino acids in table 5.4 match those in table 5.2 very well. Hydration analysis of water molecules around GABARAP has identified a large number of possible binding sites, and some of them are found to match the binding face for the GABA_A receptor $\gamma2$-unit intracellular domain. Figure 5.7 shows a global comparison of the results from docking and from hydration patch analysis.

However, in both cases, we have discovered hydration patches that might suggest a binding site, but we could not find any known binding molecule. In the case of the GABA_A receptor intracellular domain, there are hydration patches next to the $\beta2$-subunit (green and yellow patches in figure 5.5) which are distant from the GABARAP-binding site, and do not seem to bind any known protein. In the case of the GABARAP, we have discovered hydration patches which suggest binding sites, but we could not find any protein that binds. Some of the GABARAP hydration patches are involved with binding tubulin, calreticulin and various other peptides, though there is some degree of overlap between the GABA_A receptor binding site and the site for other proteins. It is interesting to note that the first-pass and third-pass sites are often involved in binding autophagy-related proteins, but the second-pass sites are used for dimerisation and trimerisation under crystallography conditions. Figure 5.8 also shows the hydration patches classified around GABARAP. The hydration patches from the 1GNU structure do not exactly match those from the 1KOT structure; the patches are defined by the 1KOT structure. Nevertheless, table 5.5 shows that the first-pass and second-pass sites around 1KOT and 1GNU are very similar. Moreover, all the possible locations for hydration are identified in both cases, though they appear at different passes.

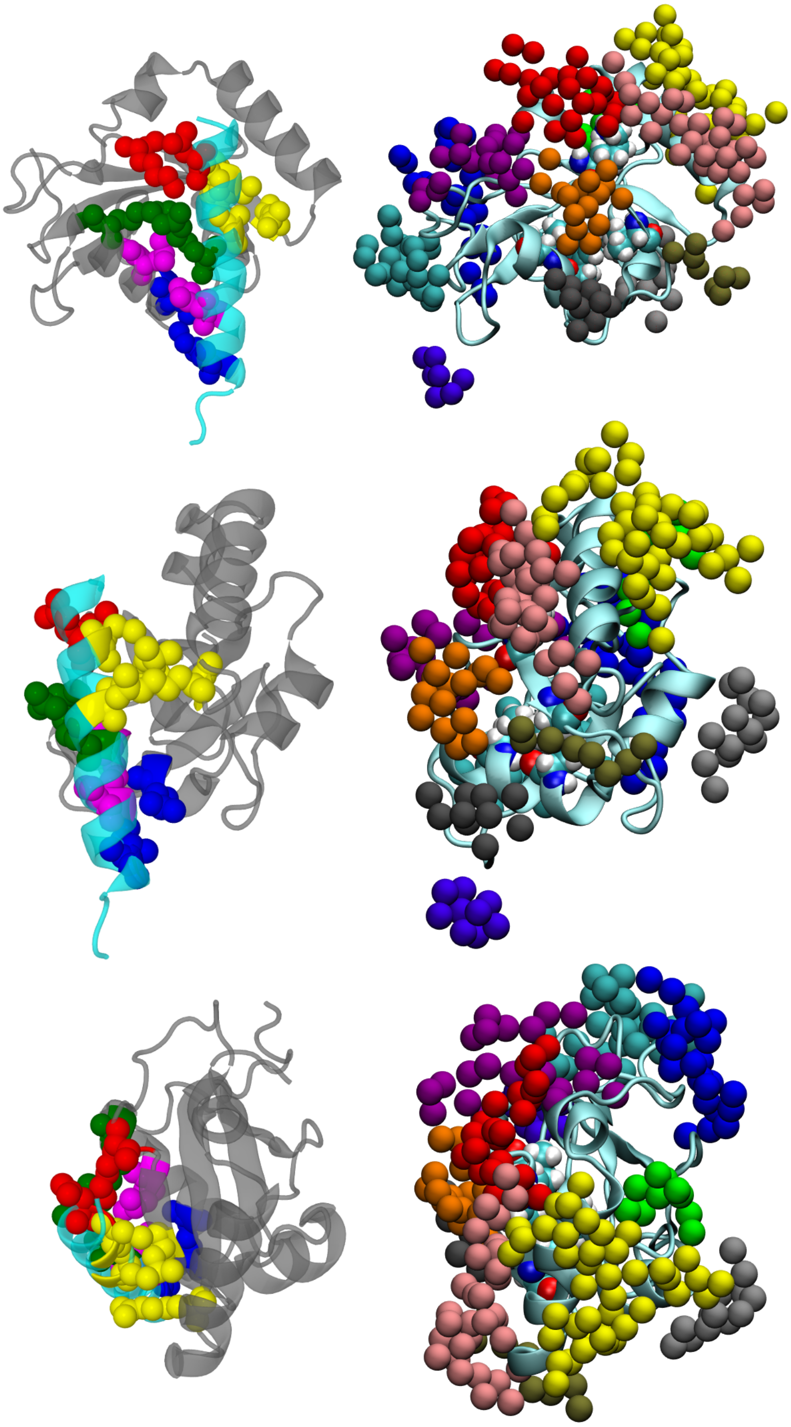


Fig. 5.7 GABARAP with the hydration sites listed in table 5.5. The CPK-coloured atoms are from residues Lys 48, Val 51, Phe 60 and Ile 64. The three panels on the left show the docking results with the intracellular MA helix of the γ2-subunit of the GABA$_A$ receptor present (transparent blue), and the three panels on the right show the results from hydration patch analysis. The angles of view on each row for the two structures are identical.

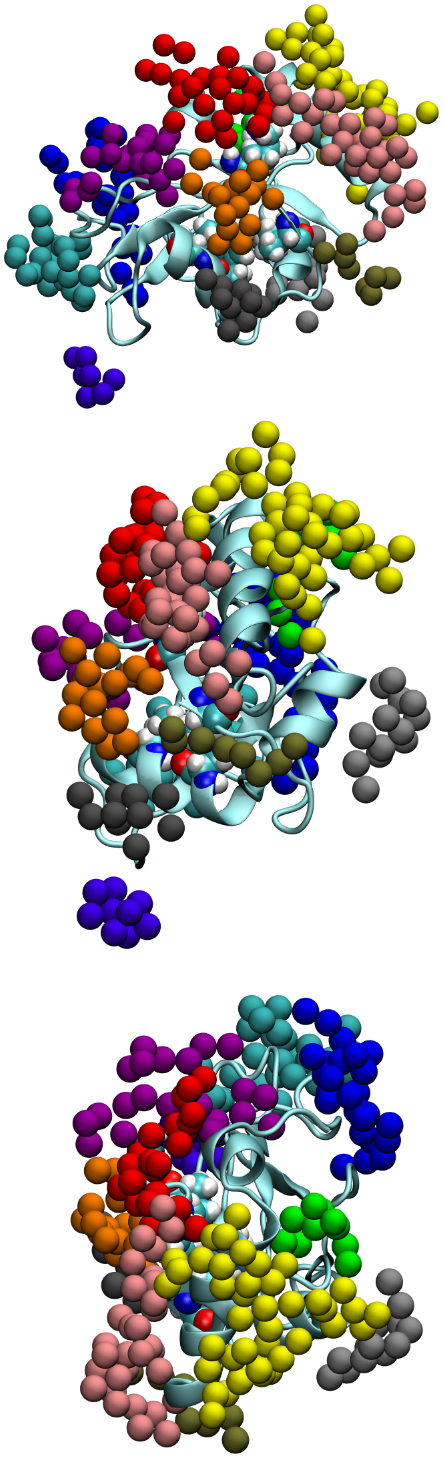


Fig. 5.8 GABARAP with the hydration sites listed in table 5.5. The CPK-coloured atoms are from residues Lys 48, Val 51, Phe 60 and Ile 64. The angles of view of these three panels are approximately the same as those for the three panels in figure 5.4.

5.6 Discussion

Cys-loop ligand-gated ion channels often interact with cytoplasmic proteins, and this interaction serves many purposes, amongst them the clustering of ion channels and the modulation of channel function.

One of the best studied examples is the interaction between the nicotinic acetylcholine receptor (nAChR) and the cytoplasmic protein rapsyn. Rapsyn has a molecular weight of about 43000 [133], and it interacts with the intracellular domain of the nAChR [134]. Electron microscopy showed that the nAChR are interconnected by rapsyn dimers. Up to three rapsyn dimers can contact each nAChR in specific regions in the nAChR intracellular domain. This tight network probably underlies the low mobility of nAChR in the plane of the cell membrane, and also allows nAChR to be concentrated at the neuromuscular junction motor end-plate [134].

The interaction between the glycine receptor and gephyrin has been studied experimentally. Gephyrin was first identified as a protein which bridged the glycine receptor and tubulin [135]. Sola et al. [136] co-crystallised a segment of the glycine receptor β-subunit and a partial dimer of the cytoplasmic protein gephyrin (Protein Data Bank code: 1T3E). They were able to resolve the structure of a pentapeptide portion of the glycine receptor β-subunit and the gephyrin domain E dimer. They proposed a network of gephyrin molecules linking the glycine receptors. Unfortunately, only the structure of five amino acids of the receptor was resolved, so it is difficult to draw any conclusion from this dataset.

Gephyrin also interacts with the GABA$_\mathrm{A}$ receptor through its α2-subunit [137] and α3-subunit [138]. It is unclear if gephyrin binds the α1-subunit of the GABA$_\mathrm{A}$ receptor; some experiments failed to show any interaction [139], but others showed a weak interaction [140]. Maric et al. [141] co-crystallised segments of the α3-subunit of the GABA$_\mathrm{A}$ receptor with segments of gephyrin, and identified the undecapeptide T^{367}FNIVGTTYPIN^{381} from the GABA$_\mathrm{A}$ receptor as important for interaction with gephyrin. They showed that there were similarities between the binding of the GABA$_\mathrm{A}$ receptor and of the glycine receptor to gephyrin: T^{367}FNIVGTT^{374} from the GABA$_\mathrm{A}$ receptor, and F^{398}SIVGSL^{404} the glycine receptor β-subunit adopted similar conformations.

Two other cytoplasmic proteins are known to interact with the GABA$_\mathrm{A}$ receptor: collybistin and GABARAP. Collybistin consists of two types, which consist of 413 and 493 amino acids, respectively [142]. Saiepour et al. [139] showed that collybistin interacted with the intracellular domain of the $\alpha2$-subunit of the $\mathrm{GABA_A}$ receptor, and its binding site for the $\alpha2$-subunit overlapped that for gephyrin. Collybistin was later shown to be important for clustering gephyrin and the $\mathrm{GABA_A}$ receptor [143].

GABARAP is a protein of 117 amino acids [110], and it binds specifically to the $\gamma2$-subunit of the $\mathrm{GABA_A}$ receptor. Coyle et al. [113] showed that GABARAP also binds tubulin, and this is believed to position the synaptic $\mathrm{GABA_A}$ receptors correctly in the membrane. Binding of GABARAP to the $\mathrm{GABA_A}$ receptor caused receptor clustering [115, 116], so some of its functions are similar to gephyrin and collybistin. However, GABARAP is unique in that its binding also caused the conductance of the $\mathrm{GABA_A}$ receptor to increase from about 30 pS to 40 pS–60 pS, and the mean opening times from about 2 ms to about 6 ms [144]. It thus appears that gephyrin has more general actions on both the $\mathrm{GABA_A}$ receptor and the glycine receptor, and that the action of gephyrin and collybistin appear to be confined to receptor clustering. The action of GABARAP is more specific to the $\mathrm{GABA_A}$ receptor, and, in addition to receptor positioning, it also modulates the electrophysiology of this ion channel.

In this work, we have used a flexible protein-protein docking programme to identify the interaction between the $\mathrm{GABA_A}$ receptor and GABARAP. We have also used a novel method to predict hydration sites on the two proteins, and suggest docking poses. We have identified possible binding faces on the $\mathrm{GABA_A}$ receptor and on GABARAP. To confirm our theoretical predictions would require a high-resolution structure of the $\mathrm{GABA_A}$ receptor with an intact intracellular domain.

Some of the GABARAP binding faces we have identified are at the GABARAP/$\mathrm{GABA_A}$ receptor interface, but others are involved in binding other proteins. In addition, we have also identified possible faces not known to bind any protein. It is interesting to note that, in the case of GABARAP, hydration patches appear on five out of six faces of this protein. As so many interfaces are involved in different types of interaction, it is possible that the last face is not active to remove the burden of constraints on protein architecture.

Currently, this method we have used only examines the hydration properties around proteins. We could envisage including details such as shape and electrostatic properties, and develop a molecular docking method based on this hydration site survey.

The $\mathrm{GABA_A}$ receptors in neurons have different ion channel properties from recombinant receptors [145]. Luu *et al* [144] and Everitt et al. [116] showed that $\mathrm{GABA_A}$ receptor conductances in neurons is similar to that obtained from recombinant receptors associated with GABARAP. GABARAP is thus of importance in physiological functioning of the GABA$_A$ receptor in the central nervous system, and this underlies the importance of understanding the physiological role of the intracellular domain of this receptor. It would be interesting to investigate the interaction between GABARAP and the GABA$_A$ receptor further, to understand how GABARAP changes the ion channel functioning of the receptor. This would require a high-resolution structure of the GABA$_A$ receptor with an intact intracellular domain.

Two important questions should be raised:

1. How do we know that GABARAP and the GABA$_A$ receptor only have one docking pose?
2. How do we know that the docking is a direct face to face interaction as depicted in this study. Could there be an indirect interaction that leads to the same set of experimental observations?

The first question can be addressed as follows:

1. The final set of docking poses do show some slight differences. This could imply degrees of flexibility in the dock and small range of structures.
2. Otherwise, the SwarmDock results covered a large number of potential docks. After applying the criteria which, to our understanding, best reflect the experimental observations only one set of docks was left. Of course the sample pool of poses may not be wide enough, but we believe it to be thorough enough to rule out additional poses.

The second question leads on from this. How do we know the criteria imposed on the docks are enough?

1. Experiments conclude the central section of the $\gamma 2$ helix is involved. This arose not only from truncating both ends of the helix, but also NMR data showing the centre amino acids to be important.
2. The *most hydrophobic* patch from all 5 pentamers was situated in front of the $\gamma 2$ helix.

3. If there was a docking pose somewhere else, that could somehow lead to chemical shifts at the key amino acids on the helix, this binding location should have to be more energetically favourable that the hydration exclusion provided by the docking pose concluded by this study.

4. This situation is unlikely because according to the IFST data, there are no ’more hydrophobic’ patches local to the model of the IC domain. This would imply a non-local interaction which although possible, is not the most parsimonious summary of the findings of this study.

5.6.1 Main Conclusions and Future Work

The main conclusions of this work are:

1. We have found theoretical docking poses of GABARAP and the $GABA_A$ receptor IC $\gamma2$ helix which are supported by experimental evidence from a number of sources.

2. We have identified sites of interest on GABARAP and classified its interactions with other proteins known to interact.

3. This map will help to construct a larger complex of proteins, for example, dimerized $GABA_A$ receptors, co-bound to GABARAP, and connected to tubulin, which would help to infer exactly how the receptors are transported to the membrane and anchored into the cytoskeleton.

4. Other possibilities are models of gephyrin interaction with the IC domain of the $GABA_A$ receptor.

5. The docking pose can be validated in future work using simulated conductivity and binding studies.

6. This was the first demonstration of using hydration sites and clusters thereof to infer protein-protein binding. This method should show future use in predicting binding conformations of other protein pairs.

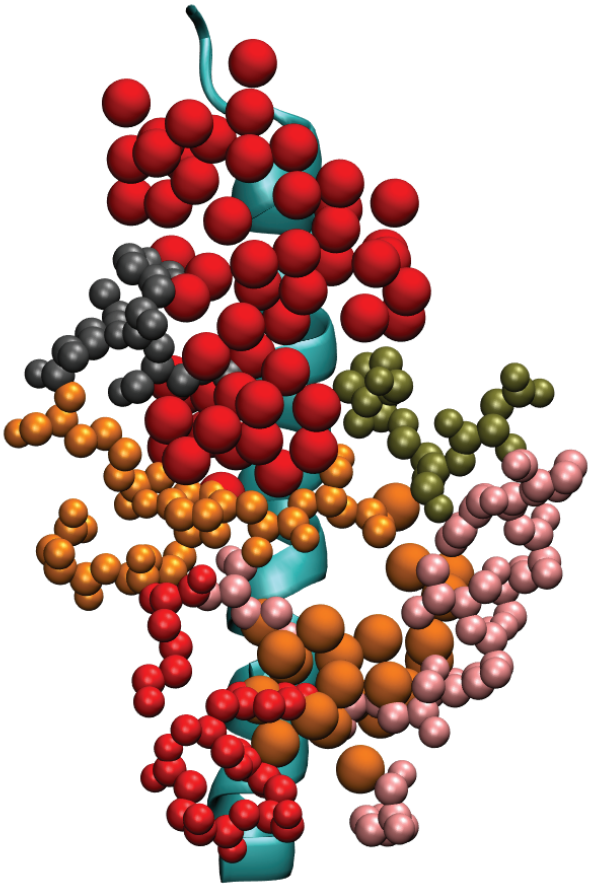


Fig. 5.9 Diagram showing the GABA$_A$ receptor red and orange hydration sites on the surface of its γ2-subunit. In this diagram, the GABA$_A$ receptor ‘close’, ‘regular’ and ’far’ red hydration sites, as described in table 5.3 are combined to give the red sites, and the ‘close’, ‘regular’ and ‘far’ orange sites are combined to give the orange sites. The GABARAP residues are coloured to correspond to their nearest sites according to the convention in table 5.5: GABARAP sites 11, 32, 33, 41 and 42 are involved in this interaction.

Chapter 6

Categorising Hydration Environments

6.1 Introduction

In the previous chapter the concept of a hydration site was used to quantitatively classify areas of hydrophobicity on the surface of a protein. The hydration free energy was calculated for each site using IFST as introduced in previous chapters. IFST gives an enthalpic and entropic breakdown of the site hydration free energy, as discussed in chapter 5 this breakdown goes further, decomposing the entropy into translational and orientational terms. Those terms can be broken down yet further into solute-water and water-water terms. In chapter 5, such an IFST calculation was applied to each hydration site around a protein which can also be viewed as a decomposition of the total protein solvation free energy, split into local regions. This further decomposition revealed the location of attractive patches on the $\gamma 2$ subunit of the GABA$_A$ receptor. These patches were associated with hydrophobic amino acids. From this and other results using the hydration sites method we can conclude there is a correlation between the 'local chemistry' of the solvated region and the free energy of water molecules around that region. The present chapter is divided into two parts. Part one will attempt to classify the hydration free energy of sites according to their local chemistry by looking at distributions of the properties of HS around many different proteins. Part two will focus on a related problem: in chapter 5 when the HS method was used, the easiest patches of hydration sites to displace were specifically hydrophobic areas. This is because the IFST calculations made at each hydration site only took into account the (de)solvation of a water molecule at that site. For sites which are close to polar parts of the protein, water molecules will be more likely to bind strongly in place, and have a large desolvation penalty [124]. In the problem of estimating ligandability, the water near a binding site will not only have to be removed, but also replaced with some atoms from the ligand. If the ligand also has charged atoms and is binding near a polar part of the protein there could be very favourable or unfavourable connections. This will translate to very different free energy of binding values. In some cases the gain in binding free energy could outweigh the desolvation penalty of the water molecules around the binding site. In these situations the previous method of using combinatoric searches on the hydration free energy becomes insensitive to highly ligandable sites. A new model is required to estimate general ligandable hotspots on the protein surface which treats polar regions of the protein in a more balanced way. Part two of this chapter will make progress towards rectifying this problem and propose an adaptation to the algorithm.

6.2 Motivation

By calculating the free energy of hydration of each HS using IFST for a range of proteins, statistical information can be gleaned about the average nature of water in the local protein environment. Examples of these local environments are carboxyl oxygen groups and amide nitrogen groups. Regions above and below the plane of amide groups which contribute unfavourably to hydration free energies have been noticed in IFST calculations on small molecules [34]. The geometry of such motifs is similar in protein amide residues and may contribute to the presence of highly displaceable hydration sites. A notable class of such sites that are connected with displacability are the so called dehydrons [146].

Dehydrons are biologically important structural motifs related to packing defects in the protein backbone which lead to solvent exposed hydrogen bonds [146–148]. They are highly sensitive to the local solvent environment through the changing Coulombic interactions associated with the displacement of water molecules and can be stabilised by bringing a hydrophobe toward to a backbone hydrogen bond which has fewer than average hydrophobic groups [146, 149]. On the other hand, they act as sticky sites which may become hydrated leading to conformational changes in the protein [149, 150]. The presence of dehydrons correlates with binding sites in proteins. They are important features in protein-protein complexation, protein-ligand interactions, high level structural arrangement and other biochemical phenomena and bioinformatical applications such as measuring proteomic complexity and selective inhibitor design [146, 148]. These examples support and motivate the goal of understanding the specific hydration of protein binding sites. Although dehydrons are specific to hydrogen bonding near the protein backbone, one would expect to find both analogous and alternative hydration features near other protein side-chains. Amide groups in asparagine and glutamine should share similar features to the protein backbone, whereas carboxyl groups in aspartate and glutamate residues will have different hydration properties.

Similar studies of hydration sites have been made on a small set of proteins in the past. Grid cell theory was used by Gergiokas et al. on 17 proteins which had large variations in structure [151]. These authors comment that for water molecules near polar or hydrophobic regions of the protein, stability strongly depends on the coordination of the local hydration environment. Beuming et al. used WaterMap (a GIST implementation) to conduct a distributional analysis on a larger set of 27 proteins [152]. These authors found that high energy hydration sites exist near backbone amide and other hydrophilic groups and concluded that hydration sites near the backbone of the protein are much less favourable than HS near polar side chains. They also found that the water molecules in the vicinity of secondary structures are less strongly bound. They also conclude that there is no direct relation between the energetics of a hydration site and the degree of burial of the HS within a protein. Both the strong dependence on coordination and the existence of high energy sites near the protein backbone agree with the work of Fernandez et al. on dehydron environments [146, 149, 147]. However, Fernandez et al. find that the degree of wrapping of hydrogen bonds near the protein backbone by hydrophobic groups strongly affects the stability of the protein, and the connection of this to the general 'degree of burial' is not clear.

Beuming et al. [152] found no correlation between the solvent accessible area and the hydration free energy of hydration sites. Whereas Gerogiokas et al. [151] state that the solvent accessible volume is large in binding sites. Gerogiokas et al. also state that there is no correlation between average water thermodynamic properties and the classification of a protein face as binding or non-binding [151]. However, Beuming et al. [152] state that regions with unstable HS indicate binding sites for drug-like molecules. This agrees with studies by Vukovic et al. that have reported success in identifying binding sites using HS for the bromodomains family of proteins [108]. Furthermore, there is a correlation between the net free energy of the most displaceable site on a protein as calculated using HS and an experimentally derived measure of 'ligandability' [153]. Irwin et al. extended this method to protein-protein interactions where HS with hydration free energies calculated using IFST revealed the location of attractive patches on the $\gamma 2$ subunit of the GABA_A receptor [107]. These patches were associated with hydrophobic amino acids and helped infer a docking pose for the two proteins. From this and other results using the hydration sites method we can conclude there is a correlation between the local environment of the solvated region and the free energy of water molecules around that region so that a detailed study of many hundreds of proteins will yield useful and interpretable results. To this end, this work applies IFST hydration site analysis to 380 proteins which gives substantially more data from which to not only understand HS in greater detail with statistical measures of their properties but also infer important properties that can be deduced from a knowledge of the HS. We attempt to classify the hydration free energy of sites according to their local environment by looking at distributions of the hydration free energy of HS around many different types of protein. From these distributions a notion of relative displacability of water molecules can be obtained which may be useful to medicinal chemists when designing ligands to displace certain targets. We also investigate a proxy for burial as the number of hydrogen atoms close to a HS, and the impact this has on HS stability.

6.3 The Dataset

380 proteins were simulated using MD. The water molecules in the simulation were clustered using the same algorithm used in chapter 5. The resulting hydration sites were analysed using the HS method also used in chapter 5. This gave a total of 357,467 HS with free energy, enthalpy and entropy values. Each hydration site is in a slightly different protein environment. This is a large data set which will contain useful information to characterise the distributions of water protein surface displacement free energies as a function of the local chemical groups. The largest group of proteins in the dataset belonged to the bromodomain family. This bromodomain part of the data set was used previously to test combinatoric searching for ligandibility prediction [108]. In total the data set has proteins from the 14-3-3 protein, Adenosine A2A receptor (A2AR), Acetylcholinesterase 1 (AChE), Angiotensin-converting enzyme (ACE), acetylcholine receptor, aldose reductase, APE-1, bromodomain, cAbl kinase, carbonic anhydrase 1, carbonic anhydrase 2, caspase 1 (ICE), cathepsin K (CTSK), cathepsin S, CDK2, cell division ZipA, cyclooxygenase 2, DNA gyrase B, EGRF kinase, enoyl reductase, factor Xa, fungal Cyp51, HIV-1 protease (HIV-1p), HIV integrase, HIV reverse transcriptase, HMG CoA reductase, HVC serine protinase, Inosine-5'-monophosphate dehydrogenase, inStem, kinase, K Ras, Mouse double minute 2 homolog, neuraminidase (NEU), P38 kinase, phosphodiesterase 4D (PDE 4D), phosphodiesterase 5A (PDE 5A), penicilin, Phaedon, PT phosphatase 1B (PTP1B),

serum albumin, Methylcytosine dioxygenase, thrombin, tyrosine kinase, urokinase and XTAL protein families. Some PDB entries had multiple structures and the analysis was performed on these replicates. A full table of PDB tags, conformer numbers and protein names and abbreviations is included in appendix A.

A specific subset of 13 families was chosen for the second part of the study. These were: HIV-1p, PDE5A, ACE, factor Xa, CTSK, NEU, CDK2, ICE, PDE4D, AChE, thrombin, c-Abl and PTP1B. These proteins were the subject of a study which measured the correlation of the best connected combinations of HS with a data driven metric for experimental ligandability [153]. This metric takes proteins with well tabulated K_d or K_a data and explores how well understood the binding process is and how completed the task of creating strong binders is. This information is derived from the number of recorded attempts at generating ligands, and the progressive improvement/success in binding.

6.4 Part I: Statistics of Water Sites by Chemical Group

Some general insights into the behaviour of water around a protein can be obtained by looking at general trends in the data from these HS. Histograms were made from the hydration free energy across all HS. The HS were split into groups based on the local chemistry including the number of potential hydrogen bonds that could be made and/or the number of close non-polar atoms (carbon and sulphur) that are near to the hydration site. The definition of a hydrogen bond and close atom in this context is purely based on a distance cut-off of 3.2 Å from the relevant O or N atom for hydrogen bonds and 4.5 Å for close heavy atoms (which could be of any type C,N,O,S). This method may be basic, but a number of more sophisticated treatments were tried to define hydrogen bonding, but these were not found to be effective. Previous studies have used cone angles of 120 degrees from heavy atoms to hydrogen atoms to define hydrogen bonds [154]. In the case of the dataset used in this work, there is only a single .pdb structure of each system. The positions of hydrogen atoms in this single structure may be misleading. Certain residues will have relatively stable hydrogen atoms, others may not and it is not clear how to treat these cases without further investigation. Some strategies tried were:

- Defining and parametrising cones around specific atoms in each amino acid side group for which hydrogen bonds were allowed.

- Defining planar constraints on rings and other flat parts of side chains e.g. histidine rings and amide $CONH_2$ triads.

- Using both cones and planar restraints simultaneously.

None of these advanced techniques of considering hydrogen bonding appeared to strongly filter sites which genuinely had hydrogen bonding activity from the sites which were very easy to displace. Hydrogen bonds range in energy and distance and are reported in the literature [155] of having length and energy properties within:

- 2.2-2.5 Å (strong, covalent)

- 2.5-3.2 Å (moderate, electrostatic)

- 3.2-4.0 Å (weak, electrostatic)

- 2-7 kcal/mol (in organic systems)

If a site does not fit in with these definitions, i.e. length less than 2 Å or energy less than 2 kcal/mol, then it most likely does not hydrogen bond with a neighbour. When either conical or planar restraints were applied checks were made to the resulting histograms to see if HS with energies below the energetic threshold were filtered.

The general distance characteristics of HS were investigated with respect to neighbouring atoms. Figure 6.1 shows that hydration sites with a hydrogen bonding atom as their nearest heavy atom neighbour are likely to be closer to that nearest heavy atom neighbour, which is to be expected. There are very few HS near O or N atoms with a distance of more than 4 Å. The 4.5 Å close atom cut-off is defined to keep the entire peak of the C and S atom distribution in figure 6.1. It can be seen that the range of distances under the hydrogen bonding atom curve is much wider than the curve for carbon and sulphur.

6.4.1 Defining Local Chemical Environments

The HS were first classified into 19 classes of local chemical environment. After this initial classification, 6 of these classes were considered to either have too few data points or be too similar to another class. The data were merged into 13 classes which are shown in table 6.1. The 6 classes that did not make final selection were, methionine sulphur, indole

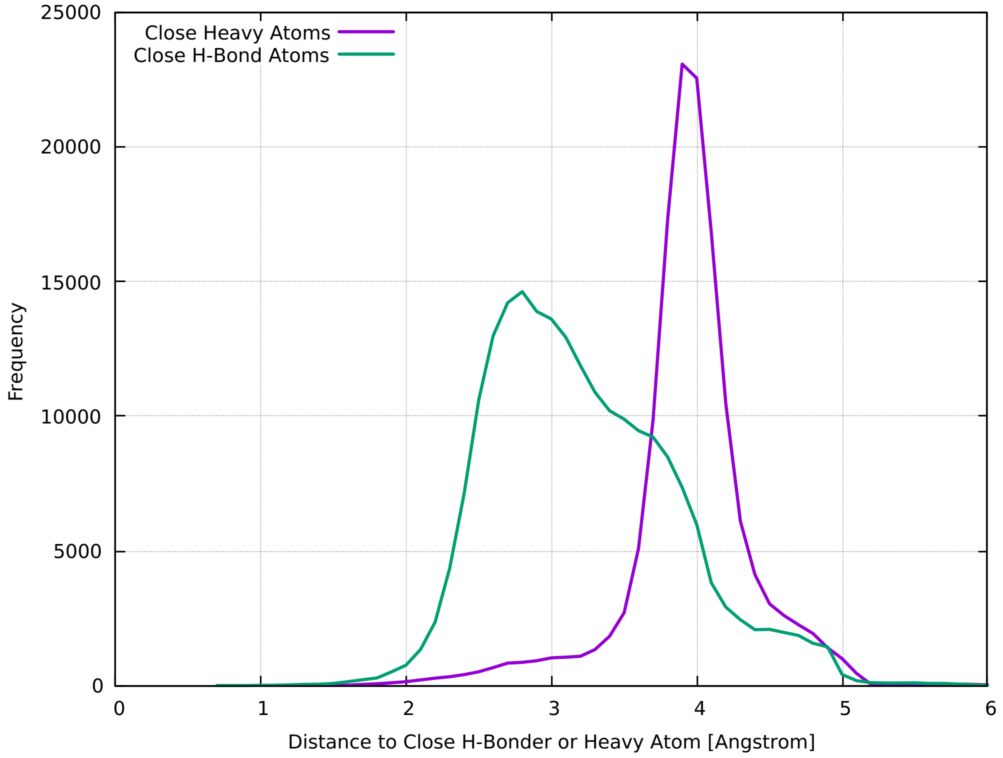


Fig. 6.1 This plot shows a comparison of the average nearest neighbour distance for hydration sites to their nearest O or N in the case of hydrogen bonding, or C or S in the case of heavy atoms. The sites near hydrogen bonders are generally closer to their nearest heavy atom.

nitrogen, neutral histidine donor nitrogen, positive histidine nitrogen, arginine chain nitrogen and oxygens on the terminus of the protein model, these classes were merged into the most similar class, which were cysteine sulphur, amide nitrogen, amide nitrogen, positive lysine nitrogen, positive arginine nitrogen and carboxylate oxygen respectively. The classes roughly cover all ‘interesting’ types of C,N,O and S atoms which are likely to show chemical differences.

There were also sites which had more than 1 hydrogen bonding atom or close atom (or both), which created a more complex environment. These were not split up into exhaustive pairs of the classes in table 6.1 otherwise the data per class in each of the combinations were too few. An additional labelling of sites is given in table 6.2, which covers all of the hydration sites in the data set. It can be seen from the count column in table 6.2 that sites with many hydrogen bonds are increasingly rare. This may be due to

Class	Definition
Amide Oxygen	HS < 2.4 Å of an O within amide part of Asn, Gln
Amide Nitrogen	HS < 2.4 Å of an N within amide part of Asn, Gln
Hydroxyl Oxygen	HS < 2.4 Å of Ser, Thr, Tyr OH group
Carboxylate Oxygen	HS < 2.4 Å of Asp, Glu =O group
Non-polar Aliphatic Carbon	HS < 3.2 Å of aliphatic C
Non-polar Aromatic Carbon	HS < 3.2 Å of aromatic C (His, Phe, Tyr, Trp)
Sulphur	HS < 3.2 Å of S on Cys, Met
Neutral Histidine Acceptor Nitrogen	HS < 2.4 Å of N on Neutral His
Lysine Nitrogen	HS < 2.4 Å of N on Lys
Arginine Nitrogen	HS < 2.4 Å of N on Arg
Backbone Carbon	HS < 3.2 Å of C on any backbone
Backbone Oxygen	HS < 2.4 Å of O on any backbone
Backbone Nitrogen	HS < 2.4 Å of N on any backbone

Table 6.1 The final set of classes used to define chemical environments around hydration sites. Sites with either 1 hydrogen bonding atom within range (for nitrogen and oxygen) or 1 close heavy atom in range (for carbon and sulphur) could be classified in this way.

multiplication of probabilities, but evidence will be shown later that indicates these sites are unstable. There are also many sites with no close atoms or hydrogen bonds.

6.5 Part I: Results

The results of the first part of this chapter are mostly histograms of the distribution of HS hydration free energies with respect to different chemical environments. The results will be discussed as these histograms are presented. All of the sites near different types of oxygen or nitrogen atoms were considered. These sites are expected to show the greatest activity according to the local chemistry as the water will be strongly affected by larger charges. The results presented in figure 6.2 would suggest that HS near amide nitrogens are in general more displaceable than for other types; the broad part of the distribution is around 2 kcal/mol more positive than the other types of nitrogen. A large population of very displaceable HS with free energies $-1 < \Delta G < 0$ kcal/mol can also be seen. This effect is somewhat less in sites near neutral histidine nitrogens but is still relatively pronounced and all classes appear to have some members in this range of energies. Backbone nitrogen sites show the deepest dip in free energy values between -3 and -1 kcal/mol. Based on the height of the peak at the broadest part of its distribution, backbone nitrogens are the

Count	#HB	HB Type	#CA	CA Type
71849	0	-	0	-
47463	0	-	1	Aliphatic C
5187	0	-	1	Aromatic C
951	0	-	1	Sulpur
32582	0	-	1	Not C,S
41737	0	-	2	Any
44716	0	-	>2	Any
2627	1	Amide Nitrogen	≥ 1	Any Type
5839	1	Amide Oxygen	≥ 1	Any Type
5521	1	Arginine Nitrogen	≥ 1	Any Type
4349	1	Backbone Nitrogen	≥ 1	Any Type
28943	1	Backbone Oxygen	≥ 1	Any Type
7541	1	Hydroxyl Oxygen	≥ 1	Any Type
9036	1	Lysine Nitrogen	≥ 1	Any Type
1298	1	Neutral Histidine Acceptor N	≥ 1	Any Type
27441	1	Carboxylate Oxygen	≥ 1	Any Type
49	1	Other	≥ 1	Any Type
16077	2	Any N,O	≥ 2	Any Type
3498	3	Any N,O	≥ 3	Any Type
763	>3	Any N,O	≥ 4	Any Type
357467	Total			

Table 6.2 Partitioning of the hydration sites into classes based on the number of hydrogen bonding atoms (O,N) within 3.2 Å denoted #HB and the number of heavy atoms (O,N,S,C) within 4.5 Å denoted #CA.

least displaceable of the nitrogen types. All of the distributions have a sharp decline as the hydration free energy becomes very negative.

Figure 6.3 shows that in general sites around the carboxylate groups (in aspartic and glutamic acid) are much harder to displace and show a broader distribution of hydration free energies. The hydration free energies that range from -3 to -14 kcal/mol would suggest that often two separate hydrogen bonds can be made. There is a very low chance of finding carboxylate sites with hydration free energies between -3 and 0 kcal/mol especially compared to the amide oxygen environment which appears to be the most displaceable common oxygen environment. This kind information could be used to alter the search for displaceable sites. If a carboxylate oxygen site was found with a hydration free energy of -2 kcal/mol the *relative* displaceability of the site is large; medicinal chemists may wish to target such sites if the surrounding environment was also favourable. Figure 6.3 also shows that the behaviour of backbone and hydroxyl oxygen sites is similar. There exist more very displaceable hydroxyl sites, but sites which are in the broad part of the distribution are marginally more stable than backbone oxygen sites.

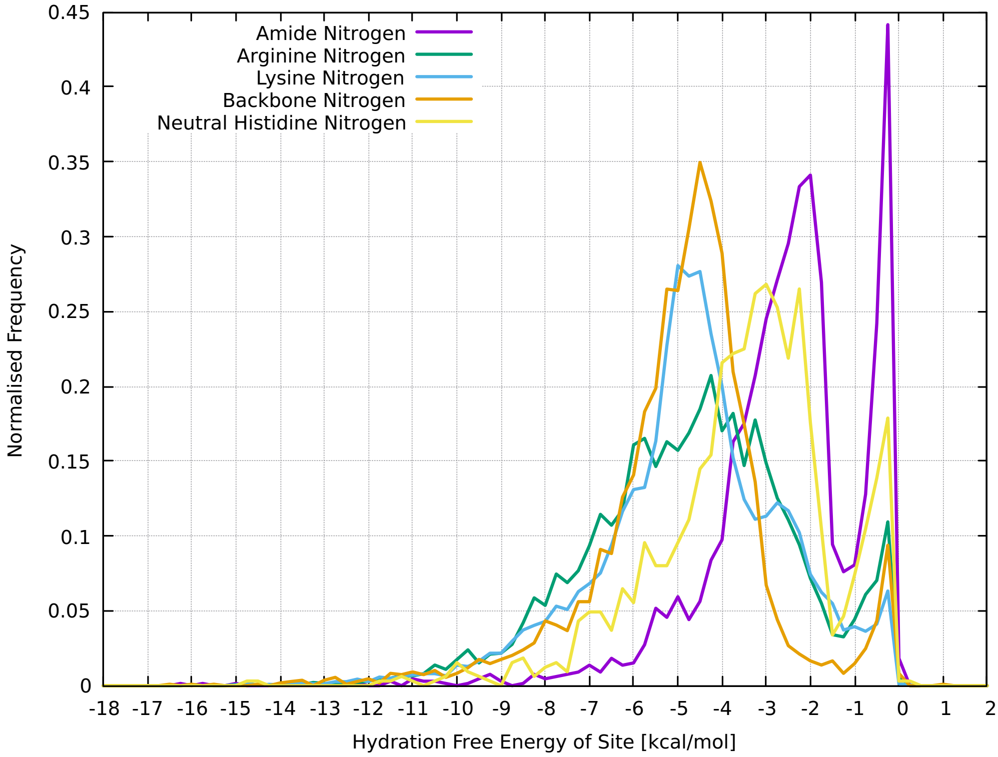


Fig. 6.2 Histogram of the hydration free energy values for the 5 main types of nitrogen. Amide and histidine nitrogens can be seen to be quite different environments for water.

Figure 6.4 shows the change in distribution of sites with no hydrogen bondable atoms nearby, but varying numbers of close atoms. The average behaviour of these sites is a sharp distribution with a long but quickly decaying left tail. These sites are fairly displaceable with energies only ranging from -1 to -4 kcal/mol. The peak of the average distribution is closer to 0 than any of the oxygen or nitrogen distributions in figures 6.3 and 6.2. This implies that having a single hydrogen bondable atom nearby improves the stability of a water molecule in the site. Increasing the number of close heavy atoms appears to strengthen the binding of water at a site. This can be seen in a leftward shift in the peak of the distributions in figure 6.4, along with a leftward broadening leading to very heavy tails and possibilities of sites with hydration free energies up to -7 kcal/mol in the case of more than two close atoms. This energy is comparable to the distributions for amide, hydroxyl and backbone oxygen sites and amide nitrogen sites. However, larger numbers of close atoms can also destabilise the site, creating very weakly binding water. This feature

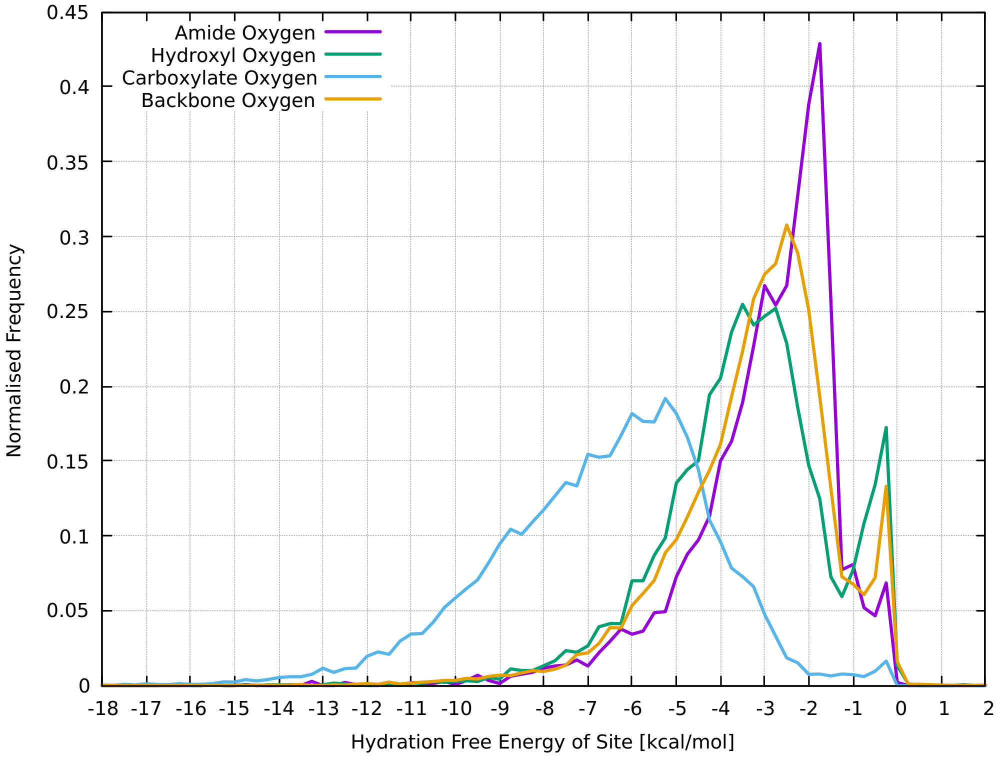


Fig. 6.3 Histogram of the hydration free energy values for the 4 main types of oxygen. Carboxylate oxygens can be seen to be quite different environments for water around proteins.

is present as a small sharp peak near 0 kcal/mol on the distribution with more than two heavy atoms only. It should be noted that the scale on figure 6.4 is different to figures 6.2 and 6.3 and care should be taken when comparing the plots.

Figure 6.5 shows that there is little difference in the free energy distribution of sites around different types of carbon and sulphur. This suggests that water will behave in a similar way in these sites, and carbon need not be partitioned into aliphatic and aromatic types when there are no potential hydrogen bonds nearby. The shape of the distribution is similar to that in figure 6.4 as this is essentially a further splitting of that dataset. There was not enough data to discern a difference between sulphur atoms on methionine and cysteine amino acids. All of these sites should be considered easy to displace.

Figure 6.6 shows that increasing the number of hydrogen bondable atoms around hydration sites in general leads to a rapid and extreme broadening of the distribution of hydration free energies. This plot mixes oxygen and nitrogen like sites and includes

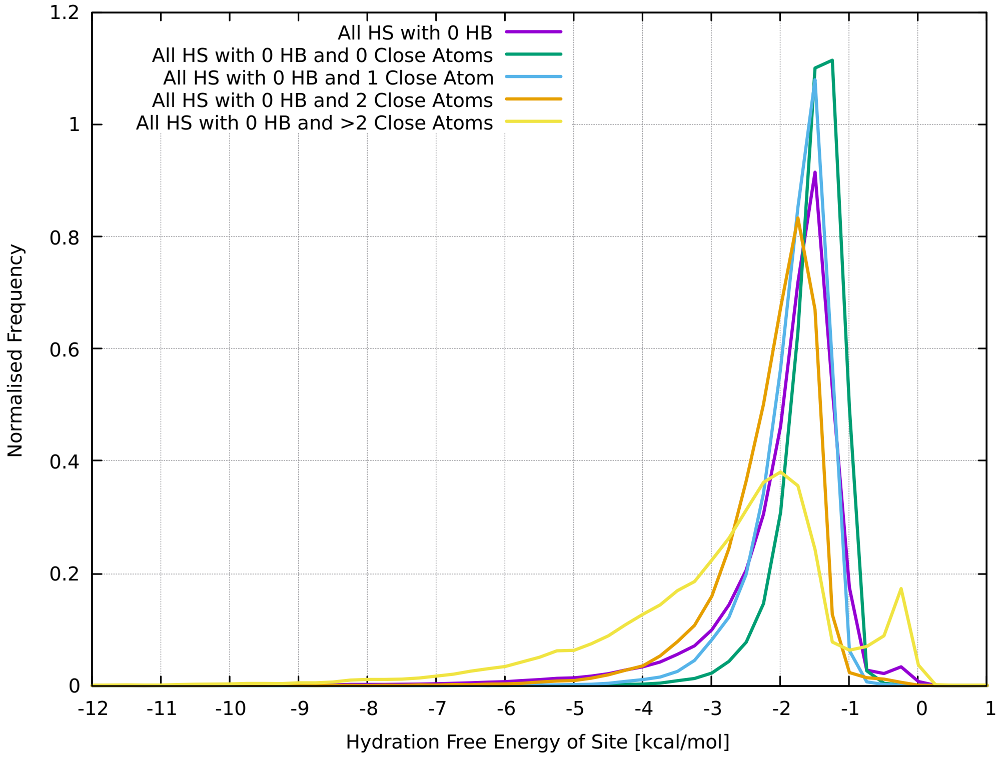


Fig. 6.4 Histogram of the hydration free energy values for the sites with no hydrogen bonders in range, but varying numbers of close heavy atoms.

mixed classes from table 6.2. Transitioning from zero to one hydrogen bond doubles the maximum expected free energy of the site from around −5 kcal/mol to −10 kcal/mol. Increasing from one to two hydrogen bonding atoms again adds −5 kcal/mol allowing sites with large −15 kcal/mol scores. However, the strongest effect when increasing the number of hydrogen bonding atoms is a very pronounced sharpening of distribution around the extremely displaceable sites with free energies between −1 and 0 kcal/mol. This implies that sites with many, possibly charged neighbours are unstable for water molecules. It is conceivable that these additional important degrees of freedom do lead to unpredictable behaviour. There could be a dynamic effect in which water is drawn to the site from afar, and once reaching the site quickly becomes expelled. Such dynamic effects would not be well captured by these distributions. Due to a lack of data counts for sites with many hydrogen bonding atoms (as seen from table 6.2) the distributions are noisy. However there is a strong flattening of density for higher distributions at very negative free energies. There is a prominent dip around −3 kcal/mol such that sites with one hydrogen bonder are fairly common, but sites with two hydrogen bonders are fairly rare. Once again

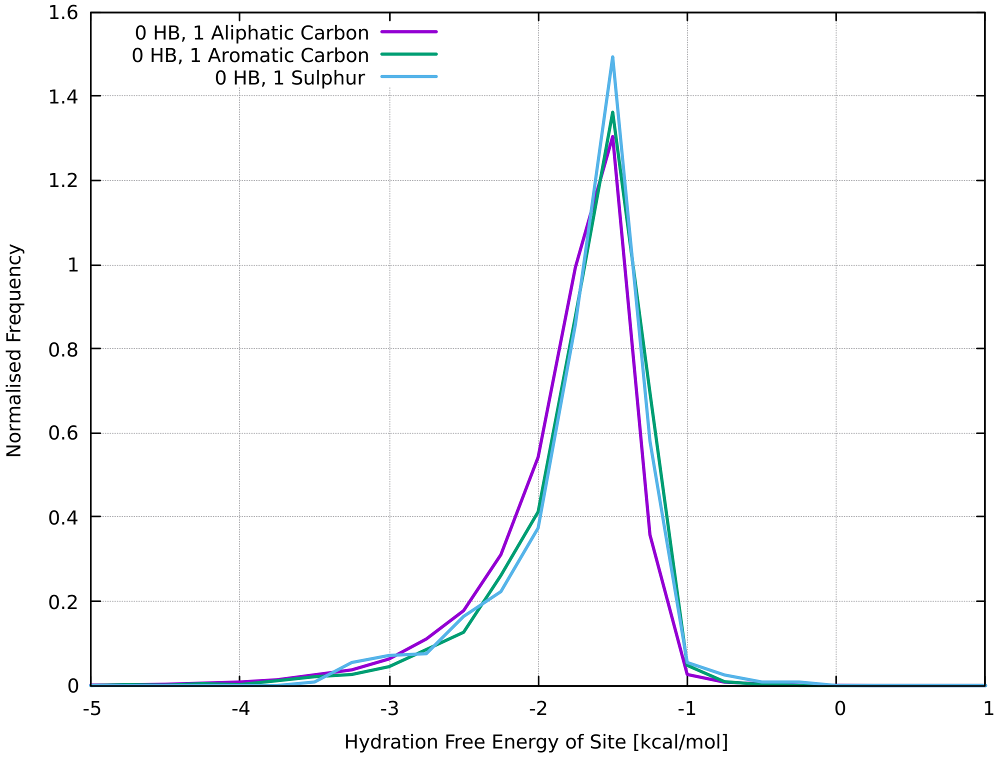


Fig. 6.5 Histogram of the hydration free energy values for the sites near a single carbon or sulphur with no hydrogen bonders in range. There is little difference in the behaviour of water around these atoms.

this kind of information could be put to use to assist the algorithm that finds ligandable sites. Finally it must be considered that sites with many hydrogen bonding atoms in the vicinity are more likely to be highly embedded within the protein. This will change the solvent exposure of the site and if a water molecule finds its way into such a site it may lose contact with the stabilising network of other water molecules. This could go some way to explaining the shapes of the distributions in this figure.

In order to investigate this destabilising effect further plots have been made with respect to an additional parameter: the number of hydrogen atoms within 4.0 Å of the site. This parameter will again correlate with the degree of protein embedding, with large numbers of hydrogen atoms found deep within the protein.

Figures 6.7, 6.8 and 6.9 show the behaviour of the hydration site free energy distribution as a function of the number of hydrogen atoms within 4.0 Å. There are additional plots for other chemical environments in appendix C. All of the plots show the same broad

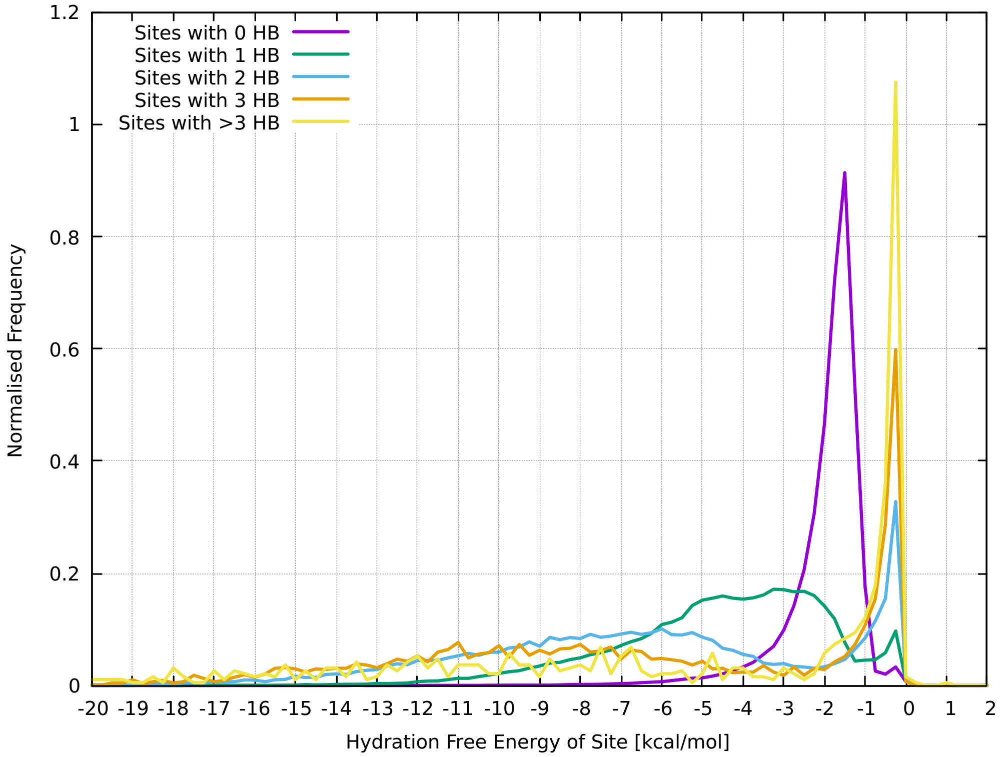


Fig. 6.6 Histogram of the hydration free energy values for the sites near varying numbers of hydrogen bondable atoms. Adding more hydrogen bonders allows much strong binding of the water, but can also lead to destabilisation.

behaviour. As the number of hydrogen atoms nearby is increased, the sites have a broader distribution of hydration free energies in which the peak of the distribution shifts leftwards and the left tail becomes increasingly heavy. Each additional hydrogen atom appears to shift the peak by approximately 0.5 kcal/mol. The backbone oxygen site distributions in figure 6.7 are very smooth as there were more data points in this class. The plots start from distributions for 0 hydrogen atoms up to 7, not all data sets start from 0 hydrogen atoms as there was very little data for either hydroxyl oxygen or arginine nitrogen like sites with 0 or 1 hydrogen atom. This is purely because the OH group or NH_2 groups always have hydrogen atoms nearby, whereas the backbone oxygen has a double bond. Once again these distributions all show an increase in extreme displacability with more neighbouring atoms, this effect appears reduced in the case of backbone oxygen in figure 6.7 but highly pronounced for hydroxyl and arginine nitrogen sites in figures 6.8 and 6.9. This comparison is not entirely fair as the data for 9 hydrogen atom neighbours is missing and the trend continues for higher number of neighbours which is not shown

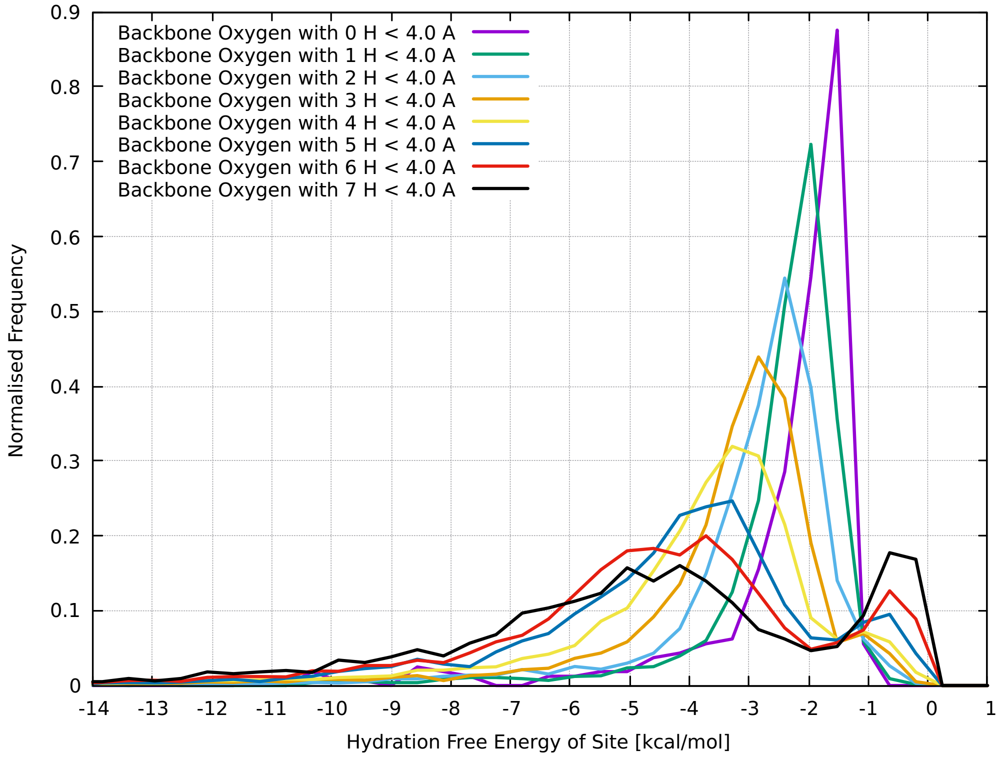


Fig. 6.7 Histograms of the hydration free energy for sites near a backbone oxygen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.

here. For each of the plots there is a region between −1 and −3 kcal/mol which becomes decreasingly occupied for a larger number of neighbours.

Figure 6.10 shows the hydration free energy of hydration sites as a function of the distance from the hydration site to the nearest O or N atom. This plot shows a few distinct features. Firstly, there is a spike of very displaceable sites with low hydration free energies (between −2 and 0 kcal/mol). These sites appear at all distances above 1 Å. Some of them should be considered very close to their nearest O or N atom (< 2.2 Å) with respect to expected hydrogen bond distances. The rest of the sites are broadly scattered across a wide region which is dense for energies ranging from 0 to −20 kcal/mol. It appears that sites with a positive hydration free energy, which are actively hydrophobic do not appear at distances less than 2.5 Å from the nearest O or N atom. This plot indicates that distance can be considered as a variable for filtering unusually displaceable sites.

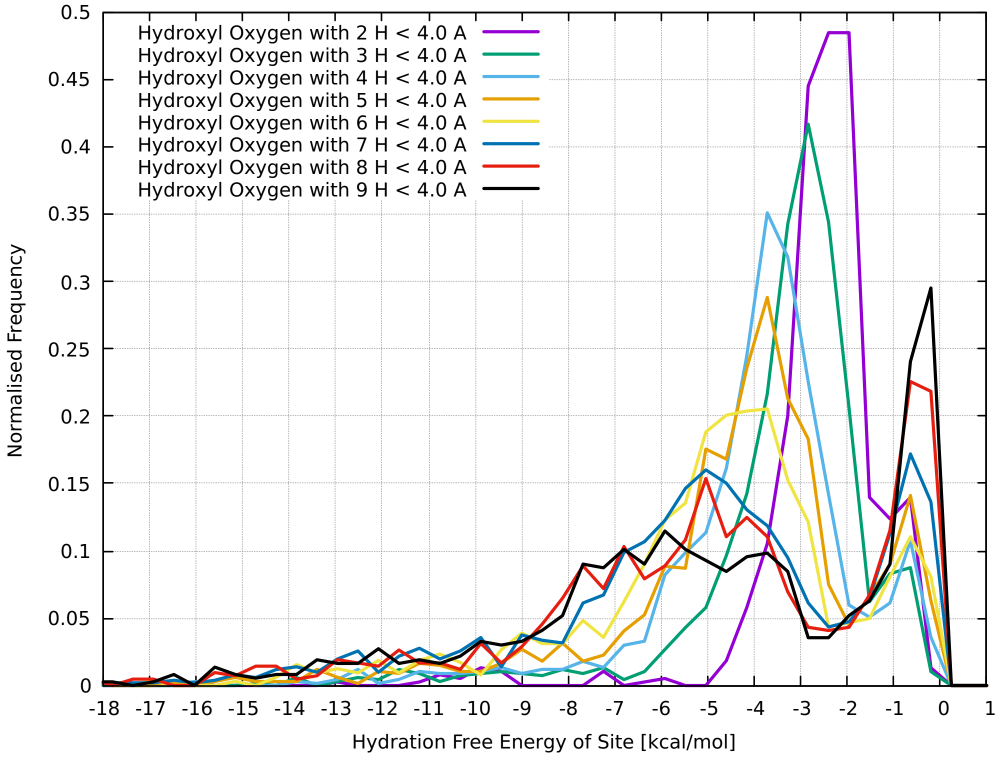


Fig. 6.8 Histograms of the hydration free energy for sites near a hydroxyl oxygen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.

The free energy of each site can also be decomposed into an enthalpy and an entropy. Figure 6.11 shows the distribution of the enthalpic component of the hydration free energy for backbone oxygen sites. The backbone oxygen class had many data points as it is contained in every amino acid. This resulted in a smoother distribution. The distribution follows a double exponential shape quite closely. The functional form of this distribution is simply

$$f(x) = a_1(-x)^{a_2}e^{a_3 x} + a_4(-x)^{a_5}e^{a_6 x}, \tag{6.1}$$

where the x terms have been negated due to the hydration free energies mostly ranging across negative values. The six parameters are $a_1 = 4.45$, $a_2 = 1.50$, $a_3 = 2.26$, $a_4 = 0.13$, $a_5 = 4.43$, $a_6 = 1.84$. There was no a priori reason for justifying this functional form and six parameters is fairly high for a one dimensional function. However, there is good agreement between $f(x)$ and the data which suggests it may be possible to parametrize distributions for the general hydration statistics of these sites if enough data were collected. Other

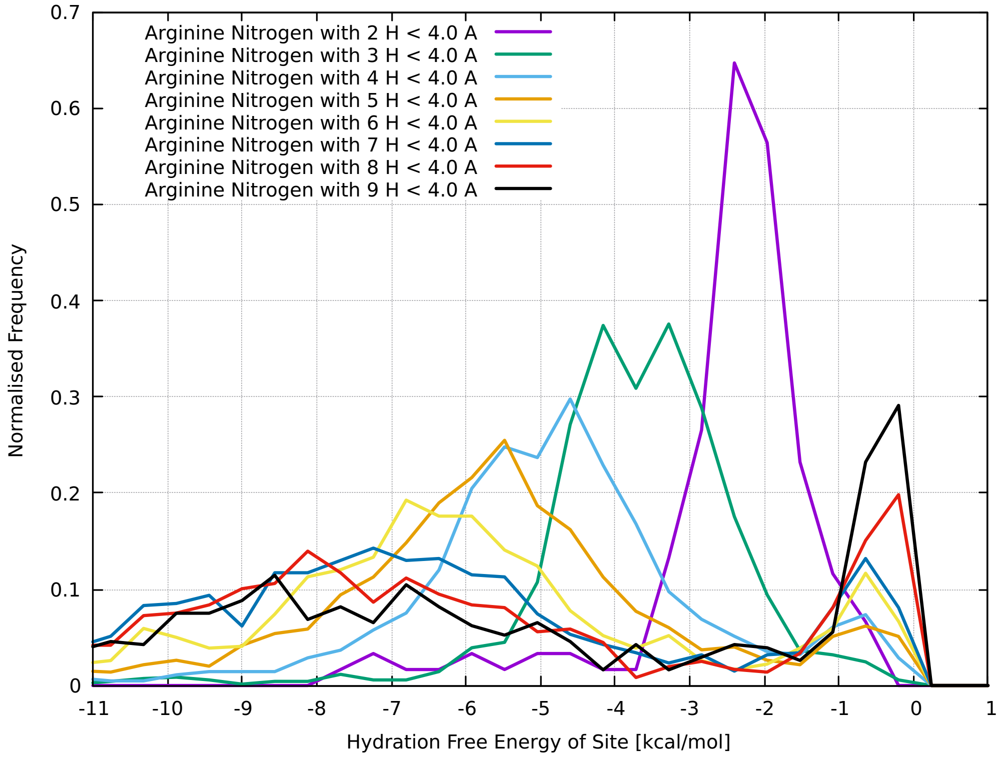


Fig. 6.9 Histograms of the hydration free energy for sites near a arginine nitrogen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.

classes had much noisier distributions. This distribution is for close oxygen atoms which extends the distance beyond hydrogen bonding atoms to 4.5 Å.

A fitting procedure can also be carried out for the $-T\Delta S$ term associated with the hydration free energy. Figure 6.12 shows the distribution of this entropic component of the hydration free energy, again for backbone oxygen sites. As in the enthalpic case, the backbone oxygen class had many data points. The distribution follows a Gaussian distribution somewhat closely, however there is some asymmetry in the tails which is not captured by the Gaussian. This again suggests a possible parametrization of distributions for the general hydration statistics of these sites if enough data were collected. The fitting routine used the function

$$g(x) = b_1 e^{-\frac{(x+b_2)^2}{b_3^2}} \tag{6.2}$$

with the parameters taking the values $b_1 = 1.9$, $b_2 = 1.2$ and $b_3 = 0.281$. This distribution only uses three parameters, however the fit is not as close as the six parameter distribution

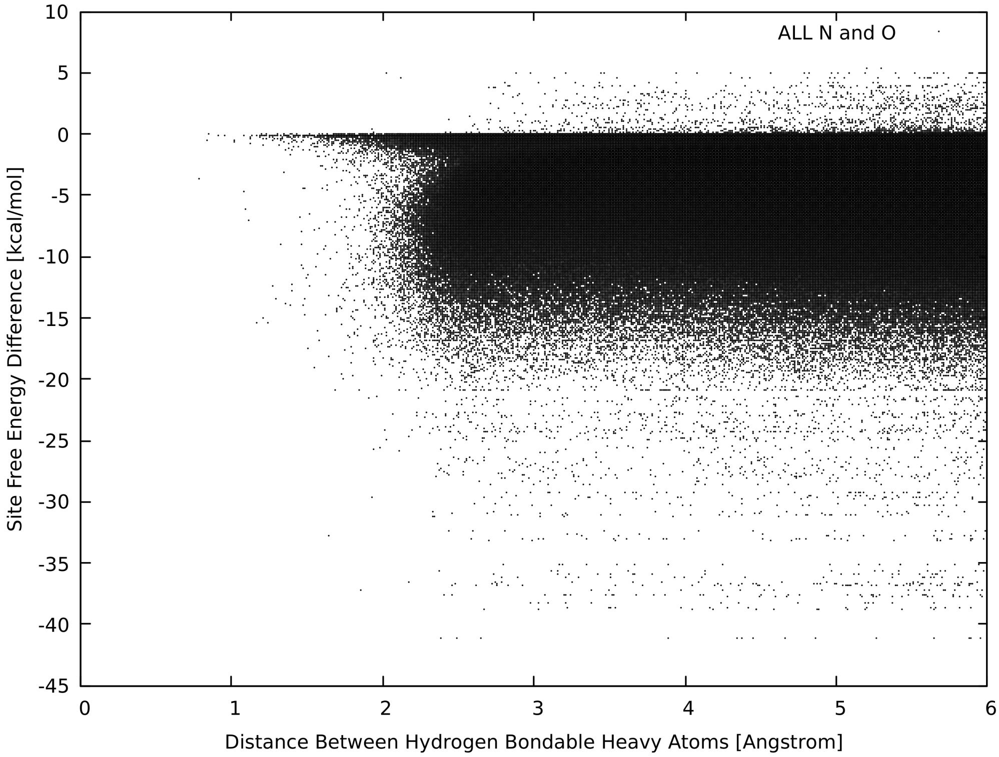


Fig. 6.10 This plot shows the hydration free energy of sites as a function of how far they are from their nearest nitrogen or oxygen atom. A peak can be seen consisting of very displacable sites which are noticeably closer to their respective heavy atom than the majority of sites.

used for the enthalpy. A resulting analytic form for the distribution hydration free energy, $h(x)$, could be found by the convolution of the enthalpy and entropy term distributions.

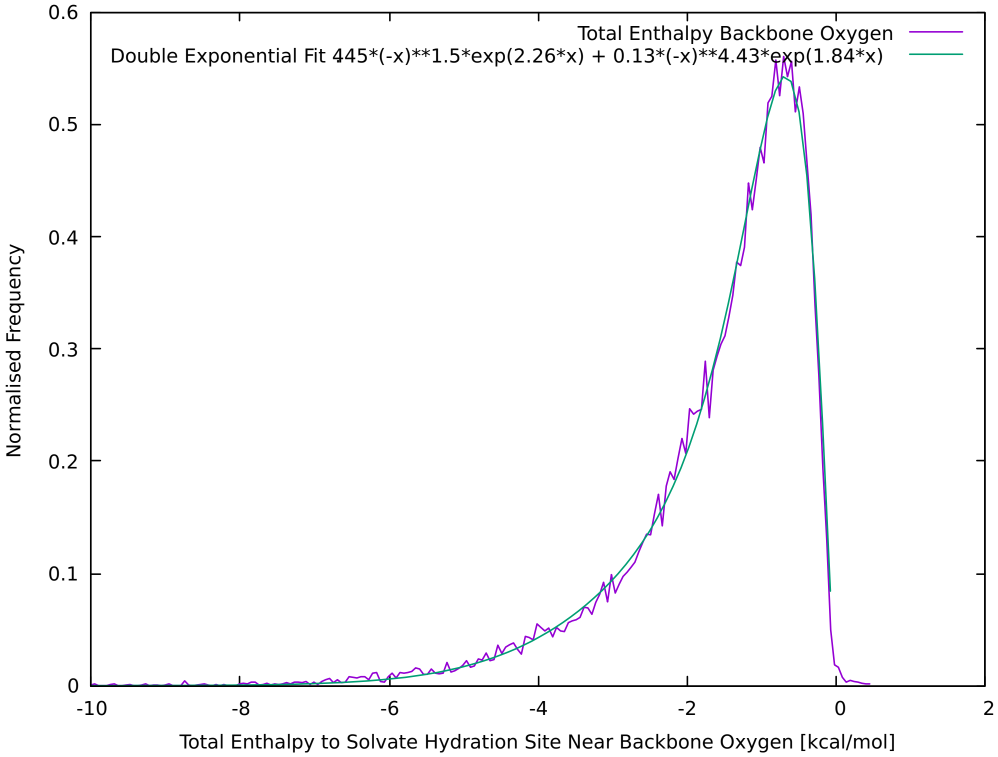


Fig. 6.11 This plot shows the distribution of hydration site enthalpies for sites with backbone oxygen as close atoms (within 4.5 Å). The shape of the distribution is quite smooth and can be fitted with a double exponential distribution.

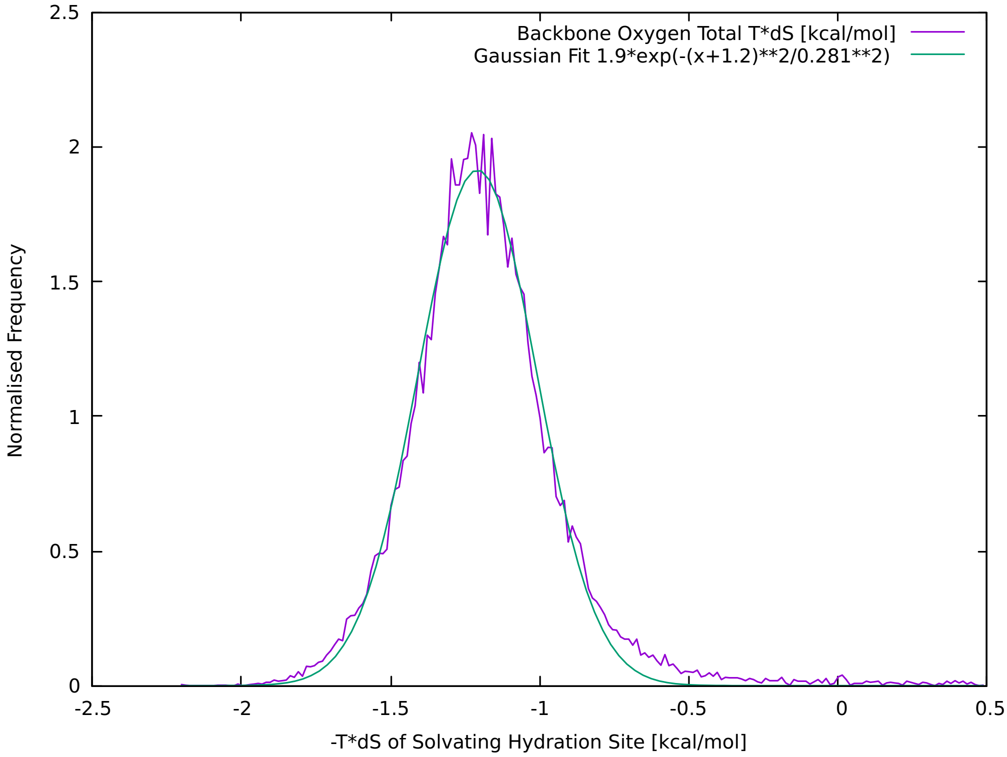


Fig. 6.12 This plot shows the distribution of hydration site entropies for sites with backbone oxygen as close atoms (within 4.5 Å). The shape fits a Gaussian reasonably well, albeit with a longer than expected right tail.

6.6 Part I: Summary

The statistics gathered from this decomposition of hydration free energy have shown numerous insights into the behaviour of water at the surface of a protein. In general, hydrogen bonding atoms create stability for local water molecules but having too many heavy atoms nearby causes either extreme stability or instability. This analysis provides a general picture for the kind of environments around proteins but there will be bias because of the selection of proteins used. Bromodomains make up a large proportion of the dataset as shown in the tables in appendix A. Some of the results may be peculiar to bromodomains, however it is likely that these distributions could be used as a useful prior for the distributions around similar groups in unknown proteins.

It may be possible to combine the insights gathered from the decomposition of free energy to adapt the inputs to the algorithm which finds the most displaceable clusters as applied in chapter 5. Changes made in this way would not be systematic, and would require some degree of tinkering to create an improved model. While this could be a valuable route to pursue in future work, it is not the route that was taken in this thesis. Instead a more systematic method was tried that is described in the next section.

6.7 Part II: Towards a Metric for Generalised Ligandability

A method to predict the ligandability of proteins with confidence through molecular dynamics simulations would be a useful tool. Decompositions of free energy are likely to be a useful tool in this effort because of the local understanding required in finding hydration patches, often around a binding site. The decomposition of the free energy is likely to express the ligandability because it represents how likely nature is to undertake the change. In chapter 5 a hydration sites method combined with a combinatoric search found the most hydrophobic sites and allowed a relative scoring between proteins. Any improvement on the method suggested by Vukovič and Huggins [153] would then be useful. That method specifically calculates the hydration free energy of the hydration sites, and there is some correlation with experimental ligandability scores. By adding a ‘charged site compensation’ as described earlier in this chapter there is potential for improvement. The adaptation attempts to insert probe atoms with various MD force field parameters into the positions of the hydration sites. By assessing the in situ energetic contribution from the probes and the bulk hydration free energy of the probe a new score is calculated for each HS. This investigation uses the same 13 targets from a dataset by Cheng [156] which were used in the study by Vukovič and Huggins [153]. These are: HIV-1p, PDE5A, ACE, factor Xa, CTSK, NEU, CDK2, ICE, PDE4D, AChE, thrombin, c-Abl and PTP1B.

The combinatoric search algorithm can then be used on this data and instead of finding the easiest patch of hydration sites to displace it will find the patch of hydration sites that has the most potential for ligandability. The metric used to score the algorithm is devised from how well results correlate with ligandability scores derived from database information which were performed by Vukovič and Huggins [153]. This study used clusters of 18 hydration sites.

6.7.1 Rescoring the Hydration Sites

The goal is to find parameters associated with hydration sites which give a useful score for $\Delta\mathcal{G}$ that represents ligandability. The original algorithm only considered $\Delta\mathcal{G} = \Delta G_{\text{hyd}}$ for the HS. If this is very negative the site is hard to displace, if it is positive the site is beneficial to displace. The first adaptation to consider is replacing the HS with a probe atom, the probe atom will have an energy associated with it being in the location of the original HS, ΔE_{prob}. This can be factored into the metric as $\Delta\mathcal{G} = \Delta G_{\text{hyd}} - \Delta E_{\text{prob}}$. Using this new value, if there is a beneficial (negative) change in energy this will adjust the hydration free energy to be more positive, and therefore more beneficial to displace.

One can consider that any probe atom representing a ligand will have come from the solution. The ligand will generally be soluble, and therefore taking the molecule from solution will incur a penalty as hydrogen bonds with the bulk water will have to be broken. This was factored in for each of the prospective probe atoms by performing an FEP calculation of the hydration free energy, $\Delta G_{\text{solv,prob}}$, of the probe in a box of water. These data and the corresponding probe parameters can be found tabulated in appendix A. The new metric that also takes into account this desolvation penalty is then:

$$\Delta\mathcal{G} = \Delta G_{\text{hyd}} - \Delta E_{\text{prob}} + \Delta G_{\text{solv,prob}}. \tag{6.3}$$

A new sweep was made on all replicates of the 13 protein types in the Cheng dataset. For each HS all of the probes were tried. In the ligandability study by Vukovič and Huggins [153] the average sum of $\Delta\mathcal{G}$ was taken across replicates for the best cluster of 18 sites. The 13 protein types were scored on a relative, unit-less scale between 0.1 and 0.77 and distributed linearly according to the relative cluster displacement scores. This was the final scoring method used to compare the original algorithm to the experimental measure for ligandability [153].

6.7.2 Technical Requirements

If a protein in the dataset has an overall net charge summed over all atoms this would normally be somewhat neutralised by ions in solution during an MD simulation. In the case of single .pdb file structures of the protein used as inputs, there is no time averaging and no solvent. As a result, upon calculation of ΔE_{prob} the effect of any ions would be spurious as they would be based on the locations they happened to be in in the .pdb snapshot. To remove this problem, the ions were removed and to keep the charge effects as non-local as possible, so as not to influence any sites in particular, all protein charges were scaled proportionally such that the net charge of the protein was close to 0 (approx 10^{-14}). This very small scaling helps to lower the Coulombic energy bias in favour of negatively charged probes (for positively charged proteins).

One could also consider the dielectric effect of the water which was not present in the final .pdb file when ΔE_{prob} was calculated. To compensate this a dielectric with $\eta = 80$ was applied. However, this adds a conceptual complication. One could justify using such a constant in a bulk like environment to mimic the screening effect of the solvent, but close to the protein this would be a variable quantity because the surrounding protein will heavily screen interactions with the solvent. To remove the complexity of varying η it was approximated as 80 for all interactions.

6.7.3 Choice of probes

A range of probe atoms that represent ligand atoms will be required to replace the HS. The main parameters for probes will be the size and strength of interaction for the non-bonded parameters. This is essentially three parameters: the VDW/repulsion radius, r_{min} and well depth ε and the probe charge q. It is not clear which probes would be best to select from the space of all possible probes. It was decided that probes should be chosen to *mimic already existing atoms* in parametrised ligands. A source of these parameters is the CHARMM general force field (or CGENFF) parameter set [12]. Arbitrarily generated probes may not reflect the kinds of atoms that would be present in ligands, for example a

Name	Atom Type	ε	$\frac{\text{rmin}}{2}$	q
IGR1	I	-0.55	2.19	-0.08
SG3O2	S	-0.35	2.0	0.65
NG311	N	-0.045	2.0	-0.69
CG2R53	C	-0.02	2.2	0.67
CG323	C	-0.11	2.2	-0.47
CG2O1	C	-0.11	2.0	0.68
NG301	N	-0.035	2.0	-0.63
NG2S2	N	-0.2	1.85	-0.69

Table 6.3 The top section are very common probes, the lower section are probes that were picked on occasion but not very often.

small strong negatively charged probe might be seen to represent fluorine, but there is not a similar atom with equivalent size but a positive charge, only proton like hydrogen atoms.

This force field has the parameters stored for all atom types in a .rtf file. All of the possible atoms in the CGENFF parameter file (.rtf) were taken with all variations of charges seen in the .rtf files for existing default molecules. This gives an initial set of 826 probes after duplications of parameters with the same name are removed. These probes were used to run through some test proteins using the rescoring metric given in equation 6.3. Common probes to appear were: SG302, with charge -0.8, NG331 with charge -1.125 and OG2D2 with charge -1.14. These all have unusually high charges and may be seen as somewhat extreme examples of probe atoms that are not representative of the average ligand atoms. It was then decided that the maximum charge for any probe should be within ±0.7 electon charges. Removing probes with charges higher than this, there were still 745 unique probes remaining. It was then decided that hydrogen is not a ligand like atom. After hydrogen like probes were removed, there were 615 probe types remaining. The hydration free energy $\Delta G_{\text{solv,prob}}$ was measured for each using an FEP simulation. These were the final probes and are shown along with thier parameters in tables in appendix A.

6.7.4 Commonly Occuring Probes

Some test examples of commonly occurring probes are shown in table 6.3.

The common probes tend to be extreme cases with large charges, radii or well depths, or minimal charges and shallow wells:

- IGR1 -0.08, The largest atom in the set of probes with barely any charge, deep well, large radius.
- SG3O2 0.65, One of the larger positive charges in the set, large radius, relatively deep well.
- NG311 -0.69, One of the larger negative charges, large radius, shallow well
- CG2R53 0.67, One of the larger positive charges, large radius, shallow well
- CG323 -0.47, moderate negative charge, larger radius, deeper well
- CG2O1 0.68, strong positive charge, larger radius, deeper well
- NG301 -0.63, strong negative charge, shallow well, large radius
- NG2S2 -0.69, strong negative charge, moderate well depth and radius

This selection of probes already allows a rough classification of sites. For sites with mixed charges, it is likely that placing a single charge of any given type will not be desirable. These sites will accept the iodine IGR1 which does a good job of filling the volume without incurring a charge penalty. It is worth noting that some may not consider iodine be a drug like atom and may argue it should not be present in the set of probes. The other probes generally classify a site as favouring a positive or negative charge. Interestingly there were not many sites that favour a small radius. This may indicate that the repulsion/VDW energy terms in the energy calculation were outweighed by Coulombic energy terms. Whether this is realistic or not is an important consideration in the design of the $\Delta\mathcal{G}$ metric.

6.8 Part II: Results

The rescored relative ligandability of the 13 proteins is shown in table 6.4. There is some agreement with experimentally predicted ligandabilities. There are also noticable differences from the original hydration only predictions by Vukovič and Huggins [153].

Table 6.4 helps highlight the proteins which required improvement over the original work by Vukovič and Huggins [153]. These proteins most likely had essential parts which were polar and were insensitive to the hydration only binding prediction algorithm. For

Target Name	VH Displacement Score	VH Predicted Ligandability	Experimental Ligandability	Rescored Ligandability
HIV-1p	-17.8	0.76	0.77	0.59
PDE5A	-17.7	0.77	0.75	0.62
ACE	-27.4	0.16	0.59	0.60
Factor Xa	-23.5	0.40	0.55	0.43
CTSK	-21.2	0.55	0.53	0.53
NEU	-25.8	0.26	0.52	0.68
CDK2	-28.3	0.10	0.44	0.56
ICE	-21.1	0.55	0.44	0.1
PDE4D	-19.7	0.65	0.42	0.62
AChE	-21.5	0.53	0.37	0.77
Thrombin	-22.2	0.48	0.37	0.54
c-Abl	-21.7	0.51	0.33	0.64
PTP1B	-26.0	0.24	0.1	0.45
MUD	-	2.11	0.00	2.24
SSQ	-	0.53	0.00	0.55

Table 6.4 Table of Cheng's targets [156] with the scores from the hydration only analysis of Vukovič and Huggins [153] and the rescored values from this work. MUD is the mean unsigned difference between the column and the experimental figures. SSQ is the sum of square differences.

example, ACE has a large difference in predicted and experimental ligandability from this original assay. After rescoring this has improved greatly with the new algorithm agreeing almost exactly. This agreement may have been by chance. Many of the other proteins ligandability predictions have only improved marginally or are worse (for example Thrombin).

Overall the new rescored results have a mean unsigned difference of 2.24 (compared to the original 2.11) and a sum of square difference of 0.55 (compared to the original 0.53). This is slightly worse than the original scaling using the hydration scores only. This is still encouraging because it means there may be real improvements to be made if the energy terms are included in a more sensible manner. The rescoring may have 'overshot' when correcting, and may be focusing too greatly on the probe energies.

A degree of overall validation is required. To rule out the scoring methods giving fortuitous correlations with experiment 300 tests were taken where the best sites energy sums were randomly ordered. The distribution of experimental mean unsigned differences and sum of square differences were taken from this random data to compare with the actual results. The comparison is shown in figure 6.13. This figure demonstrates that the rescoring

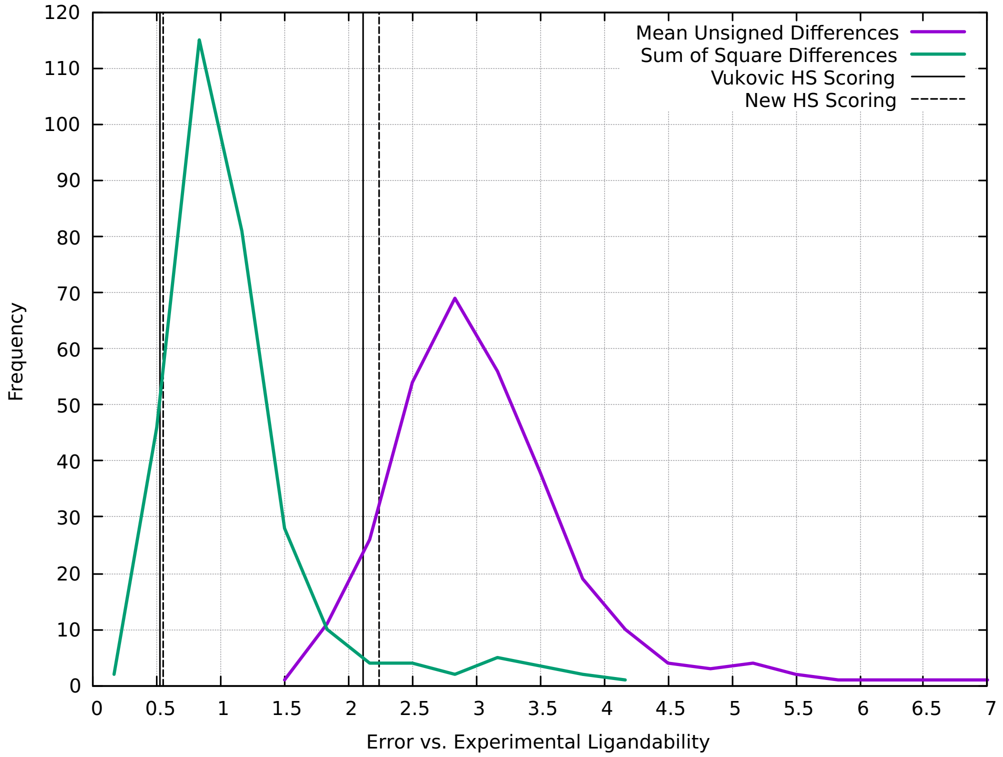


Fig. 6.13 The distributions of mean unsigned difference and sum of square differences to the experimental ligandability scores for randomly ordered data. The scores obtained by the two rescoring methods appear in the left tails of each distribution showing they are more predictive than expected for random assignments.

from the two methods is better than would be given from random assignments of the ligandability of each target from 0.1 to 0.77. The two methods scores are both in the left tail of each distribution and show they are more predictive than the random data by a large margin. There is still a chance that both data points are flukes, this can be ruled out with further studies, potentially with better $\Delta\mathcal{G}$ metrics. There are a number of ways to potentially improve this method and will be discussed in detail in the next section.

6.9 Part II: Analysis

In this section we discuss potential problems with the current implementation of this method.

1. In equation 6.3, the full interaction energy ΔE_{probe} is used, but this has the potential to be quite large. There would normally be an entropic offset $-T\Delta S_{\text{probe}}$ that may reduce this. It was considered whether the entropy of the HS itself could be used as a proxy for the probe entropy. However, this is unlikely to be accurate, as the entropy of the ligand will be a fixed penalty depending on the size and shape of the ligand and the bonding/stiffness of the molecule. To use the water/HS entropy term would be wrong. There was no overall reduction to the energetic terms meaning they may have been considered too strongly in the weighting search, this may justify including a parameter to scale down the ΔE_{prob} term.

2. The desolvation penalty was included but it assumes that the probe is fully desolvated in each case. Different HS have different levels of embedding into the protein as is evident from figures such as 6.8. It is possible to calculate or approximate the degree of solvation around hydration sites, for example by looking in a 10 Å range around each hydration site and using a quick Monte-Carlo estimation. This idea is described in further detail in section 6.9.2 below. This may be considered fine tuning for most probes, however the $\Delta G_{\text{solv,prob}}$ terms can grow to a few tens of kcal/mol.

3. There was no long range correction made to the probe desolvations during the FEP calculation. The values shown in the tables in appendix A are solvation free energies of probes from pure MD FEP. There should be a correction made to consider the long range charge interactions across the (falsely periodic) box. There are standard ways of doing this, but they were not implemented because the effect is likely be small. A more advanced study following this work may wish to take this into account.

4. It seems that highly charged atoms were dominating the most favoured list of probes in table 6.3. Probes with large charges will have larger energy contributions as the Coulombic interaction terms are generally stronger than the VDW terms. This may be because dipoles are needed; most ligand molecules have alternating charges that will soften the interactions to some degree. Again to bypass this a weighting factor could be included on ΔE_{probe} but also specifically on the Coulombic energy term. This is discussed in the following section but it may also make sites more sensitive to VDW/repulsion energies.

5. To find the parameters one could perform a sweep over metrics $\Delta\mathcal{G}$. This kind of parameter sweep will be a somewhat expensive and time consuming calculation for future work, but will probably find a good metric to predict ligandability.

6. It would appear that detailed further investigations are required to develop a better metric for ligandability. Some insights could also be drawn from the distributions displayed in the first part of this chapter.

6.9.1 Future Work: Compensation Strategies

Whenever a HS is displaced we have a number of factors to take into account. The solvation free energy of the hydration site ($\Delta G_{HS,IFST}$) is usually negative as it is favourable to have water there. If the HS is replaced with a probe there is a contribution ΔG_{probe}. This can be split into an enthalpic and entropic part $\Delta G_{probe} = \Delta E_{probe} - T\Delta S_{probe}$. Estimating ΔS_{probe} is hard, but calculating ΔE_{probe} from an MD force field is relatively easy. One could approximate that $\Delta G \approx \Delta E$ as used in the rescoring scheme of this chapter. It may be better to assign a constant such that $\Delta G \approx A\Delta E$ where A probably lies between 0 and 1. This constant accounts for the entropy decreasing the energetic term. The probe energy will come from two sources, the Coulombic interactions with the protein and the VDW interactions. We can write $\Delta E_{\text{probe}} = \Delta E_{\text{VDW}} + \Delta E_{\text{Coulomb}}$. If we are to assign the constant A as above, we may wish to treat these two terms separately. The VDW term is more sensitive to the local interactions. We could write $\Delta G_{\text{probe}} \approx A\Delta E_{\text{probe,VDW}} + B\Delta E_{\text{probe,Coulomb}}$. Finally, there is the probe desolvation penalty which arose from bringing the probe atoms from solution to the protein environment. This free energy $\Delta G_{\text{probe,solv}}$ was calculated for all probes using FEP simulations and the results are in appendix A. The problem with this approach is that if we simply subtract the full penalty, this assumes that the probe was fully desolvated upon reaching the protein. For HS on the surface of the protein we would expect at least some volume to be solvated. Only for fully embedded HS would the full desolvation cost need to be subtracted. In this case a parameter could be introduced which mediates the partial desolvation. The parameter κ is such that if κ is 1 then there is no protein nearby and the hydration site is in bulk water, if κ is 0 then the HS is fully embedded in the protein and a full desolvation penalty is paid. In addition to this if another parameter C was introduced to control the importance of the overall desolvation penalty the metric for rescoring sites would have the following form:

$$\Delta G_{total} = \Delta G_{HS,IFST} - A\Delta E_{probe,vdw} - B\Delta E_{probe,coulomb} - C(1-\kappa)\Delta G_{probe,solv} \tag{6.4}$$

In this form a parameter sweep can be made which finds the best A, B and C to define a metric that fits the 13 test cases from the Cheng dataset.

6.9.2 Calculation of κ

the bulk-like parameter κ could be calculated by a rough Monte-Carlo sampling around each site. An example calculation was made where $N_t = 10000$ points were randomly generated within 10 Å around each hydration site. The number of points that were within 1 Å of a protein atom N_p were counted then $\kappa = 1 - \frac{N_p}{N_t}$ was used to calculate the bulk-like index. This lead to an average value of around 0.6, and the measure seemed to correlate well with protein embedding when sites were checked by eye. The parameters 10 Å and 1 Å could be modified, if necessary.

6.9.3 Averaging over Probes

Talking the best probe for each hydration site may not be reflective of the drug design process. The best probe may not be a feasible atom for most sites. Instead the average of the best few probes could be taken. This average could possibly be weighted.

6.10 Part II: Summary

In summary, the hydration free energy decomposition has offered a new avenue to understand ligandability. Although the metric proposed here did not succeed in improving the existing search algorithm, there is reasonable evidence that a better metric could be developed. Such a metric will require a delicate balance of many factors which will take some time to determine. If this were achieved then it would be possible to estimate purely from simulation how much effort should be spent of developing new drugs for a specific target. The most ligandable site can be found in the computer and compared to the current best ligandability result in the literature for that target. This would act as a management tool for drug design projects and could highlight wasted effort for targets which have already achieved their maximum likely ligandability.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The decomposition of free energy was featured in all four of the main chapters of this work. In chapter 3 the first type of decomposition was splitting a solvation or hydration free energy into an enthalpic and entropic part. This allowed the calculation of the free energy in terms of an entropy integral series, which was truncated to facilitate calculation. The k-nearest neighbours method used to calculate the entropy integrals also allowed a further decomposition. Each small unit of volume, or voxel, in a simulation cell could have the local sum of contributions to the entropy from the atoms in that unit of volume. This allowed the hydration free energy and entropy to be displayed as a field around the solute. This also allowed the entropy integrals to be calculated in parallel on a supercomputer.

Chapter 4 dealt with a different type of decomposition. This took a free energy change associated with hydration or binding of a ligand molecule. That ligand was composed of atoms, and the free energy change was shared across the atoms in the molecule based on the interactions those individual atoms took part in throughout the MD simulation. This allowed a visualisation, showing the strongest and weakest contributors to the hydration/binding free energy. The weights given to each atom might be useful in future studies to see which parts of a molecule can be changed to allow for stronger or weaker interactions. This information would not have been present without the adaptation to the Zwanzig equation called atomwise free energy perturbation, or AFEP, which was derived in this work. This measure is specifically useful to combat the complexity of large simulations of complicated proteins. This will highlight the parts of the systems that are important with respect to the free energy change being measured. There was some evidence that splitting the contributions down to the atom level was too fine grained as neighbouring pairs of atoms often had opposing signs for their free energy contributions, which when averaged cancelled each other out. For larger systems it would be interesting to investigate a partitioning by amino acid, or by chemical group, which may provide a different kind of information. This larger grouping can be performed with the current state of the algorithm.

Chapter 5 considered the binding of two proteins. The free energy change in question was again the hydration free energy, but this time for water molecules on the surface of the proteins. The individual molecules hydration scores at points of interest on the protein surface allowed a different kind of decomposition. Hydration sites on the protein surface along with their scores helped classify the local chemistry at that point on the protein. This highlighted hydrophobic regions on the surface of the $\gamma 2$ subunit of the intracellular domain of the GABA$_A$ receptor. Out of these patches, the most displaceable patch of hydration sites was the one that agreed best with experimental evidence of important amino acids. It was also the patch that agreed well with SwarmDock docking positions of GABARAP to a section of the GABA$_A$ receptor. The hydration environment on the surface of the complementary GABARAP protein was complicated and exhibited many sites of interest which were catalogued using the free energy decomposition. Most of these sites appeared to be implicated in numerous other interactions that GABARAP takes part in with other proteins or between multiple GABARAPs in dimerisation or trimerisation interactions. The sites which overlapped in the binding created a mutual exclusion zone of hydrophobicity, which will have been one of the positive factors toward the binding of these two proteins. It is likely that the method used for this study could be used in other studies of protein binding. This work is the first example of using this kind of hydration free energy decomposition to study protein-protein binding.

The final chapter, chapter 6 looked further into this final type of free energy decomposition, taking the data from hydration sites from the surfaces of many different types of protein. This data showed that the interactions of water and proteins could be predicted to some extent by analysing the local chemical environment that the water molecule is in. Certain sites were very displacable despite being close to atoms which might be thought to encourage hydrogen bonding and the stability associated with it. Curves were fitted to distributions which had sufficient data. For those distributions which did not have sufficient data mathematical models could be built if more data were collected.

Later in chapter 6 an attempt was made to enhance the algorithm that finds highly displaceable hydration patches on the surface of proteins. The interaction energy at the local hydration site and the desolvation of a probe atom was used to rescale the free energy scores of hydration sites in key proteins involved in a ligandability study. After rescoring, the correlation of the dataset to ligandability was only slightly worse than considering hydration alone. This is encouraging because there may have been an overcompensation of site energy when including the new terms. Also, important factors were not taken into account, for example the degree of protein embedding at the hydration site. In the current state of the method much work is still needed, but if these adaptations were developed there could be improvements to ligandability prediction from MD simulations.

In summary, free energy decompositions are useful tools that generate a suite of tools to help understand complex systems and environments. Free energy is a fundamental quantity from which all other thermodynamic quantities can be derived. Decompositions of the free energy offer a deep insight into the local behaviour of systems. These tools could be used in both academic and industrial applications in the future, such as understanding protein-protein interactions and designing new ligands for binding sites.

References

[1] Karplus Martin and McCammon J. Andrew, "Molecular dynamics simulations of biomolecules," *Nature Structural Biology*, vol. 9, p. 646, sep 2002.

[2] G. J. Martyna, D. J. Tobias, and M. L. Klein, "Constant pressure molecular dynamics algorithms," *The Journal of Chemical Physics*, vol. 101, no. 5, pp. 4177–4189, 1994.

[3] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks, "Constant pressure molecular dynamics simulation: The langevin piston method," *The Journal of Chemical Physics*, vol. 103, no. 11, pp. 4613–4621, 1995.

[4] S. Plimpton, "Fast parallel algorithms for short-range molecular dynamics," *Journal of Computational Physics*, vol. 117, no. 1, pp. 1 – 19, 1995.

[5] Brooks BR, Brooks CL, MacKerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable R M, Woodcock H L, Wu X, Yang W, York DM, and Karplus M, "CHARMM: The Biomolecular Simulation Program," *Journal of computational chemistry*, vol. 30, p. 1545–1614, jul 2009.

[6] Abraham Mark James, Murtola Teemu, Schulz Roland, Páll Szilárd, Smith Jeremy C., Hess Berk, and Lindahl Erik, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, p. 19–25, 2015.

[7] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé, and K. Schulten, "Scalable molecular dynamics with NAMD," *J. Comput. Chem.*, vol. 26, no. 16, pp. 1781–1802, 2005.

[8] Salomon-Ferrer Romelia, Case David A., and Walker Ross C., "An overview of the Amber biomolecular simulation package," *WIREs Comput Mol Sci*, vol. 3, no. 2, p. 198–210, 2012. doi: 10.1002/wcms.1121.

[9] Eastman Peter, Swails Jason, Chodera John D., McGibbon Robert T., Zhao Yutong, Beauchamp Kyle A., Wang Lee-Ping, Simmonett Andrew C., Harrigan Matthew P., Stern Chaya D., Wiewiora Rafal P., Brooks Bernard R., and Pande Vijay S., "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," *PLOS Computational Biology*, vol. 13, p. e1005659, jul 2017.

[10] H. H. Loeffler, S. Bosisio, G. Duarte Ramos Matos, D. Suh, B. Roux, D. L. Mobley, and J. Michel, “Reproducibility of free energy calculations across different molecular simulation software,” Jun 2018.

[11] MacKerell Alexander D., Banavali Nilesh, and Foloppe Nicolas, “Development and current status of the CHARMM force field for nucleic acids,” *Biopolymers*, vol. 56, no. 4, p. 257–265, 2001. doi: 10.1002/1097-0282(2000)56:4<257::AID-BIP10029>3.0.CO;2-W.

[12] Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, and MacKerell A D, “CHARMM General Force Field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields,” *Journal of computational chemistry*, vol. 31, p. 671–690, mar 2010.

[13] T. P. Straatsma and J. A. McCammon, “Multiconfiguration thermodynamic integration,” *J. Chem. Phys.*, vol. 95, pp. 1175–1188, jul 1991.

[14] G. Gerogiokas, G. Calabro, R. H. Henchman, M. W. Y. Southey, R. J. Law, and J. Michel, “Prediction of small molecule hydration thermodynamics with grid cell theory,” *Journal of Chemical Theory and Computation*, vol. 10, no. 1, pp. 35–48, 2014. PMID: 26579889.

[15] R. W. Zwanzig, “High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases,” *J Chem Phys*, vol. 22, no. 8, pp. 1420–1426, 1954.

[16] C. H. Bennett, “Efficient estimation of free energy differences from Monte Carlo data,” *J. Comput. Phys.*, vol. 22, no. 2, pp. 245–268, 1976.

[17] P. Liu, F. Dehez, W. Cai, and C. Chipot, “A toolkit for the analysis of free-energy perturbation calculations,” *J. Chem. Theory Comput.*, vol. 8, no. 8, pp. 2606–2616, 2012.

[18] E. Friedgut, “Hypergraphs, entropy, and inequalities,” *Am. Math. Mon.*, vol. 111, no. 9, pp. 749–760, 2004.

[19] L. Wang, R. Abel, R. a. Friesner, and B. J. Berne, “NIH Public Access,” *October*, vol. 5, no. 6, pp. 1462–1473, 2009.

[20] T. Lazaridis, “Inhomogeneous Fluid Approach to Solvation Thermodynamics. 1. Theory,” *J. Phys. Chem. B*, vol. 102, no. 18, pp. 3531–3541, 1998.

[21] Lazaridis Themis, “Inhomogeneous Fluid Approach to Solvation Thermodynamics. 2. Applications to Simple Fluids,” *The Journal of Physical Chemistry B*, vol. 102, no. 18, p. 3542–3550, 1998. doi: 10.1021/jp972358w.

[22] T. Lazaridis and M. Karplus, “Orientational correlations and entropy in liquid water,” *The Journal of Chemical Physics*, vol. 105, no. 10, pp. 4294–4316, 1996.

[23] Fisher I Z and Kopeliovich B L *Soviet Phys. Dokl.*, 1961.

[24] Goethe Martin, Fita Ignacio, and Rubi J. Miguel, “Testing the mutual information expansion of entropy with multivariate Gaussian distributions,” *The Journal of Chemical Physics*, vol. 147, no. 22, p. 224102, 2017. doi: 10.1063/1.4996847.

[25] J. G. Kirkwood, “Statistical Mechanics of Fluid Mixtures,” *J. Chem. Phys.*, vol. 3, pp. 300–313, may 1935.

[26] J. G. Kirkwood and F. P. Buff, “The Statistical Mechanical Theory of Solutions. I,” *J. Chem. Phys.*, vol. 19, no. 6, pp. 774–777, 1951.

[27] R. E. Nettleton and M. S. Green, “Expression in Terms of Molecular Distribution Functions for the Entropy Density in an Infinite System,” *J. Chem. Phys.*, vol. 29, no. 6, p. 1365, 1958.

[28] H. J. Raveche, “Entropy and Molecular Correlation Functions in Open Systems. I. Derivation,” *J. Chem. Phys.*, vol. 55, no. 5, p. 2242, 1971.

[29] D. C. Wallace, “On the role of density fluctuations in the entropy of a fluid,” *J. Chem. Phys.*, vol. 87, no. 4, p. 2282, 1987.

[30] D. C. Wallace, “Statistical entropy and a qualitative gas-liquid phase diagram,” *Phys. Rev. A*, vol. 38, no. 1, pp. 469–472, 1988.

[31] D. C. Wallace, “Statistical theory for the entropy of a liquid,” *Phys. Rev. A*, vol. 39, no. 9, pp. 4843–4847, 1989.

[32] T. Morita and K. Hiroike, “A new approach to the theory of classical fluids. iiigeneral treatment of classical systems,” *Progress of Theoretical Physics*, vol. 25, no. 4, pp. 537–578, 1961.

[33] T. Lazaridis, “Solvent reorganization energy and entropy in hydrophobic hydration,” *J. Phys. Chem. B*, vol. 104, no. 20, pp. 4964–4979, 2000.

[34] D. J. Huggins and M. C. Payne, “Assessing the accuracy of inhomogeneous fluid solvation theory in predicting hydration free energies of simple solutes,” *J. Phys. Chem. B*, vol. 117, no. 27, pp. 8232–8244, 2013.

[35] C. N. Nguyen, T. K. Young, and M. K. Gilson, “Erratum: “Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril” [J. Chem. Phys. 137, 044101 (2012)],” *J. Chem. Phys.*, vol. 137, p. 149901, oct 2012.

[36] Lin Shu-Kun, “Correlation of Entropy with Similarity and Symmetry,” *Journal of Chemical Information and Computer Sciences*, vol. 36, no. 3, p. 367–376, 1996. doi: 10.1021/ci950077k.

[37] S. J. Irudayam and R. H. Henchman, “Entropic cost of protein-ligand binding and its dependence on the entropy in solution,” *J. Phys. Chem. B*, vol. 113, no. 17, pp. 5871–5884, 2009.

[38] R. H. Henchman, “Free energy of liquid water from a computer simulation via cell theory,” *The Journal of Chemical Physics*, vol. 126, no. 6, p. 064504, 2007.

[39] S. J. Irudayam and R. H. Henchman, “Solvation theory to provide a molecular interpretation of the hydrophobic entropy loss of noble-gas hydration,” *Journal of Physics: Condensed Matter*, vol. 22, p. 284108, jun 2010.

[40] S. J. Irudayam and R. H. Henchman, “Prediction and interpretation of the hydration entropies of monovalent cations and anions,” *Molecular Physics*, vol. 109, no. 1, pp. 37–48, 2011.

[41] B. W. J. Irwin and D. J. Huggins, “On the accuracy of one-and two-particle solvation entropies,” *J. Chem. Phys.*, 2017.

[42] C. N. Nguyen, A. Cruz, M. K. Gilson, and T. Kurtzman, “Thermodynamics of water in an enzyme active site: Grid-based hydration analysis of coagulation factor xa,” *J. Chem. Theory Comput.*, vol. 10, no. 7, pp. 2769–2780, 2014.

[43] M. K. Gilson and H.-X. Zhou, “Calculation of protein-ligand binding affinities.,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 36, pp. 21–42, 2007.

[44] G. J. Rocklin, D. L. Mobley, and K. A. Dill, “Separated topologies-A method for relative binding free energy calculations using orientational restraints,” *J. Chem. Phys.*, vol. 138, p. 085104, feb 2013.

[45] J. Kästner, H. M. Senn, S. Thiel, N. Otte, and W. Thiel, “QM/MM Free-Energy Perturbation Compared to Thermodynamic Integration and Umbrella Sampling: Application to an Enzymatic Reaction,” *J. Chem. Theory Comput.*, vol. 2, pp. 452–461, mar 2006.

[46] N. Matubayasi, L. H. Reed, and R. M. Levy, “Thermodynamics Of The Hydration Shell .1. Excess Energy Of A Hydrophobic Solute,” *J. Phys. Chem.*, vol. 98, no. 41, pp. 10640–10649, 1994.

[47] N. Matubayasi and M. Nakahara, “Theory of solutions in the energetic representation. I. Formulation,” *J. Chem. Phys.*, vol. 113, no. 15, pp. 6070–6081, 2000.

[48] D. J. Huggins, “Estimating Translational and Orientational Entropies Using the k -Nearest Neighbors Algorithm,” *J. Chem. Theory Comput.*, vol. 10, pp. 3617–3625, sep 2014.

[49] D. Huggins, “Quantifying the Entropy of Binding for Water Molecules in Protein Cavities by Computing Correlations,” *Biophys. J.*, vol. 108, no. 4, pp. 928–936, 2015.

[50] C. Velez-Vega, D. J. J. McKay, T. Kurtzman, V. Aravamuthan, R. A. Pearlstein, and J. S. Duca, “Estimation of solvation entropy and enthalpy via analysis of water oxygen-hydrogen correlations,” *J. Chem. Theory Comput.*, vol. 11, no. 11, pp. 5090–5102, 2015.

[51] F. Reinhard and H. Grubmüller, “Estimation of absolute solvent and solvation shell entropies via permutation reduction,” *J. Chem. Phys.*, vol. 126, p. 014102, jan 2007.

[52] B. M. Dickson, H. Huang, and C. B. Post, “Unrestrained computation of free energy along a path,” *J. Phys. Chem. B*, vol. 116, no. 36, pp. 11046–11055, 2012.

[53] L. Young and C. B. Post, “Free Energy Calculations Involving Internal Coordinate Constraints to Determine Puckering of a Six-Membered Ring Molecule,” *J. Am. Chem. Soc.*, vol. 115, no. 13, pp. 1964–1970, 1993.

[54] V. Hnizdo, T. Jun, B. J. Killian, and M. K. Gilson, “Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods,” *J. Comput. Chem.*, vol. 29, no. 10, pp. 1605–1614, 2008.

[55] L. F. Kozachenko and N. N. Leonenko, “Statistics of a random vector (in Russian),” *Probl. Commun.*, vol. 23, no. 2, pp. 9–16, 1987.

[56] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, “Nearest neighbor estimates of entropy,” 2003.

[57] C. N. Nguyen, T. Kurtzman, and M. K. Gilson, “Spatial Decomposition of Translational Water–Water Correlation Entropy in Binding Pockets,” *J. Chem. Theory Comput.*, vol. 12, pp. 414–429, jan 2016.

[58] C. Nguyen, M. K. Gilson, and T. Young, “Structure and Thermodynamics of Molecular Hydration via Grid Inhomogeneous Solvation Theory,” *arXiv*, p. 16, 2011.

[59] A. Baranyai and D. J. Evans, “Direct entropy calculation from computer simulation of liquids,” *Phys. Rev. A*, vol. 40, no. 7, pp. 3817–3822, 1989.

[60] H. Reiss, “Superposition approximations from a variation principle,” *J. Stat. Phys.*, vol. 6, no. 1, pp. 39–47, 1972.

[61] V. Hnizdo, E. V. A. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh, “Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules,” *J. Comput. Chem.*, vol. 28, no. 3, pp. 655–668, 2007.

[62] U. Hensen, H. Grubmüller, and O. F. Lange, “Adaptive anisotropic kernels for nonparametric estimation of absolute configurational entropies in high-dimensional configuration spaces,” *Phys. Rev. E*, vol. 80, p. 011913, jul 2009.

[63] U. Hensen, O. F. Lange, and H. Grubmüller, “Estimating Absolute Configurational Entropies of Macromolecules: The Minimally Coupled Subspace Approach,” *PLoS One*, vol. 5, p. e9179, feb 2010.

[64] S. E. Feller, A. D. Mackerell Jr, and A. D. J. Mackerell, “An Improved Empirical Potential Energy Function for Molecular Simulations of Phospholipids Supporting 2,” *J. Phys. Chem. B*, vol. 104, no. 31, pp. 7510–7515, 2000.

[65] A. Ahmed and R. J. Sadus, “Solid-liquid equilibria and triple points of n-6 Lennard-Jones fluids,” *J. Chem. Phys.*, vol. 131, p. 174504, nov 2009.

[66] A. Pohorille, C. Jarzynski, and C. Chipot, “Good practices in free-energy calculations,” *J. Phys. Chem. B*, vol. 114, no. 32, pp. 10235–10253, 2010.

[67] W. Humphrey, A. Dalke, and K. Schulten, “VMD: Visual molecular dynamics,” 1996.

[68] J. D. Weeks, D. Chandler, and H. C. Andersen, "Role of Repulsive Forces in Determining the Equilibrium Structure of Simple Liquids," *J. Chem. Phys.*, vol. 54, pp. 5237–5247, jun 1971.

[69] J. Zielkiewicz, "Structural properties of water: Comparison of the SPC, SPCE, TIP4P, and TIP5P models of water," *J. Chem. Phys.*, vol. 123, p. 104501, sep 2005.

[70] Irwin Benedict W. J. and Huggins David J., "Estimating Atomic Contributions to Hydration and Binding Using Free Energy Perturbation," *Journal of Chemical Theory and Computation*, vol. 14, no. 6, p. 3218–3227, 2018. doi: 10.1021/acs.jctc.8b00027.

[71] D. Cappel, M. L. Hall, E. B. Lenselink, T. Beuming, J. Qi, J. Bradner, and W. Sherman, "Relative Binding Free Energy Calculations Applied to Protein Homology Models," *J. Chem. Inf. Model.*, vol. 56, pp. 2388–2400, dec 2016.

[72] R. Abel, L. Wang, E. D. Harder, B. J. Berne, and R. A. Friesner, "Advancing Drug Discovery through Enhanced Free Energy Calculations," *Acc. Chem. Res.*, vol. 50, pp. 1625–1632, jul 2017.

[73] W. L. Jorgensen, "Free energy calculations: a breakthrough for modeling organic chemistry in solution," *Acc. Chem. Res.*, vol. 22, no. 10, pp. 184–189, 1989.

[74] M. R. Shirts, D. L. Mobley, and J. D. Chodera, "Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time?," in *Annu. Rep. Comput. Chem.*, vol. 3, pp. 41–59, 2007.

[75] J. Michel and J. W. Essex, "Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations," *J. Comput. Aided. Mol. Des.*, vol. 24, pp. 639–658, aug 2010.

[76] J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, and V. S. Pande, "Alchemical free energy methods for drug discovery: progress and challenges," *Curr. Opin. Struct. Biol.*, vol. 21, pp. 150–160, apr 2011.

[77] P. V. Klimovich, M. R. Shirts, and D. L. Mobley, "Guidelines for the analysis of free energy calculations," *J. Comput. Aided. Mol. Des.*, vol. 29, pp. 397–411, may 2015.

[78] Z. Cournia, B. Allen, and W. Sherman, "Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations," *J. Chem. Inf. Model.*, vol. 57, pp. 2911–2937, dec 2017.

[79] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, and R. Abel, "Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field," *J. Am. Chem. Soc.*, vol. 137, no. 7, pp. 2695–2703, 2015.

[80] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, pp. 845–854, apr 2013.

[81] W. L. Jorgensen and J. Tirado-Rives, "Molecular modeling of organic and biomolecular systems usingBOSS andMCPRO," *J. Comput. Chem.*, vol. 26, pp. 1689–1700, dec 2005.

[82] D. D.A. Case, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, R. L. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, A. D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, X. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, D. Y. Wu, L. Xiao, and P. Kollman, "Amber17," no. April, 2017.

[83] D. L. Mobley and J. P. Guthrie, "FreeSolv: a database of experimental and calculated hydration free energies, with input files," *J. Comput. Aided. Mol. Des.*, vol. 28, pp. 711–720, jul 2014.

[84] D. Shivakumar, J. Williams, Y. Wu, W. Damm, J. Shelley, and W. Sherman, "Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field," *J. Chem. Theory Comput.*, vol. 6, no. 5, pp. 1509–1519, 2010.

[85] T. Steinbrecher, R. Abel, A. Clark, and R. Friesner, "Free Energy Perturbation Calculations of the Thermodynamics of Protein Side-Chain Mutations," *J. Mol. Biol.*, vol. 429, no. 7, pp. 923–929, 2017.

[86] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus, "Absolute binding free energies: A quantitative approach for their calculation," *J. Phys. Chem. B*, vol. 107, no. 35, pp. 9535–9551, 2003.

[87] A. E. Mark and W. F. van Gunsteren, "Decomposition of the Free Energy of a System in Terms of Specific Interactions," *J. Mol. Biol.*, vol. 240, pp. 167–176, jul 1994.

[88] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," *J. Chem. Phys.*, vol. 129, p. 124105, sep 2008.

[89] A. M. Hahn and H. Then, "Characteristic of Bennett's acceptance ratio method," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 80, no. 3, pp. 1–10, 2009.

[90] B. Sherborne, V. Shanmugasundaram, A. C. Cheng, C. D. Christ, R. L. DesJarlais, J. S. Duca, R. A. Lewis, D. A. Loughney, E. S. Manas, G. B. McGaughey, C. E. Peishoff, and H. van Vlijmen, "Collaborating to improve the use of free-energy and other quantitative methods in drug discovery," *J. Comput. Aided. Mol. Des.*, vol. 30, pp. 1139–1141, dec 2016.

[91] F. Clavel and F. Mammano, "Role of Gag in HIV Resistance to Protease Inhibitors," *Viruses*, vol. 2, pp. 1411–1426, jul 2010.

[92] S. Walmsley, B. Bernstein, M. King, J. Arribas, G. Beall, P. Ruane, M. Johnson, D. Johnson, R. Lalonde, A. Japour, S. Brun, and E. Sun, “Lopinavir-Ritonavir versus Nelfinavir for the Initial Treatment of HIV Infection,” *N. Engl. J. Med.*, vol. 346, pp. 2039–2046, jun 2002.

[93] A. K. Ghosh, H. L. Osswald, and G. Prato, “Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS,” *J. Med. Chem.*, vol. 59, pp. 5172–5208, jun 2016.

[94] G. Geenens, “Mellin-Meijer-kernel density estimation on ${R}+̂$,” *arXiv e-prints*, p. arXiv:1707.04301, July 2017.

[95] J. Huang and A. D. MacKerell, “CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data,” *J. Comput. Chem.*, vol. 34, pp. 2135–2145, sep 2013.

[96] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, “Comparison of simple potential functions for simulating liquid water,” *J. Chem. Phys.*, vol. 79, pp. 926–935, jul 1983.

[97] M. H. Abraham, G. S. Whiting, R. Fuchs, and E. J. Chambers, “Thermodynamics of solute transfer from water to hexadecane,” *J. Chem. Soc. Perkin Trans. 2*, no. 2, p. 291, 1990.

[98] A. Ben-Naim and Y. Marcus, “Solvation thermodynamics of nonionic solutes,” *J. Chem. Phys.*, vol. 81, no. 4, pp. 2016–2027, 1984.

[99] K. Vanommeslaeghe and A. D. MacKerell, “Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing,” *J. Chem. Inf. Model.*, vol. 52, pp. 3144–3154, dec 2012.

[100] K. Vanommeslaeghe, E. P. Raman, and A. D. MacKerell, “Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges,” *J. Chem. Inf. Model.*, vol. 52, pp. 3155–3168, dec 2012.

[101] J. J. Irwin and B. K. Shoichet, “ZINC - A Free Database of Commercially Available Compounds for Virtual Screening ZINC - A Free Database of Commercially Available Compounds for Virtual Screening,” *J. Chem. Inf. Model*, vol. 45, no. December 2004, pp. 177–182, 2005.

[102] G. S. K. K. Reddy, A. Ali, M. N. L. Nalam, S. G. Anjum, H. Cao, R. S. Nathans, C. A. Schiffer, and T. M. Rana, “Design and Synthesis of HIV-1 Protease Inhibitors Incorporating Oxazolidinones as P2/P2’ Ligands in Pseudosymmetric Dipeptide Isosteres,” *J. Med. Chem.*, vol. 50, pp. 4316–4328, sep 2007.

[103] H. M. Berman, “The Protein Data Bank,” *Nucleic Acids Res.*, vol. 28, pp. 235–242, jan 2000.

[104] A. Ali, G. S. K. K. Reddy, H. Cao, S. G. Anjum, M. N. L. Nalam, C. A. Schiffer, and T. M. Rana, “Discovery of HIV-1 Protease Inhibitors with Picomolar Affinities Incorporating N -Aryl-oxazolidinone-5-carboxamides as Novel P2 Ligands,” *J. Med. Chem.*, vol. 49, pp. 7342–7356, dec 2006.

[105] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Res.*, vol. 35, pp. D198–D201, jan 2007.

[106] J. C. Clemente, R. M. Coman, M. M. Thiaville, L. K. Janka, J. A. Jeung, S. Nukoolkarn, L. Govindasamy, M. Agbandje-McKenna, R. McKenna, W. Leelamanit, M. M. Goodenow, and B. M. Dunn, "Analysis of HIV-1 CRF_01 A/E Protease Inhibitor Resistance: Structural Determinants for Maintaining Sensitivity and Developing Resistance to Atazanavir †," *Biochemistry*, vol. 45, pp. 5468–5477, may 2006.

[107] Irwin Benedict W J, Chau P L, Vuckovic S, and Payne M, "Prediction of GABARAP Interaction with the GABA type A Receptor," *Proteins: Structure, Function, and Bioinformatics*, 2018.

[108] S. Vukovic, P. E. Brennan, and D. J. Huggins, "Exploring the role of water in molecular recognition: predicting protein ligandability using a combinatorial search of surface hydration sites," *Journal of Physics: Condensed Matter*, vol. 28, p. article no. 344007, 2016.

[109] J. Fan, D. Li, H.-S. Chen, J.-G. Huang, J.-F. Xu, W.-W. Zhu, J.-G. Chen, and F. Wang, "Metformin produces anxiolytic-like effects in rats by facilitating gabaa receptor trafficking to membrane," *British Journal of Pharmacology*, vol. 176, no. 2, pp. 297–316.

[110] H. Wang, F. K. Bedford, N. J. Brandon, S. J. Moss, and R. W. Olsen, "GABA_A-receptor-associated protein links GABA_A receptors and the cytoskeleton," *Nature*, vol. 397, pp. 69–72, 1999.

[111] J. Nymann-Andersen, H. Wang, L. Chen, J. T. Kittler, S. J. Moss, and R. W. Olsen, "Subunit specificity and interaction domain between GABAA receptor-associated protein (GABARAP) and GABAA receptors," *J. Neurochem.*, vol. 80, no. 5, pp. 815–823, 2002.

[112] D. Knight, R. Harris, M. S. B. McAlister, J. P. Phelan, S. Geddes, S. J. Moss, P. C. Driscoll, and N. H. Keep, "The X-ray Crystal Structure and Putative Ligand-derived Peptide Binding Properties of γ-Aminobutyric Acid Receptor Type A Receptor-associated Protein," *J. Biol. Chem.*, vol. 277, pp. 5556–5561, feb 2002.

[113] J. E. Coyle, S. Qamar, K. R. Rajashankar, and D. B. Nikolov, "Structure of GABARAP in two conformations: implications for GABA_A receptor localization and tubulin binding," *Neuron*, vol. 33, pp. 63–74, 2002.

[114] O. H. Weiergräber, T. Stangler, Y. Thielmann, J. Mohrlüder, K. Wiesehan, and D. Willbold, "Ligand binding mode of GABA_A-receptor-associated protein," *Journal of Molecular Biology*, vol. 381, pp. 1320–1331, 2008.

[115] L. Chen, H. Wang, S. Vicini, and R. W. Olsen, "The γ-aminobutyric acid type A (GABA_A) receptor-associated protein (GABARAP) promotes GABA_A receptor clustering and modulates the channel kinetics," *Proceedings of the National Academy of Sciences, USA*, vol. 97, pp. 11557–11562, 2000.

[116] A. B. Everitt, V. A. L. Seymour, J. Curmi, D. R. Laver, P. W. Gage, and M. L. Tierney, “Protein interactions involving the $\gamma 2$ large cytoplasmic loop of GABA$_A$ receptors modulate conductance,” *FASEB Journal*, vol. 23, pp. 4361–4369, 2009.

[117] Y. Mokrab, V. N. Bavro, K. Mizuguchi, N. P. Todorov, I. L. Martin, S. M. J. Dunn, S. L. Chau, and P.-L. Chau, “Exploring ligand recognition and ion flow in comparative models of the human GABA type A receptor,” *Journal of Molecular Graphics and Modelling*, vol. 26, pp. 760–774, 2007.

[118] N. Unwin, “Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution,” *Journal of Molecular Biology*, vol. 346, pp. 967–989, 2005.

[119] T. Stangler, L. M. Mayr, and D. Willbold, “Solution structure of human GABA$_A$ receptor-associated protein GABARAP,” *Journal of Biological Chemistry*, vol. 277, pp. 13363–13366, 2002.

[120] M. Torchala, I. H. Moal, R. A. G. Chaleil, J. Fernandez-Recio, and P. A. Bates, “SwarmDock: a server for flexible protein-protein docking,” *Bioinformatics*, vol. 29, pp. 807–809, 2013.

[121] M. Torchala, I. H. Moal, R. A. G. Chaleil, R. Agius, and P. A. Bates, “A Markov-chain model description of binding funnels to enhance the ranking of docked solutions,” *Proteins: Structure, Function and Bioinformatics*, vol. 81, pp. 2143–2149, 2013.

[122] I. H. Moal and P. A. Bates, “SwarmDock and the use of normal modes in protein-protein docking,” *International Journal of Molecular Sciences*, vol. 11, pp. 3623–3648, 2010.

[123] S. Jo, T. Kim, V. G. Iyer, and W. Im, “Software news and updates - CHARMM-GUI: A web-based grraphical user interface for CHARMM,” *Journal of Computational Chemistry*, vol. 29, no. 11, pp. 1859–1865, 2008.

[124] K. Haider and D. J. Huggins, “Combining solvent thermodynamic profiles with functionality maps of the Hsp90 binding site to predict the displacement of water molecules,” *J. Chem. Inf. Model.*, vol. 53, no. 10, pp. 2571–2586, 2013.

[125] D. J. Huggins, “Correlations in liquid water for the TIP3P-Ewald, TIP4P-2005, TIP5P-Ewald, and SWM4-NDP models,” *J. Chem. Phys.*, vol. 136, no. 6, 2012.

[126] T. Young, R. Abel, B. Kim, B. J. Berne, and R. A. Friesner, “Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding,” *Proceedings of the National Academy of Sciences, USA*, vol. 104, pp. 808–813, 2006.

[127] Z. Li and T. Lazaridis, “Thermodynamics of buried water clusters at a protein-ligand binding interface,” *J. Phys. Chem. B*, vol. 110, no. 3, pp. 1464–1475, 2006.

[128] D. D. Robinson, W. Sherman, and R. Farid, “Understanding kinase selectivity through energetic analysis of binding site waters,” *ChemMedChem*, vol. 5, pp. 618–627, 2010.

[129] D. J. Huggins, M. Marsh, and M. C. Payne, “Thermodynamic properties of water molecules at a protein-protein interaction surface,” *Journal of Chemical Theory and Computation*, vol. 7, pp. 3514–3522, 2011.

[130] Y. Thielmann, O. H. Weiergräber, J. Mohrlüder, and D. Willbold, "Structural framework of the GABARAP-calreticulum interface – implications for substrate binding to endoplasmic reticulum chaperones," *FEBS Journal*, vol. 276, pp. 1140–1152, 2008.

[131] A. H. Lystad, Y. Ichimura, K. Takagi, Y. Yang, S. Pankiv, Y. Kanegae, S. Kageyama, M. Suzuki, I. Saito, T. Mizushima, M. Komatsu, and A. Simonsen, "Structural determinants in GABARAP required for the selective binding and recruitment of ALFY to LC3B-positive structures," *EMBO Reports*, vol. 15, pp. 557–565, 2014.

[132] H. M. Genau, J. Huber, F. Baschieri, M. Akutsu, V. Doetsch, H. Farhan, V. Rogov, and C. Behrends, "CUL3-KBTBD6/KBTBD7 ubiquitin ligase cooperates with GABARAP proteins to spatially restrict TIAM1-RAC1 signaling," *Molecular Cell*, vol. 57, pp. 995–1010, 2015.

[133] M. K. Ramarao, M. J. Bianchetta, J. Lanken, and J. B. Cohen, "Role of rapsyn tetratricopeptide repeat and coiled-coil domains in self-association and nicotinic acetylcholine receptor clustering," *Journal of Biological Chemistry*, vol. 276, no. 10, pp. 7475–7483, 2001.

[134] B. Zuber and N. Unwin, "Structure and superorganization of acetylcholine receptor-rapsyn complexes," *Proc. Natl. Acad. Sci.*, vol. 110, pp. 10622–10627, jun 2013.

[135] J. Kirsch, D. Langosch, P. Prior, U. Z. Littauer, B. Schmitt, and H. Betz, "The 93-kDa glycine receptor-associated protein binds to tubulin," *Journal of Biological Chemistry*, vol. 266, pp. 22242–22245, 1991.

[136] M. Sola, V. N. Bavro, J. Timmins, T. Franz, S. Ricard-Blum, G. Schoehn, R. W. H. Ruigrok, I. Paarmann, T. Saiyed, G. A. O'Sullivan, B. Schmitt, H. Betz, and W. Weissenhorn, "Structural basis of dynamic glycine receptor clustering by gephyrin," *EMBO Journal*, vol. 23, pp. 2510–2519, 2004.

[137] V. Tretter, T. C. Jacob, J. Mukherjee, J.-M. Fritschy, M. N. Pangalos, and S. J. Moss, "The clustering of $\mathrm{GABA_A}$ receptor subtypes at inhibitory synapses is facilitated via the direct binding of receptor $\alpha 2$ subunits to gephyrin," *Journal of Neuroscience*, vol. 28, pp. 1356–1365, 2008.

[138] V. Tretter, B. Kerschner, I. Milenkovic, S. L. Ramsden, J. Ramerstorfer, L. Saiepour, J.-M. Maric, S. J. Moss, H. Schindelin, R. J. Harvey, W. Sieghart, and K. Harvey, "Molecular basis of the γ-aminobutyric acid A receptor $\alpha 3$ subunit interaction with the clustering protein gephyrin," *Journal of Biological Chemistry*, vol. 286, pp. 37702–37711, 2011.

[139] L. Saiepour, C. Fuchs, A. Patrizi, M. Sassoe-Pognetto, R. J. Harvey, and K. Harvey, "Complex role of collybistin and gephyrin in $\mathrm{GABA_A}$ receptor clustering," *Journal of Biological Chemistry*, vol. 285, pp. 29623–29631, 2010.

[140] J. Mukherjee, K. Kretschmannova, G. Gouzer, H.-M. Maric, S. Ramsden, V. Tretter, K. Harvey, P. A. Davies, A. Triller, H. Schindelin, and S. J. Moss, "The residence time of $\mathrm{GABA_A}$Rs at inhibitory synapses is determined by direct binding of the receptor $\alpha 1$ subunit to gephyrin," *Journal of Neuroscience*, vol. 31, pp. 14677–14687, 2011.

[141] H. M. Maric, V. B. Kasaragod, T. J. Hausrat, M. Kneussel, V. Tretter, K. Stromgaard, and H. Schindelin, “Molecular basis of the alternative recruitment of $\mathrm{GABA_A}$ versus glycine receptors through gephyrin,” *Nature Communications*, vol. 5, 2014.

[142] S. Kins, H. Betz, and J. Kirsch, “Collybistin, a newly identified brain-specific GEF, induces submembrane clustering of gephyrin,” *Nature Neuroscience*, vol. 3, no. 1, pp. 22–29, 2000.

[143] T.-T. Chiou, B. Bonhomme, H. Jin, C. P. Miralles, H. Xiao, Z. Fu, R. J. Harvey, K. Harvey, S. Vicini, and A. L. D. Blas, “Differential regulation of the postsynaptic clustering of γ-aminobutyric acid type A ($\mathrm{GABA_A}$) receptors by collybistin isoforms,” *Journal of Biological Chemistry*, vol. 286, pp. 22456–22468, 2011.

[144] T. Luu, P. W. Gage, and M. L. Tierney, “GABA increases both the conductance and mean open time of recombinant $\mathrm{GABA_A}$ channels co-expressed with GABARAP,” *Journal of Biological Chemistry*, vol. 281, pp. 35699–35708, 2006.

[145] A. B. Everitt, T. Luu, B. Cromer, M. L. Tierney, B. Birnir, R. W. Olsen, and P. W. Gage, “Conductance of recombinant $\mathrm{GABA_A}$ channels is increased in cells co-expressing $\mathrm{GABA_A}$ receptor-associated protein,” *Journal of Biological Chemistry*, vol. 279, pp. 21701–21706, 2004.

[146] Fernandez Ariel and Scott Ridgway, “Dehydron: A Structurally Encoded Signal for Protein Interaction,” *Biophysical Journal*, vol. 85, pp. 1914–1928, jun 2003.

[147] Fernandez Ariel and Crespo Alejandro, “Protein wrapping: a molecular marker for association, aggregation and drug design,” *Chemical Society Reviews*, vol. 37, no. 11, pp. 2373–2382, 2008.

[148] X. Zhang, A. Crespo, and A. Fernandez, “Turning promiscuous kinase inhibitors into safer drugs,” *Trends in Biotechnology*, vol. 26, no. 6, pp. 295 – 301, 2008.

[149] A. Fernandez, “Incomplete protein packing as a selectivity filter in drug design,” *Structure*, vol. 13, no. 12, pp. 1829 – 1836, 2005.

[150] A. Crespo and A. Fernandez, “Kinase packing defects as drug targets,” *Drug Discovery Today*, vol. 12, no. 21, pp. 917 – 923, 2007.

[151] Gerogiokas Georgios, Southey Michelle W. Y., Mazanetz Michael P., Heifetz Alexander, Bodkin Michael, Law Richard J., Henchman Richard H., and Michel J., “Assessment of Hydration Thermodynamics at Protein Interfaces with Grid Cell Theory,” *The Journal of Physical Chemistry B*, vol. 120, no. 40, pp. 10442–10452, 2016. doi: 10.1021/acs.jpcb.6b07993.

[152] Beuming Thijs, Che Ye, Abel Robert, Kim Byungchan, Shanmugasundaram Veerabahu, and Sherman Woody, “Thermodynamic analysis of water molecules at the surface of proteins and applications to binding site prediction and characterization,” *Proteins*, vol. 80, no. 3, pp. 871–883, 2011. doi: 10.1002/prot.23244.

[153] Vukovic Sinisa and Huggins David J., “Quantitative metrics for drug–target ligandability,” *Drug Discovery Today*, vol. 23, no. 6, p. 1258–1266, 2018.

[154] Huggins David J., “Studying the role of cooperative hydration in stabilizing folded protein states,” *Journal of Structural Biology*, vol. 196, no. 3, p. 394–406, 2016.

[155] Jeffrey, George A., *An introduction to hydrogen bonding.* Oxford University Press.

[156] Cheng Alan C, Coleman Ryan G, Smyth Kathleen T, Cao Qing, Soulard Patricia, Caffrey Daniel R, Salzberg Anna C, and Huang Enoch S, “Structure-based maximal affinity model predicts small-molecule druggability,” *Nature Biotechnology*, vol. 25, p. 71, jan 2007.

[157] P. M. Vaidya, “An O(n log n) algorithm for the all-nearest-neighbors Problem,” *Discrete Comput. Geom.*, vol. 4, no. 1, pp. 101–115, 1989.

[158] N. Matubayasi, E. Gallicchio, and R. M. Levy, “On the local and nonlocal components of solvation thermodynamics and their relation to solvation shell models,” *J. Chem. Phys.*, vol. 109, no. 12, pp. 4864–4872, 1998.

Appendix A

Appendix A: Tables of Data

A.1 Supplementary Tables of Data

A.1.1 Probe Free Energies

The free energies are given for all of the probe atoms.

A.1.2 List of Proteins Analysed

The PDB tags for files analysed:

Atom	ε	rmin/2	q	ΔG_{solv}
BRGA1	-0.48	1.97	-0.1	0.17
BRGA2	-0.53	2.05	-0.04	1.00
BRGA3	-0.54	2	-0.01	1.13
BRGR1	-0.42	2.07	-0.1	0.5
CG1N1	-0.18	1.87	0.36	-0.31
CG1N1	-0.18	1.87	0.39	-1.23
CG1T1	-0.17	1.87	-0.08	1.85
CG252O	-0.068	2.09	0.09	4.62
CG252O	-0.068	2.09	0.1	4.66
CG252O	-0.068	2.09	0.22	3.89
CG2D1O	-0.068	2.09	-0.04	3.84
CG2D1O	-0.068	2.09	-0.06	3.54
CG2D1O	-0.068	2.09	-0.1	2.76
CG2D1O	-0.068	2.09	-0.13	2.12
CG2D1O	-0.068	2.09	-0.14	1.91
CG2D1	-0.068	2.09	-0.15	1.73
CG2D1	-0.068	2.09	-0.18	0.96
CG2D1	-0.068	2.09	-0.24	-1.34
CG2D1	-0.068	2.09	0.23	4.01
CG2D1	-0.068	2.09	0.37	1.12
CG2D2	-0.064	2.08	-0.42	-10.94
CG2D2	-0.064	2.08	-0.5	-17.05
CG2D2	-0.064	2.08	-0.53	-19.71
CG2DC1	-0.068	2.09	-0.11	2.48
CG2DC1	-0.068	2.09	-0.25	-1.62
CG2DC1	-0.068	2.09	-0.3	-3.78
CG2DC1	-0.068	2.09	0.36	1.34
CG2DC3	-0.064	2.08	-0.46	-14.01
CG2DC3	-0.064	2.08	-0.49	-16.07
CG2DC3	-0.064	2.08	-0.52	-18.8
CG2DC3	-0.064	2.08	-0.58	-24.61
CG2N1	-0.11	2.0	0.59	-7.69
CG2N1	-0.11	2.0	0.64	-10.36
CG2N1	-0.11	2.0	0.66	-11.31
CG2O1	-0.11	2.0	0.42	-1.13
CG2O1	-0.11	2.0	0.43	-1.41
CG2O1	-0.11	2.0	0.51	-4.3
CG2O1	-0.11	2.0	0.52	-4.75
CG2O1	-0.11	2.0	0.55	-5.73
CG2O1	-0.11	2.0	0.58	-7.32
CG2O1	-0.11	2.0	0.63	-9.88
CG2O1	-0.11	2.0	0.68	-12.74
CG2O2	-0.098	1.7	0.38	-1.61
CG2O2	-0.098	1.7	0.46	-4.92
CG2O2	-0.098	1.7	0.63	-14.42
CG2O3	-0.07	2.0	0.3	2.35
CG2O3	-0.07	2.0	0.34	1.44
CG2O3	-0.07	2.0	0.52	-4.96
CG2O3	-0.07	2.0	0.62	-9.28
CG2O4	-0.06	1.8	0.2	3.04
CG2O4	-0.06	1.8	0.24	2.44
CG2O4	-0.06	1.8	0.27	1.98
CG2O5	-0.09	2.0	0.34	1.33
CG2O5	-0.09	2.0	0.36	0.85
CG2O5	-0.09	2.0	0.38	0.35
CG2O5	-0.09	2.0	0.4	-0.49
CG2O6	-0.07	2.0	0.2	3.71
CG2O6	-0.07	2.0	0.22	3.67
CG2O6	-0.07	2.0	0.23	3.44
CG2O6	-0.07	2.0	0.6	-8.43
CG2O7	-0.058	1.563	0.6	-14.98
CG2R51	-0.05	2.1	-0.01	4.38
CG2R51	-0.05	2.1	-0.02	4.28
CG2R51	-0.05	2.1	-0.03	4.2
CG2R51	-0.05	2.1	-0.04	3.97
CG2R51	-0.05	2.1	-0.05	3.96
CG2R51	-0.05	2.1	-0.06	3.8
CG2R51	-0.05	2.1	-0.08	3.39
CG2R51	-0.05	2.1	-0.09	3.22
CG2R51	-0.05	2.1	-0.15	1.85
CG2R51	-0.05	2.1	-0.16	1.71
CG2R51	-0.05	2.1	-0.17	1.28
CG2R51	-0.05	2.1	-0.18	0.91
CG2R51	-0.05	2.1	-0.2	0.33
CG2R51	-0.05	2.1	-0.22	-0.24
CG2R51	-0.05	2.1	-0.23	-0.63
CG2R51	-0.05	2.1	-0.24	-1.07
CG2R51	-0.05	2.1	-0.25	-1.44
CG2R51	-0.05	2.1	-0.27	-2.41
CG2R51	-0.05	2.1	-0.28	-2.89
CG2R51	-0.05	2.1	-0.3	-3.74
CG2R51	-0.05	2.1	-0.33	-5.36
CG2R51	-0.05	2.1	-0.35	-6.39
CG2R51	-0.05	2.1	-0.36	-7.12
CG2R51	-0.05	2.1	-0.4	-9.46
CG2R51	-0.05	2.1	-0.41	-10.02

Atom	ε	rmin/2	q	ΔG_{solv}
CG2R51	-0.05	2.1	-0.43	-11.34
CG2R51	-0.05	2.1	0.02	4.70
CG2R51	-0.05	2.1	0.05	4.85
CG2R51	-0.05	2.1	0.07	4.79
CG2R51	-0.05	2.1	0.1	4.94
CG2R51	-0.05	2.1	0.11	4.98
CG2R51	-0.05	2.1	0.12	4.70
CG2R51	-0.05	2.1	0.15	4.75
CG2R51	-0.05	2.1	0.17	4.68
CG2R51	-0.05	2.1	0.19	4.53
CG2R51	-0.05	2.1	0.2	4.33
CG2R51	-0.05	2.1	0.22	4.15
CG2R51	-0.05	2.1	0.25	3.66
CG2R51	-0.05	2.1	0.28	3.19
CG2R52	-0.02	2.2	0.07	5.39
CG2R52	-0.02	2.2	0.1	5.46
CG2R52	-0.02	2.2	0.14	5.35
CG2R52	-0.02	2.2	0.18	5.02
CG2R52	-0.02	2.2	0.2	4.88
CG2R52	-0.02	2.2	0.23	4.17
CG2R52	-0.02	2.2	0.28	3.6
CG2R52	-0.02	2.2	0.32	2.98
CG2R52	-0.02	2.2	0.35	2.37
CG2R53	-0.02	2.2	0.16	5.16
CG2R53	-0.02	2.2	0.22	4.72
CG2R53	-0.02	2.2	0.24	4.34
CG2R53	-0.02	2.2	0.25	4.28
CG2R53	-0.02	2.2	0.26	4.19
CG2R53	-0.02	2.2	0.29	3.42
CG2R53	-0.02	2.2	0.3	3.20
CG2R53	-0.02	2.2	0.34	2.56
CG2R53	-0.02	2.2	0.37	1.72
CG2R53	-0.02	2.2	0.42	0.32
CG2R53	-0.02	2.2	0.45	-0.82
CG2R53	-0.02	2.2	0.47	-1.46
CG2R53	-0.02	2.2	0.49	-2.04
CG2R53	-0.02	2.2	0.52	-3.29
CG2R53	-0.02	2.2	0.61	-7.83
CG2R53	-0.02	2.2	0.67	-10.91
CG2R61	-0.07	1.9924	-0.01	3.91
CG2R61	-0.07	1.9924	-0.02	3.74
CG2R61	-0.07	1.9924	-0.06	3.26
CG2R61	-0.07	1.9924	-0.08	3.03
CG2R61	-0.07	1.9924	-0.1	2.43
CG2R61	-0.07	1.9924	-0.11	2.31
CG2R61	-0.07	1.9924	-0.115	2.29
CG2R61	-0.07	1.9924	-0.12	2.03
CG2R61	-0.07	1.9924	-0.13	1.86
CG2R61	-0.07	1.9924	-0.14	1.46
CG2R61	-0.07	1.9924	-0.15	1.33
CG2R61	-0.07	1.9924	-0.155	1.12
CG2R61	-0.07	1.9924	-0.16	0.91
CG2R61	-0.07	1.9924	-0.17	0.57
CG2R61	-0.07	1.9924	-0.18	0.58
CG2R61	-0.07	1.9924	-0.19	0.11
CG2R61	-0.07	1.9924	-0.2	-0.16
CG2R61	-0.07	1.9924	-0.21	-0.59
CG2R61	-0.07	1.9924	-0.22	-0.8
CG2R61	-0.07	1.9924	-0.23	-1.27
CG2R61	-0.07	1.9924	-0.24	-1.8
CG2R61	-0.07	1.9924	-0.25	-2.2
CG2R61	-0.07	1.9924	-0.26	-2.63
CG2R61	-0.07	1.9924	-0.27	-2.97
CG2R61	-0.07	1.9924	-0.28	-3.39
CG2R61	-0.07	1.9924	-0.29	-4.02
CG2R61	-0.07	1.9924	-0.3	-4.45
CG2R61	-0.07	1.9924	-0.32	-5.45
CG2R61	-0.07	1.9924	-0.33	-5.9
CG2R61	-0.07	1.9924	-0.34	-6.56
CG2R61	-0.07	1.9924	-0.35	-7.13
CG2R61	-0.07	1.9924	-0.36	-7.88
CG2R61	-0.07	1.9924	-0.38	-9.19
CG2R61	-0.07	1.9924	-0.4	-10.24
CG2R61	-0.07	1.9924	-0.43	-12.54
CG2R61	-0.07	1.9924	-0.44	-13.41
CG2R61	-0.07	1.9924	-0.45	-14.23
CG2R61	-0.07	1.9924	-0.46	-14.8
CG2R61	-0.07	1.9924	-0.6	-28.11
CG2R61	-0.07	1.9924	0.02	4.35
CG2R61	-0.07	1.9924	0.03	4.22
CG2R61	-0.07	1.9924	0.04	4.37
CG2R61	-0.07	1.9924	0.05	4.26
CG2R61	-0.07	1.9924	0.07	4.40
CG2R61	-0.07	1.9924	0.08	4.39
CG2R61	-0.07	1.9924	0.09	4.45
CG2R61	-0.07	1.9924	0.1	4.30

Atom	ε	rmin/2	q	ΔG_{solv}
CG2R61	-0.07	1.9924	0.11	4.37
CG2R61	-0.07	1.9924	0.12	4.40
CG2R61	-0.07	1.9924	0.13	4.15
CG2R61	-0.07	1.9924	0.14	4.26
CG2R61	-0.07	1.9924	0.16	4.03
CG2R61	-0.07	1.9924	0.17	4.02
CG2R61	-0.07	1.9924	0.18	3.84
CG2R61	-0.07	1.9924	0.185	3.85
CG2R61	-0.07	1.9924	0.19	3.97
CG2R61	-0.07	1.9924	0.21	3.67
CG2R61	-0.07	1.9924	0.215	3.69
CG2R61	-0.07	1.9924	0.22	3.67
CG2R61	-0.07	1.9924	0.23	3.39
CG2R61	-0.07	1.9924	0.24	3.29
CG2R61	-0.07	1.9924	0.25	3.04
CG2R61	-0.07	1.9924	0.28	2.65
CG2R61	-0.07	1.9924	0.29	2.37
CG2R61	-0.07	1.9924	0.3	2.27
CG2R61	-0.07	1.9924	0.31	1.93
CG2R61	-0.07	1.9924	0.32	1.85
CG2R61	-0.07	1.9924	0.33	1.36
CG2R61	-0.07	1.9924	0.34	1.29
CG2R61	-0.07	1.9924	0.345	1.13
CG2R61	-0.07	1.9924	0.35	0.95
CG2R61	-0.07	1.9924	0.36	0.8
CG2R61	-0.07	1.9924	0.4	-0.44
CG2R61	-0.07	1.9924	0.45	-2.1
CG2R61	-0.07	1.9924	0.47	-2.83
CG2R61	-0.07	1.9924	0.49	-3.45
CG2R62	-0.09	1.9	-0.05	2.94
CG2R62	-0.09	1.9	-0.1	2.09
CG2R62	-0.09	1.9	-0.13	1.34
CG2R62	-0.09	1.9	-0.15	0.95
CG2R62	-0.09	1.9	-0.2	-0.65
CG2R62	-0.09	1.9	-0.22	-1.51
CG2R62	-0.09	1.9	-0.26	-3.21
CG2R62	-0.09	1.9	0.05	4.01
CG2R62	-0.09	1.9	0.11	3.96
CG2R62	-0.09	1.9	0.15	3.80
CG2R62	-0.09	1.9	0.16	3.63
CG2R62	-0.09	1.9	0.17	3.62
CG2R62	-0.09	1.9	0.18	3.55
CG2R62	-0.09	1.9	0.2	3.42
CG2R63	-0.1	1.9	0.16	3.51
CG2R63	-0.1	1.9	0.4	-0.87
CG2R63	-0.1	1.9	0.44	-2.41
CG2R63	-0.1	1.9	0.5	-4.6
CG2R63	-0.1	1.9	0.51	-5.04
CG2R63	-0.1	1.9	0.52	-5.58
CG2R63	-0.1	1.9	0.53	-6.13
CG2R63	-0.1	1.9	0.54	-6.6
CG2R63	-0.1	1.9	0.55	-7.06
CG2R64	-0.04	2.1	0.44	-0.93
CG2R64	-0.04	2.1	0.46	-1.6
CG2R64	-0.04	2.1	0.48	-2.37
CG2R64	-0.04	2.1	0.5	-2.98
CG2R64	-0.04	2.1	0.52	-3.82
CG2R64	-0.04	2.1	0.6	-7.75
CG2R64	-0.04	2.1	0.62	-9.02
CG2R64	-0.04	2.1	0.65	-10.52
CG2R66	-0.07	1.9	0.11	4.11
CG2R66	-0.07	1.9	0.17	3.83
CG2R66	-0.07	1.9	0.22	3.16
CG2R66	-0.07	1.9	0.27	2.17
CG2R66	-0.07	1.9	0.28	2.21
CG2R71	-0.067	1.9948	-0.15	1.4
CG2R71	-0.067	1.9948	-0.22	-0.95
CG2R71	-0.067	1.9948	-0.24	-1.72
CG2RC0	-0.099	1.86	-0.06	2.66
CG2RC0	-0.099	1.86	-0.11	1.69
CG2RC0	-0.099	1.86	0.01	3.62
CG2RC0	-0.099	1.86	0.03	3.75
CG2RC0	-0.099	1.86	0.06	3.79
CG2RC0	-0.099	1.86	0.11	3.75
CG2RC0	-0.099	1.86	0.15	3.57
CG2RC0	-0.099	1.86	0.17	3.49
CG2RC0	-0.099	1.86	0.2	3.16
CG2RC0	-0.099	1.86	0.21	2.99
CG2RC0	-0.099	1.86	0.23	2.67
CG2RC0	-0.099	1.86	0.24	2.68
CG2RC0	-0.099	1.86	0.25	2.38
CG2RC0	-0.099	1.86	0.26	2.27
CG2RC0	-0.099	1.86	0.28	1.7
CG2RC0	-0.099	1.86	0.29	1.62
CG2RC0	-0.099	1.86	0.3	1.32
CG2RC0	-0.099	1.86	0.31	1.29

Atom	ε	rmin/2	q	ΔG_{solv}
CG2RC0	-0.099	1.86	0.33	0.93
CG2RC0	-0.099	1.86	0.41	-1.52
CG2RC0	-0.099	1.86	0.43	-2.22
CG2RC0	-0.099	1.86	0.44	-2.91
CG2RC0	-0.099	1.86	0.45	-3.09
CG2RC0	-0.099	1.86	0.64	-12.59
CG2RC7	-0.099	1.86	0.07	3.86
CG301	-0.032	2.0	-0.04	3.9
CG301	-0.032	2.0	-0.12	2.36
CG301	-0.032	2.0	-0.24	-1.58
CG301	-0.032	2.0	0.174	4.35
CG301	-0.032	2.0	0.23	3.66
CG301	-0.032	2.0	0.26	3.23
CG301	-0.032	2.0	0.3	2.43
CG301	-0.032	2.0	0.4	-0.42
CG302	-0.02	2.3	0.34	3.07
CG302	-0.02	2.3	0.38	1.97
CG311	-0.032	2.0	-0.01	4.15
CG311	-0.032	2.0	-0.02	4.19
CG311	-0.032	2.0	-0.09	3.09
CG311	-0.032	2.0	-0.19	0.44
CG311	-0.032	2.0	0.07	4.7
CG311	-0.032	2.0	0.1	4.54
CG311	-0.032	2.0	0.12	4.65
CG311	-0.032	2.0	0.14	4.57
CG311	-0.032	2.0	0.17	4.31
CG312	-0.042	2.05	0.21	4.15
CG312	-0.042	2.05	0.24	3.75
CG314	-0.031	2.165	0.21	4.64
CG314	-0.031	2.165	0.29	3.57
CG321	-0.056	2.01	-0.01	4.1
CG321	-0.056	2.01	-0.02	4.03
CG321	-0.056	2.01	-0.03	3.93
CG321	-0.056	2.01	-0.04	3.74
CG321	-0.056	2.01	-0.05	3.52
CG321	-0.056	2.01	-0.06	3.35
CG321	-0.056	2.01	-0.07	3.25
CG321	-0.056	2.01	-0.08	3.13
CG321	-0.056	2.01	-0.09	2.95
CG321	-0.056	2.01	-0.1	2.67
CG321	-0.056	2.01	-0.11	2.43
CG321	-0.056	2.01	-0.12	2.17
CG321	-0.056	2.01	-0.13	2.06
CG321	-0.056	2.01	-0.14	1.78
CG321	-0.056	2.01	-0.17	0.87
CG321	-0.056	2.01	-0.18	0.72
CG321	-0.056	2.01	-0.22	-0.73
CG321	-0.056	2.01	-0.26	-2.39
CG321	-0.056	2.01	-0.28	-3.44
CG321	-0.056	2.01	-0.3	-4.13
CG321	-0.056	2.01	0.02	4.23
CG321	-0.056	2.01	0.04	4.41
CG321	-0.056	2.01	0.05	4.48
CG321	-0.056	2.01	0.052	4.47
CG321	-0.056	2.01	0.06	4.60
CG321	-0.056	2.01	0.07	4.70
CG321	-0.056	2.01	0.08	4.56
CG321	-0.056	2.01	0.11	4.44
CG321	-0.056	2.01	0.14	4.50
CG321	-0.056	2.01	0.16	4.39
CG321	-0.056	2.01	0.18	4.15
CG321	-0.056	2.01	0.21	3.75
CG321	-0.056	2.01	0.22	3.68
CG321	-0.056	2.01	0.29	2.59
CG322	-0.06	1.9	-0.06	3.05
CG323	-0.11	2.2	-0.38	-7.39
CG323	-0.11	2.2	-0.47	-12.56
CG324	-0.055	2.175	-0.1	3.25
CG324	-0.055	2.175	0.03	4.94
CG324	-0.055	2.175	0.13	5.04
CG324	-0.055	2.175	0.15	4.84
CG324	-0.055	2.175	0.2	4.63
CG324	-0.055	2.175	0.21	4.60
CG324	-0.055	2.175	0.26	3.85
CG331	-0.078	2.05	-0.01	3.96
CG331	-0.078	2.05	-0.02	3.93
CG331	-0.078	2.05	-0.03	3.84
CG331	-0.078	2.05	-0.04	3.70
CG331	-0.078	2.05	-0.05	3.86
CG331	-0.078	2.05	-0.06	3.50
CG331	-0.078	2.05	-0.07	3.15
CG331	-0.078	2.05	-0.08	3.10
CG331	-0.078	2.05	-0.09	2.86
CG331	-0.078	2.05	-0.1	2.68

Atom	ε	rmin/2	q	ΔG_{solv}
CG331	-0.078	2.05	-0.11	2.41
CG331	-0.078	2.05	-0.12	2.34
CG331	-0.078	2.05	-0.15	1.57
CG331	-0.078	2.05	-0.16	1.26
CG331	-0.078	2.05	-0.17	0.94
CG331	-0.078	2.05	-0.18	0.75
CG331	-0.078	2.05	-0.19	0.34
CG331	-0.078	2.05	-0.2	-0.11
CG331	-0.078	2.05	-0.21	-0.30
CG331	-0.078	2.05	-0.22	-0.63
CG331	-0.078	2.05	-0.23	-1.02
CG331	-0.078	2.05	-0.24	-1.48
CG331	-0.078	2.05	-0.27	-2.76
CG331	-0.078	2.05	-0.3	-4.11
CG331	-0.078	2.05	-0.31	-4.36
CG331	-0.078	2.05	-0.35	-6.62
CG331	-0.078	2.05	-0.37	-7.94
CG331	-0.078	2.05	-0.41	-10.38
CG331	-0.078	2.05	-0.45	-13.27
CG331	-0.078	2.05	-0.51	-18.07
CG331	-0.078	2.05	-0.52	-19.00
CG331	-0.078	2.05	-0.53	-19.90
CG331	-0.078	2.05	0.01	4.21
CG331	-0.078	2.05	0.02	4.31
CG331	-0.078	2.05	0.05	4.48
CG331	-0.078	2.05	0.07	4.47
CG331	-0.078	2.05	0.16	4.23
CG334	-0.077	2.215	-0.35	-5.57
CG334	-0.077	2.215	0.11	5.09
CG334	-0.077	2.215	0.15	4.96
CG334	-0.077	2.215	0.16	4.99
CG334	-0.077	2.215	0.18	4.71
CG334	-0.077	2.215	0.2	4.44
CG3AM0	-0.07	1.97	-0.06	3.39
CG3AM1	-0.078	1.98	-0.06	3.18
CG3AM2	-0.08	1.99	-0.06	3.29
CG3C41	-0.065	2.02	-0.01	3.99
CG3C41	-0.065	2.02	-0.18	0.61
CG3C41	-0.065	2.02	0.01	4.22
CG3C51	-0.036	2.01	-0.01	4.24
CG3C51	-0.036	2.01	-0.06	3.53
CG3C51	-0.036	2.01	-0.09	2.95
CG3C51	-0.036	2.01	-0.12	2.40
CG3C51	-0.036	2.01	0.01	4.55
CG3C51	-0.036	2.01	0.02	4.55
CG3C51	-0.036	2.01	0.05	4.61
CG3C51	-0.036	2.01	0.08	4.70
CG3C51	-0.036	2.01	0.11	4.65
CG3C51	-0.036	2.01	0.14	4.58
CG3C51	-0.036	2.01	0.16	4.37
CG3C52	-0.06	2.02	-0.01	3.82
CG3C52	-0.06	2.02	-0.04	3.65
CG3C52	-0.06	2.02	-0.08	3.15
CG3C52	-0.06	2.02	-0.09	2.91
CG3C52	-0.06	2.02	-0.1	2.62
CG3C52	-0.06	2.02	-0.12	2.30
CG3C52	-0.06	2.02	-0.15	1.48
CG3C52	-0.06	2.02	-0.18	0.58
CG3C52	-0.06	2.02	-0.19	0.20
CG3C52	-0.06	2.02	-0.21	-0.46
CG3C52	-0.06	2.02	0.02	4.34
CG3C52	-0.06	2.02	0.05	4.58
CG3C52	-0.06	2.02	0.07	4.51
CG3C52	-0.06	2.02	0.08	4.57
CG3C52	-0.06	2.02	0.1	4.45
CG3C52	-0.06	2.02	0.12	4.59
CG3C52	-0.06	2.02	0.16	4.30
CG3C52	-0.06	2.02	0.17	4.21
CG3C52	-0.06	2.02	0.38	0.48
CG3C52	-0.06	2.02	0.45	-1.9
CG3C53	-0.035	2.175	0.11	5.26
CG3C53	-0.035	2.175	0.16	5.06
CG3C54	-0.059	2.185	-0.17	1.77
CG3C54	-0.059	2.185	-0.19	1.03
CG3C54	-0.059	2.185	-0.22	0.08
CG3C54	-0.059	2.185	-0.3	-3.13
CG3C54	-0.059	2.185	-0.33	-4.6
CG3C54	-0.059	2.185	-0.35	-5.62
CG3C54	-0.059	2.185	-0.41	-9.15
CG3C54	-0.059	2.185	0.16	4.97
CG3RC1	-0.032	2	0.03	4.44
CLGA1	-0.343	1.91	-0.04	1.71
CLGA1	-0.343	1.91	-0.1	0.81
CLGA3	-0.31	1.91	0.14	2.74
CLGR1	-0.32	1.93	-0.13	0.22
CLGR1	-0.32	1.93	-0.18	-1.07
FGA1	-0.135	1.63	-0.22	-2.98
FGA2	-0.105	1.63	-0.17	-0.82
FGA2	-0.105	1.63	-0.19	-1.58
FGA2	-0.105	1.63	-0.28	-5.94
FGA3	-0.097	1.6	-0.14	0.08
FGA3	-0.097	1.6	-0.15	-0.33
FGP1	-0.097	1.6	-0.54	-30.89
FGR1	-0.12	1.7	-0.19	-1.34
FGR1	-0.12	1.7	-0.21	-2.11
IGR1	-0.55	2.19	-0.08	0.55

Atom	ε	rmin/2	q	ΔG_{solv}
NG1T1	-0.18	1.79	-0.46	-17.33
NG1T1	-0.18	1.79	-0.47	-18.13
NG2D1	-0.2	1.85	-0.31	-6.36
NG2D1	-0.2	1.85	-0.6	-30
NG2O1	-0.2	1.85	0.11	3.11
NG2O1	-0.2	1.85	0.14	3.06
NG2O1	-0.2	1.85	0.4	-1.6
NG2P1	-0.2	1.85	-0.4	-11.84
NG2R43	-0.2	1.85	-0.54	-23.9
NG2R50	-0.2	1.85	-0.1	1.44
NG2R50	-0.2	1.85	-0.32	-6.92
NG2R50	-0.2	1.85	-0.37	-9.89
NG2R50	-0.2	1.85	-0.41	-12.63
NG2R50	-0.2	1.85	-0.42	-13.25
NG2R50	-0.2	1.85	-0.45	-15.79
NG2R50	-0.2	1.85	-0.49	-18.94
NG2R50	-0.2	1.85	-0.53	-22.64
NG2R50	-0.2	1.85	-0.57	-26.7
NG2R50	-0.2	1.85	-0.58	-28
NG2R50	-0.2	1.85	-0.61	-31.12
NG2R50	-0.2	1.85	-0.64	-34.42
NG2R50	-0.2	1.85	-0.66	-37.2
NG2R51	-0.2	1.85	-0.02	2.4
NG2R51	-0.2	1.85	-0.04	2.22
NG2R51	-0.2	1.85	-0.05	2.24
NG2R51	-0.2	1.85	-0.06	1.89
NG2R51	-0.2	1.85	-0.19	-1.05
NG2R51	-0.2	1.85	-0.22	-2.12
NG2R51	-0.2	1.85	-0.28	-4.74
NG2R51	-0.2	1.85	-0.36	-9.38
NG2R51	-0.2	1.85	-0.51	-20.78
NG2R51	-0.2	1.85	0.28	1.41
NG2R52	-0.2	1.85	-0.27	-4.38
NG2R53	-0.2	1.85	-0.18	-0.86
NG2R53	-0.2	1.85	-0.25	-3.53
NG2R60	-0.06	1.89	-0.56	-26.45
NG2R60	-0.06	1.89	-0.58	-28.58
NG2R60	-0.06	1.89	-0.6	-31.18
NG2R60	-0.06	1.89	-0.61	-32.03
NG2R60	-0.06	1.89	-0.64	-36.36
NG2R60	-0.06	1.89	-0.69	-42.57
NG2R61	-0.2	1.85	-0.07	1.88
NG2R61	-0.2	1.85	-0.13	0.77
NG2R61	-0.2	1.85	-0.46	-16.66
NG2R61	-0.2	1.85	-0.52	-21.88
NG2R62	-0.05	2.06	-0.41	-10.48
NG2R62	-0.05	2.06	-0.44	-13.02
NG2R62	-0.05	2.06	-0.45	-13.45
NG2R62	-0.05	2.06	-0.49	-16.62
NG2R62	-0.05	2.06	-0.5	-17.79
NG2R62	-0.05	2.06	-0.53	-20.24
NG2R62	-0.05	2.06	-0.58	-25.38
NG2R62	-0.05	2.06	-0.65	-32.91
NG2R62	-0.05	2.06	-0.66	-33.67
NG2S0	-0.2	1.85	-0.29	-5.23
NG2S0	-0.2	1.85	-0.33	-7.47
NG2S0	-0.2	1.85	-0.35	-8.84
NG2S1	-0.2	1.85	-0.34	-8.16
NG2S1	-0.2	1.85	-0.38	-10.63
NG2S1	-0.2	1.85	-0.47	-17.58
NG2S2	-0.2	1.85	-0.62	-32.29
NG2S2	-0.2	1.85	-0.69	-40.79
NG2S3	-0.2	1.85	-0.68	-39.4
NG301	-0.035	2	-0.27	-2.82
NG301	-0.035	2	-0.63	-32.91
NG311	-0.045	2	-0.36	-8.03
NG311	-0.045	2	-0.47	-16.02
NG311	-0.045	2	-0.48	-16.83
NG311	-0.045	2	-0.54	-22.61
NG311	-0.045	2	-0.55	-23.49
NG311	-0.045	2	-0.56	-24.56
NG311	-0.045	2	-0.57	-25.71
NG311	-0.045	2	-0.6	-29.14
NG311	-0.045	2	-0.69	-39.4
NG321	-0.06	1.99	-0.46	-15.09
NG321	-0.06	1.99	-0.6	-28.79
NG3P1	-0.2	1.85	0.13	3.21
NG3P3	-0.2	1.85	-0.173	-0.68
NG3P3	-0.2	1.85	-0.3	-5.85
OG2D1	-0.12	1.7	-0.39	-12.88
OG2D1	-0.12	1.7	-0.4	-13.79
OG2D1	-0.12	1.7	-0.41	-14.7
OG2D1	-0.12	1.7	-0.43	-16.58
OG2D1	-0.12	1.7	-0.45	-18.11
OG2D1	-0.12	1.7	-0.46	-19.35
OG2D1	-0.12	1.7	-0.47	-20.41
OG2D1	-0.12	1.7	-0.49	-21.84
OG2D1	-0.12	1.7	-0.51	-24.15
OG2D1	-0.12	1.7	-0.52	-25.37
OG2D1	-0.12	1.7	-0.54	-27.79
OG2D1	-0.12	1.7	-0.55	-28.71
OG2D1	-0.12	1.7	-0.57	-30.96
OG2D1	-0.12	1.7	-0.58	-32.78
OG2D1	-0.12	1.7	-0.63	-38.78
OG2D2	-0.12	1.7	-0.6	-34.95
OG2D2	-0.12	1.7	-0.67	-44.43
OG2D3	-0.05	1.7	-0.46	-20.06
OG2D3	-0.05	1.7	-0.47	-21.39

Atom	ε	rmin/2	q	ΔG_{solv}
OG2D3	-0.05	1.7	-0.48	-22.44
OG2D4	-0.12	1.7	-0.48	-21.19
OG2D5	-0.165	1.692	-0.3	-6.8
OG2N1	-0.12	1.7	-0.29	-6.11
OG2N1	-0.12	1.7	-0.34	-9.22
OG2P1	-0.12	1.7	-0.28	-5.38
OG2P1	-0.12	1.7	-0.32	-7.9
OG2P1	-0.12	1.7	-0.36	-10.43
OG2P1	-0.12	1.7	-0.42	-15.51
OG2P1	-0.12	1.7	-0.64	-40.15
OG2P1	-0.12	1.7	-0.65	-41.87
OG301	-0.1	1.65	-0.28	-6.01
OG301	-0.1	1.65	-0.34	-9.71
OG301	-0.1	1.65	-0.38	-12.91
OG301	-0.1	1.65	-0.39	-13.7
OG301	-0.1	1.65	-0.43	-17.52
OG301	-0.1	1.65	-0.54	-29.19
OG302	-0.1	1.65	-0.24	-3.77
OG302	-0.1	1.65	-0.33	-9.12
OG302	-0.1	1.65	-0.49	-23.62
OG303	-0.1	1.65	-0.36	-11.32
OG303	-0.1	1.65	-0.4	-14.73
OG303	-0.1	1.65	-0.56	-31.45
OG303	-0.1	1.65	-0.57	-33.16
OG303	-0.1	1.65	-0.62	-39.73
OG304	-0.1	1.65	-0.63	-41.09
OG304	-0.1	1.65	-0.68	-48.4
OG311	-0.192	1.765	-0.17	-0.72
OG311	-0.192	1.765	-0.49	-20.44
OG311	-0.192	1.765	-0.51	-22.31
OG311	-0.192	1.765	-0.53	-24.12
OG311	-0.192	1.765	-0.59	-31.07
OG311	-0.192	1.765	-0.6	-32.13
OG311	-0.192	1.765	-0.62	-34.92
OG311	-0.192	1.765	-0.63	-35.53
OG311	-0.192	1.765	-0.649	-38.18
OG311	-0.192	1.765	-0.65	-38.21
OG311	-0.192	1.765	-0.67	-40.87
OG312	-0.12	1.75	-0.37	-10.87
OG2R5	-0.12	1.7	-0.17	-0.63
OG2R5	-0.12	1.7	-0.18	-1.04
OG3C5	-0.1	1.65	-0.31	-7.83
OG3C5	-0.1	1.65	-0.5	-24.6
OG3R6	-0.1	1.65	-0.26	-4.8
OG3R6	-0.1	1.65	-0.32	-8.54
OG3R6	-0.1	1.65	-0.37	-11.79
SG2D1	-0.565	2.05	-0.24	-3.97
SG2D1	-0.565	2.05	-0.27	-5.13
SG2R5	-0.45	2.0	-0.1	0.3
SG2R5	-0.45	2.0	-0.16	-0.99
SG2R5	-0.45	2.0	-0.25	-3.78
SG2R5	-0.45	2.0	0.01	1.99
SG2R5	-0.45	2.0	0.14	2.17
SG301	-0.38	1.975	-0.08	1.02
SG311	-0.45	2.0	-0.12	0.05
SG311	-0.45	2.0	-0.14	-0.54
SG311	-0.45	2.0	-0.15	-0.71
SG311	-0.45	2.0	-0.17	-1
SG311	-0.45	2.0	-0.18	-1.53
SG311	-0.45	2.0	-0.22	-2.81
SG311	-0.45	2.0	-0.23	-3.23
SG311	-0.45	2.0	-0.24	-3.67
SG3O2	-0.35	2.0	0.14	2.54
SG3O2	-0.35	2.0	0.22	2.01
SG3O2	-0.35	2.0	0.24	1.63
SG3O2	-0.35	2.0	0.33	0.31
SG3O2	-0.35	2.0	0.42	-2.13
SG3O2	-0.35	2.0	0.56	-6.76
SG3O2	-0.35	2.0	0.58	-7.92
SG3O2	-0.35	2.0	0.6	-8.73
SG3O2	-0.35	2.0	0.61	-9.08
SG3O2	-0.35	2.0	0.65	-11.12
SG3O3	-0.35	2.0	0.31	0.73

Table A.1 Table of data for each probe type.

Protein Type	PDB Name	Tag	Comments
14-3-3	1YZ5	A-B	14-3-3-σ
14-3-3	3P1N		
A2AR	3PWH		Adenosine A2A receptor
A2AR	3QAK		
A2AR	3VG9		
ACE	1O86		Angiotensin-converting enzyme
ACE	2IUL	mut	
AChE	1E66		Acetylcholinesterase
AChE	1EA5		
AChR	3UON		acetylcholine receptor
AChR	4DAJ	A-D	
ALR2	1ADS		aldose reductase
ALR2	1PWL		
ALR2	1PWM		
ALR2	1XGD		
APE-1	1BIX		
BRD	1EQF	A-B	Bromodomain
BRD	1F68	A	
BRD	1N72		
BRD	1X0J	A-C	
BRD	2D9E	A-T	
BRD	2DAT	A-T	
BRD	2DVQ	C	
BRD	2DVR	C	
BRD	2DVS	C	
BRD	2DWW		
BRD	2E3K	A	
BRD	2E7N	A	
BRD	2E7O	A	
BRD	2F6J	A-C	
BRD	2F6N	A-B	
BRD	2FSA	A-C	
BRD	2G4A	A,C,K,S-U	
BRD	2GRC		
BRD	2I7K		
BRD	2I8N	A	
BRD	2NXB	A-B	
BRD	2OO1	A-D	
BRD	2OSS		
BRD	2OUO		
BRD	2YW5	A	
BRD	2YYN	A-D	

Table A.2 Table

Protein Type	PDB Name	Tag	Comments
BRD	3AQA	B-C	bromodomains
BRD	3D7C	A-B	
BRD	3DAI		
BRD	3DWY	A-B	
BRD	3G0J	A-B	
BRD	3G0L		
BRD	3GG3	A-B	
BRD	3HME	A-B	
BRD	3HMF		
BRD	3HMH		
BRD	3I3J	A-L	
BRD	3IU5		
BRD	3JVL		
BRD	3JVM		
BRD	3K2J	A-B	
BRD	3LJW	A-B	
BRD	3LXJ		
BRD	3MB3		
BRD	3MQM	A-B	
BRD	3MXF	no ligand	
BRD	3NXB	A-B	
BRD	3O33	A-D	
BRD	3O37	A-D	
BRD	3ONI	no ligand	
BRD	3QZS	A-B no ligand	
BRD	3RCW	A-H	
BRD	3S91	no ligand	
BRD	3S92	no ligand	
BRD	3TLP	A-B	
BRD	3UV5	A-B	
BRD	4ALG	no ligand	
BRD	4CUP		
BRD	4CUQ		
BRD	4CUR		
BRD	4CUS		
BRD	4CUT		
BRD	4CUU		
BRD	4FLP	A-B no ligand	
BRD	4IOR		
BRD	4LC2		
BRD	4LDF	A-B	
BRD	4LYI		
BRD	4LZ2		

Protein Type	PDB Name	Tag
bromodomains	4NXJ	A-C
bromodomains	4PTB	A-B
bromodomains	4QEU	
bromodomains	4QZS	no ligand
bromodomains	4UIT	no ligand
bromodomains	4UIW	no ligand
bromodomains	5AME	A
bromodomains	5AMF	B
cAbl kinase	1IEP	A-B
cAbl kinase	1M52	A-B
cAbl kinase	2G2F	A-B
carbonic anhydrase 1	1HCB	
carbonic anhydrase 1	2CAB	
carbonic anhydrase 1	2NN7	A-B
carbonic anhydrase 2	1CA2	
carbonic anhydrase 2	2NNG	
carbonic anhydrase 2	4CAC	
carbonic anhydrase 2	5CAC	
caspase 1	1BMQ	
caspase 1	1SC1	A-B
cathepsin K	1MEM	
cathepsin K	1NLJ	A-B
cathepsin K	4LEG	
cathepsin S	1NPZ	A-B
cathepsin S	4P6G	A-D
Cyclin-dependent kinase 2 (CDK2)	1E1X	
CDK2	1H07	
CDK2	4EK3	
cell division ZipA	1F46	A-B
cell division ZipA	1F47	
cell division ZipA	1S1S	A-B
cyclooxygenase 2	4COX	A-D
cyclooxygenase 2	4COX	heme
cyclooxygenase 2	5COX	A-D
DNA gyrase B	1KIJ	
DNA gyrase B	1KIJ	A-B
DNA gyrase B	1KZN	
EGRF kinase	1M14	
EGRF kinase	1M17	
enoyl reductase	1C14	A-B
enoyl reductase	1DFI	A-D
factor Xa	1EZQ	
factor Xa	1EZQ	A
factor Xa	1LPZ	

Protein Type	PDB Name	Tag
fungal Cyp51	1E9X	
fungal Cyp51	1EA1	
fungal Cyp51	3JUV	
HIV-1 protease	1DMP	
HIV-1 protease	1ODW	
HIV-1 protease	3PHV	
HIV-1 protease	3PHV	A
HIV integrase	1BI4	
HIV integrase	1BI4	A-C
HIV integrase	1QS4	
HIV integrase	1QS4	A-C
HIV integrase	3L3U	
HIV integrase	3L3U	A-B
HIV Reverse Transcriptase (RT)	1T03	
HIV RT	1T03	1
HIV RT	1T03	2
HIV RT	1T05	
HIV RT	1T05	1
HIV RT	1T05	2
HIV RT	3DLK	
HIV RT	3DLK	1
HIV RT	3DLK	2
HIV RT	4G1Q	
HIV RT	4G1Q	1
HIV RT	4G1Q	2
HMG CoA reductase	1HW8	A-D
HMG CoA reductase	1HW8	AB
HVC serine proteinase	4KTC	A
HVC serine proteinase	4KTC	C
IMP dehydroginase (DH)	1NF7	A-B
inStem	1JNX	
inStem	1T15 pS	
inStem	1T15 pT	
inStem	2AZM A pS	
inStem	2AZM A pT	
inStem	3AL3 pS	
inStem	3AL3 pS	chopped
inStem	3AL3 pT	
inStem	3AL3 pT	chopped
inStem	3K0K pS	
inStem	3K0K pT	
kinase	1HCK	
kinase	1HCL	

Protein Type	PDB Name	Tag
KRas	2PMX	A
KRas	3GFT	
KRas	4DST	
Mouse double minute 2 homolog (MDM2)	1RV1	A-C
MDM2	1Z1M	
neuraminidase	1IVG	A-D
neuraminidase	2HU4	A-D
P38 kinase	1KV1	
P38 kinase	1P38	
P38 kinase	1WFC	
phosphodiesterase (PDE) 4D	1OYN	A-D
PDE 4D	3SL3	A-D
PDE 5A	1T9R	A-B
PDE 5A	1T9S	
PDE 5A	1UDT	
penicilin	1HVB	
penicilin	1QME	
penicilin	1QMF	
penicilin	3PTE	
penicilin	4BLM	A-B
Phaedon	NC	A-B
Phaedon	U	A-B
Protein tyrosine (PT) phosphatase 1B	1G1F	
PT phosphatase 1B	1JF7	A-B
PT phosphatase 1B	1NNY	
PT phosphatase 1B	1ONZ	
PT phosphatase 1B	1PTY	
PT phosphatase 1B	1Q1M	
PT phosphatase 1B	3A5J	
serum albumin	1AO6	A-B
serum albumin	1BM0	A-B
serum albumin	1E78	A1
serum albumin	1E78	A2
serum albumin	1E78	B1
serum albumin	1E78	B2
Methylcytosine dioxygenase (TET2)	4NM6	
TET2	5D9Y	
TET2	5DEU	
thrombin	1KTS	
thrombin	1MKX	HL
thrombin	1XMN	AB
thrombin	1XMN	CD
thrombin	2UUF	

Protein Type	PDB Name	Tag
tyrosine kinase	1FMK	
tyrosine kinase	1Y57	
tyrosine kinase	2SRC	
urokinase	2NWN	
urokinase	4FU7	
Tuberculosis zinc metalloprotease (XTAL)	3ZUK	
Human aldehyde dehydrogenase (XTAL)	5FHZ	
Human aldehyde dehydrogenase (XTAL)	5FHZ	1

Appendix B

Appendix 2: Simulation Protocols

B.1 Chapter 3: K-D Trees

B.1.1 K-D Trees

Previous studies using IFST or GIST have used a grid of voxels, or the cut-cell approach to find nearest neighbour distances [48, 49]. For the pair nearest neighbour distances required for the KNN estimators this method was not practical. In order to increase efficiency and allow the calculation of the pair terms, a K-Dimensional (K-D) Tree method was used. The tree can be used find the nearest neighbours of N different K dimensional vectors in $\mathcal{O}(KN\ln N)$ time [157]. For the neighbours of individual solvent atoms used in the conditional one particle entropy estimator $K = 3$. For pairs of solvent atoms used in the conditional two particle entropy estimator, $K = 6$. The memory associated with the construction of such a tree grows as $\mathcal{O}(N^2)$. This memory scaling is appropriate for the $K = 3$ case, however became unfavourable for the $K = 6$ case. In light of this a compromise method was developed.

The simulation cell is split into $v \times v \times v$ large cubic sub-volumes, and then each pair of sub-volumes is taken, a tree is constructed for the members of the sub-volumes and the neighbours found. This is illustrated in figure B.1. In practice a larger set of particles a distance of $L/(2V) + \delta$ away from the centre of each sub-volume is taken, where δ is chosen appropriately based on the current density of frames, an example of this is shown in Fig. B.2.

When searching for neighbours in the tree, a strict system is used to determine what counts as a nearest neighbour to give the distances $d_{ij,k}$ and $d_{i_1 i_2 j,k}$ in the KNN estimators (Eqns. 43 and 44 in the main text). The nearest neighbour to an atom i in frame j, is the closest atom k in frame l where $j \neq l$. An atom and its nearest neighbour for the density estimator cannot be in the same molecular dynamics frame, otherwise the presence of original atom will alter the density at that point. This set of rules made the tree process slightly slower than a general neighbour search as an additional check of frame number had to be made for each node in the tree.

The 6-vectors used in the conditional two particle entropy calculations takes two atoms with position 3-vectors (x_1, y_1, z_1) and (x_2, y_2, z_2) respectively and forms the 6-vector $(x_1, y_1, z_1, x_2, y_2, z_2)$ for all $N(N-1)$ combinations. A cutoff of $\delta = 0.8$ was used for $v = 6$, (a $6 \times 6 \times 6$ memory discritization). The value of δ was found by inspecting figure B.2. and finding it was highly unlikely to find a nearest neighbouring pair further than 0.8 Å away for 5000 frames. The distance metric used between two 6 vectors was the Euclidean metric.

B.1.2 Plotting as a Function of Space

Because the entropic contributions are summed based on the neighbours of given atoms, if these atoms' locations are known, the contribution can be associated with a point in space. With enough frames a histogram can be built up of all of the contributions binned into voxels of an appropriate size. 2 particle conditional entropy contributions in the half cell. Red indicates the greater than bulk entropy contributions corresponding to the peaks in the RDF, blue indicates the smaller than bulk contributions corresponding to the dips in the RDF. The excluded volume of the solute makes up most of the entropy, with the troughs in the RDF corresponding to the radial shells which become less coherent with distance to the solute. At the edges of the box, bulk like pockets overtake the shell like structures consistent with the solvation shell model [158]. If the solute had been considerably anisotropic this method would still have produced an estimate and a surface plot would highlight key areas where solvent-solvent reconstruction effects contribute to the entropy.

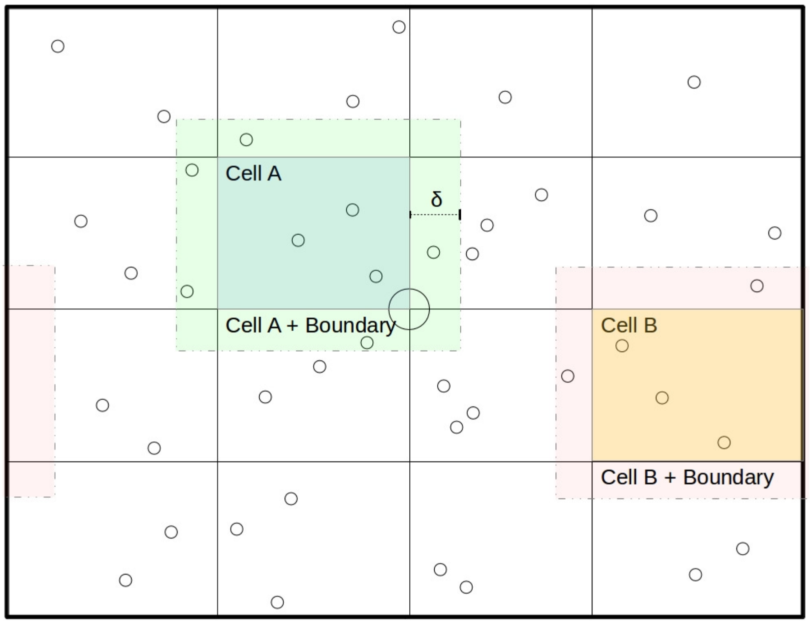


Fig. B.1 Schematic for the conditional two particle entropy tree method. The simulation box is split into cubic voxels. Every pair of these voxels are compared, for example Cell A and Cell B. Create 'List A' of all possible 6-vectors from the atoms in Cell A (blue) and Cell B (orange). A 6-D tree is made of all possible 6-vectors from the atoms in Cell A+δ (green) and Cell B+δ (red). For every vector in 'List A', the closest, different 6-vector in the tree is found, where the simulation frames of the Cell A atoms are the same and the frame of the Cell B atoms are the same. This is reported as the nearest pair. Periodic boundary conditions are used for the δ region.

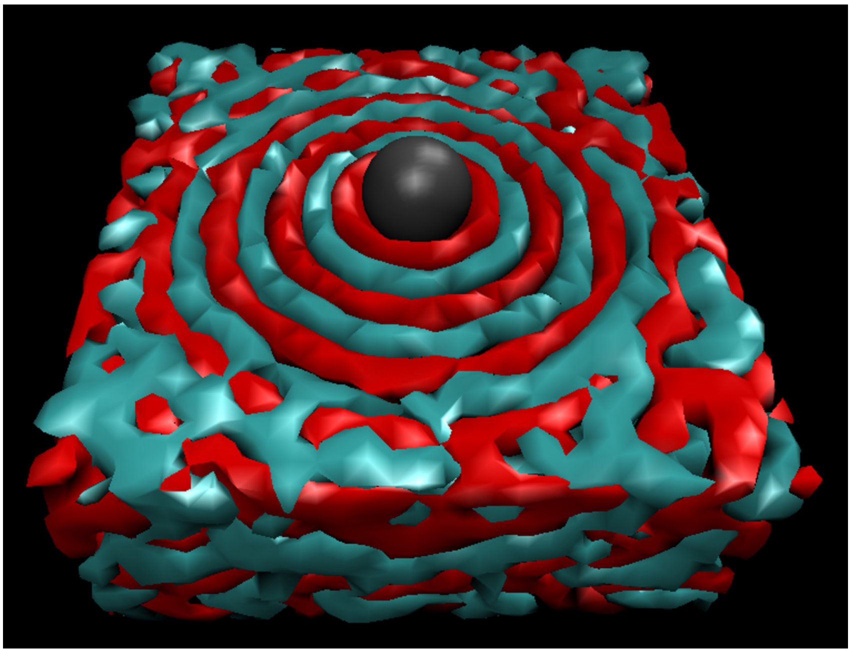


Fig. B.3 A surface plot of a 3-dimensional histogram of the conditional two particle entropy for half of the SOL4 solute simulation box. This plot contains the largest 10% of the entropies associated to each bin, and the pronounced shell like features coincide with the dips (blue) and peaks (red) of the radial distribution function shown in the main text.

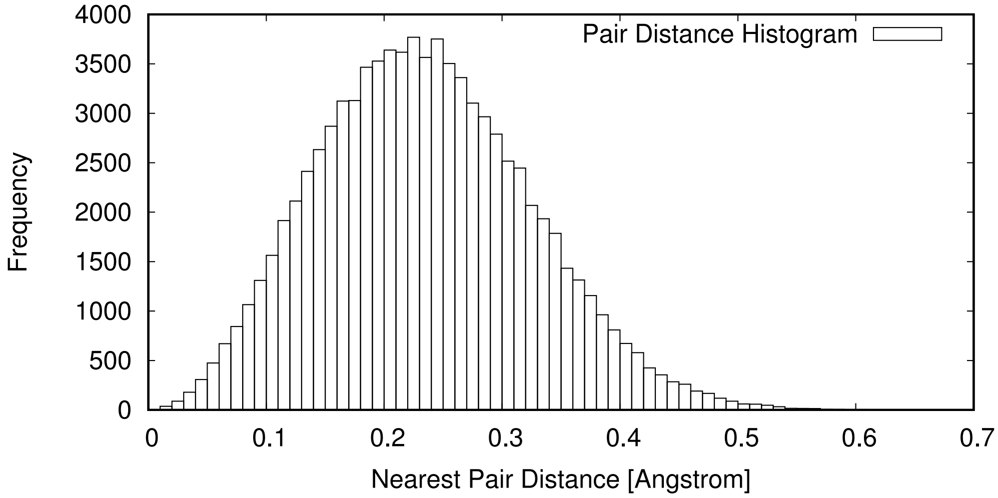


Fig. B.2 A histogram of the frequency of nearest pair distances for 1000 frames with the simulation cell being cut into 6 × 6 × 6 voxels. From this a δ parameter of 0.8 Å was decided. The smaller the δ parameter can be made, the quicker the code will run, however, if the right tail of this figure becomes truncated then the code is approximating the conditional two particle entropy.

B.2 Chapter 4: MD Simulations for the AFEP Results

B.2.1 Caffeine Simulations

Lambda Schedule

Caffeine was run with the CHARMM36 forcefield in NAMD excluding scaled 1-4 interactions with scaling constant 1.0, PME electrostatics with a grid spacing of 1 Å and a dielectric constant of 1.0. The TIP3P water model was used. Switching was used from 9 Å till the cut-off of 11 Å. The pair list distance was 12.5 Å and 20 steps per cycle were used.

The simulation box for the start of the NpT equilibration was a rhombic dodecahedron with cell vectors (31 0 0),(0 31 0),(15.5 15.5 21.9203) and origin (0,0,0).

A .dcd frame was saved every 2500 steps. Each timestep was 2fs; nonbonded terms were re-evaluated every step and electrostatic terms every 2 steps. All bonds were rigid. Langevin temperature control was used with a damping constant of 1 and a target temperature of 300 K.

For the NpT equilibration, Langevin piston control was used with a target pressure of 1.01325 bar, and piston period decay and temperature of 100, 50 and 300 respectively.

λ-window	start	end	λ-window	start	end
1	0	0.01	17	0.19	0.21
2	0.01	0.02	18	0.21	0.22
3	0.02	0.03	19	0.22	0.23
4	0.03	0.04	20	0.23	0.24
5	0.04	0.05	21	0.24	0.26
6	0.05	0.06	22	0.26	0.28
7	0.06	0.07	23	0.28	0.3
8	0.07	0.08	24	0.3	0.32
9	0.08	0.09	25	0.32	0.38
10	0.09	0.1	26	0.38	0.48
11	0.1	0.115	27	0.48	0.6
12	0.115	0.13	28	0.6	0.8
13	0.13	0.145	29	0.8	0.89
14	0.145	0.16	30	0.89	0.95
15	0.16	0.175	31	0.95	0.975
16	0.175	0.19	32	0.975	1

Table B.1 The 32-window λ schedule used for the caffeine FEP simulations.

The energy of the system for the BAR FEP calculation was output every 500 steps. An alchemical softening constant of 5 was used for the Lennard-Jones interactions. Each FEP window was run for 10^6 steps of which 10^5 were equilibration steps. There were 32 λ-windows in the whole simulation shown in table B.1. To vary the electrostatics and LJ terms seperately in one FEP simulation the NAMD controls alchElecLambdaStart=0.6 and alchVdwLambdaEnd=1.0 were set.

B.2.2 Simulations of HIV-1P

Binding and Unbinding the Lennard-Jones Terms

The binding and unbinding FEP simulations were run with the CHARMM36 forcefield in NAMD excluding scaled 1-4 interactions with scaling constant 1.0, PME electrostatics with a grid spacing of 1 Å and a dielectric constant of 1.0. The TIP3P water model was used. Switching was used from 9 Å till the cutoff of 11 Å. The pair list distance was 12.5 Å and 20 steps per cycle were used.

The simulation box for the start of the NpT equilibration was a rhombic dodecahedron with cell vectors (91,0,0),(0,91,0) and (45.5,45.5,64.347) and origin (0,0,0). For the NpT

λ-window LJ	start	end	λ-window LJ	start	end	λ-window Electrostatic	start	end
1	0.00	0.01	17	0.16	0.18	1	0.0000	0.0625
2	0.01	0.02	18	0.18	0.20	2	0.0625	0.1250
3	0.02	0.03	19	0.20	0.22	3	0.1250	0.1875
4	0.03	0.04	20	0.22	0.24	4	0.1875	0.2500
5	0.04	0.05	21	0.24	0.28	5	0.2500	0.3125
6	0.05	0.06	22	0.28	0.32	6	0.3125	0.3750
7	0.06	0.07	23	0.32	0.36	7	0.3750	0.4375
8	0.07	0.08	24	0.36	0.40	8	0.4375	0.5000
9	0.08	0.09	25	0.40	0.44	9	0.5000	0.5625
10	0.09	0.10	26	0.44	0.52	10	0.5625	0.6250
11	0.10	0.11	27	0.52	0.60	11	0.6250	0.6875
12	0.11	0.12	28	0.60	0.68	12	0.6875	0.7500
13	0.12	0.13	29	0.68	0.76	13	0.7500	0.8125
14	0.13	0.14	30	0.76	0.84	14	0.8125	0.8750
15	0.14	0.15	31	0.84	0.92	15	0.8750	0.9375
16	0.15	0.16	32	0.92	1.00	16	0.9375	1.0000

Table B.2 Table of the 32-window λ schedule for the binding and unbinding of the Lennard-Jones (LJ) interactions. The 16-window λ schedule for the swapping of charges between the ligands.

equilibration, Langevin piston control was used with a target pressure of 1.01325 bar, and piston period, decay and temperature of 100, 50 and 300 respectively.

A .dcd frame was saved every 50 steps. Each timestep was 2 fs, nonbonded terms were re-evaluated every step and electrostatic terms every 2 steps. All simulated bonds were rigid.

Langevin temperature control was used with a damping constant of 1 and a target temperature of 300 K. The energy of the system for the BAR FEP calculation was output every 250 steps. An alchmical softening constant of 5 was used for the Lennard-Jones interactions. Each FEP window was run for 50000 steps of which 10000 were equilibration steps. The lambda schedule had 32 windows as shown in table B.2.

Swapping Charges

For the charge swapping stage the lambda schedule had 16 windows as shown in table B.2. All other parameters were the same as the turning on and off of the Lennard-Jones terms, as described above.

B.3 Chapter 5: MD Simulations of the GABA$_A$ Receptor and GABARAP

In the simulation of the intracellular helices of the GABA$_A$ receptor we first selected atoms from the following amino acids: α1-subunit Lys 391–Leu 422, β2-subunit His 421–Ile 449 and γ2-subunit Asp 413–Ser 443. The helices were placed in a periodic box with at least 10 Å between the protein and its image. The system consisted of 19708 water molecules, 56 K^+ ions and 73 Cl^- ions to achieve a [KCl] of 0.15 mM. The system comprised a total of 61857 atoms.

The system was minimised for 10000 steps with all the protein atoms frozen. Molecular dynamics was initialised for 10000 time-steps of 0.1 fs each, with all main-chain nitrogen atoms frozen. Langevin dynamics was applied; the thermostat was set with a time constant of 1 ps^{-1}, and the barostat set with a piston decay time of 10 ps and a piston period of 20 ps. The van der Waals cut-off was 12 Å, and Ewald summation was used for long-range electrostatics. The time-step was lengthened to 2 fs over 30000 time-steps, while all main-chain nitrogen atoms were frozen. A 2-ns equilibration was carried out on the initialised system. A data collection simulation was then carried out for 5 ns, again with all main-chain nitrogen atoms fixed. Configurations were output every 0.5 ps. We obtained a total of 10000 configurations of the intracellular helices of the GABA$_A$ receptor.

For the simulation of GABARAP, we chose model 3 of 1KOT and the 1GNU structure (AAA) as the starting structures. The 1KOT structure of 117 amino acids was placed in a periodic box with at least 10 Å between the protein and its image; 9161 water molecules, 24 K^+ ions and 26 Cl^- ions were placed in this box. The system consisted of a total of 29508 atoms. The 1GNU structure of 117 amino acids was placed in a periodic box with at least 10 Å between the protein and its image; 9115 water molecules, 25 K^+ ions and 27 Cl^- ions were placed in this box. The system consisted of a total of 29372 atoms.

These systems were minimised for 10000 steps with all main-chain nitrogen atoms frozen. Langevin dynamics was applied; the thermostat was set with a time constant of 1 ps^{-1}, and the barostat set with a piston decay time of 1 ps and a piston period of 2 ps. The van der Waals cut-off was 12 Å, and Ewald summation was used for long-range electrostatics. The time-step was lengthened to 2 fs over 40000 time-steps. The system was then equilibrated for 2 ns. Data collection was carried out for 5 ns, again with all main-chain nitrogen atoms frozen, with configurations output every 0.5 ps. We obtained a total of 10000 configurations for each model of the hydrated GABARAP.

Appendix C

Appendix 3: Additional Plots

C.1 Chapter 6: Additional Plots of Hydration Site Data

This section contains some additional plots which it was felt would not fit in the main text without interrupting the flow of the discussion. In general they provide extra evidence for the broadening and leftward shift of free energy distributions with increasing number of hydrogen atom neighbours to a hydration site. They also show an increased number of very displaceable sites with increasing neighbours.

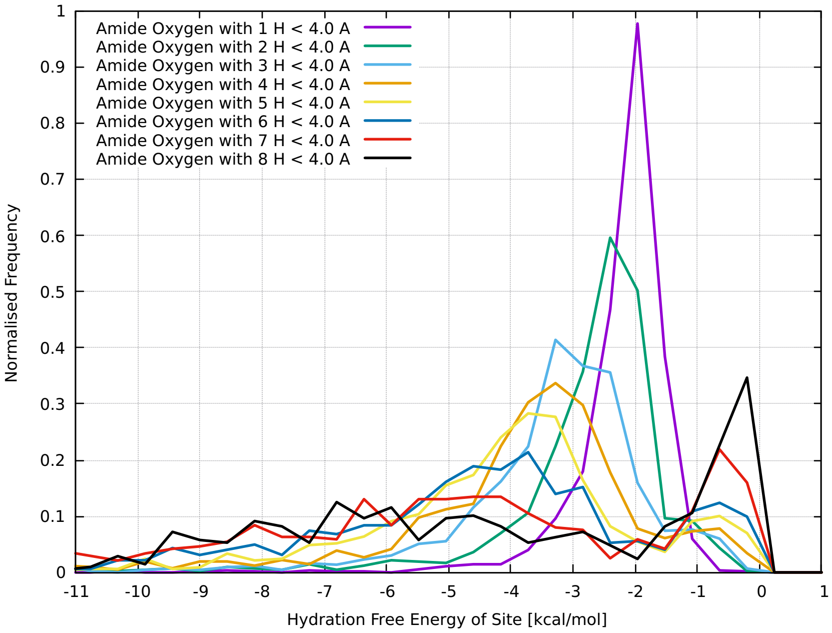


Fig. C.1 Histograms of the hydration free energy for sites near an amide oxygen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.

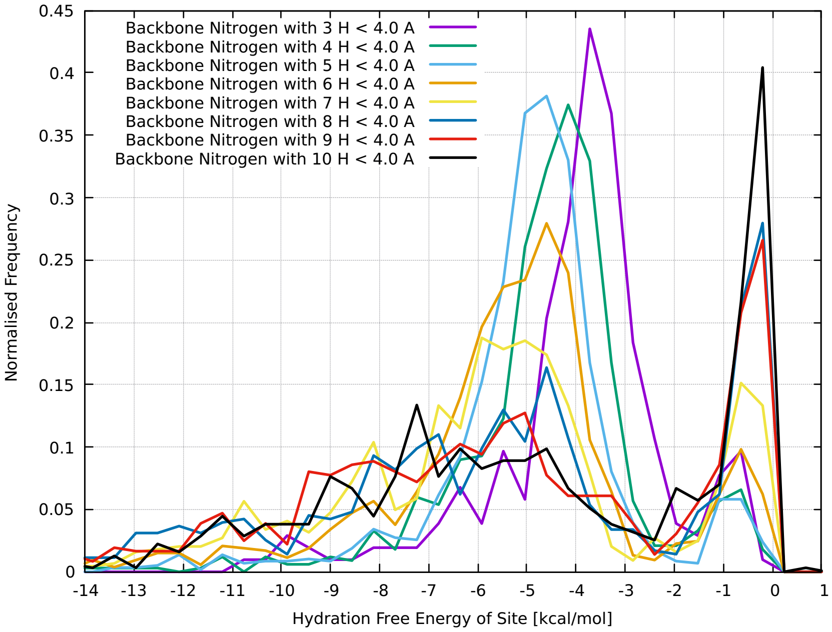


Fig. C.2 Histograms of the hydration free energy for sites near a backbone nitrogen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.

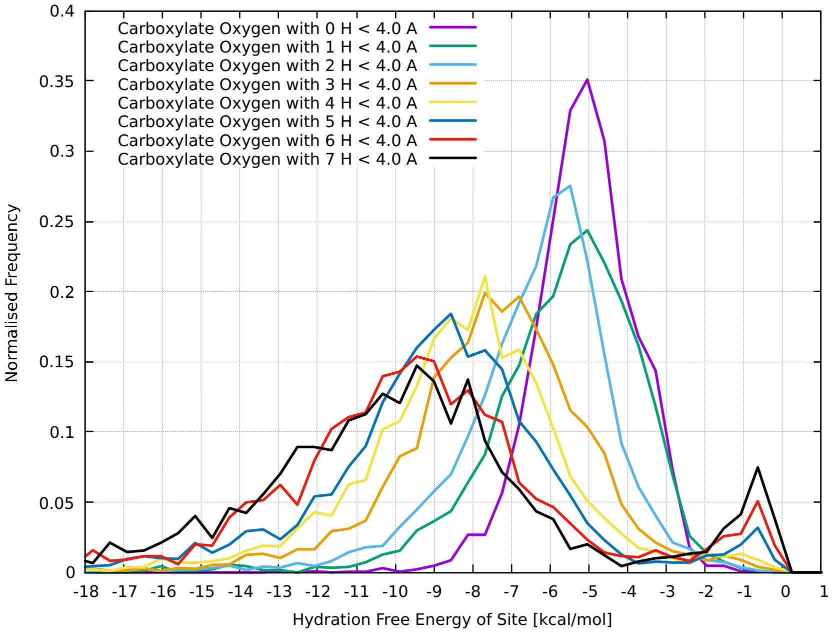


Fig. C.3 Histograms of the hydration free energy for sites near a carboxylate oxygen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.

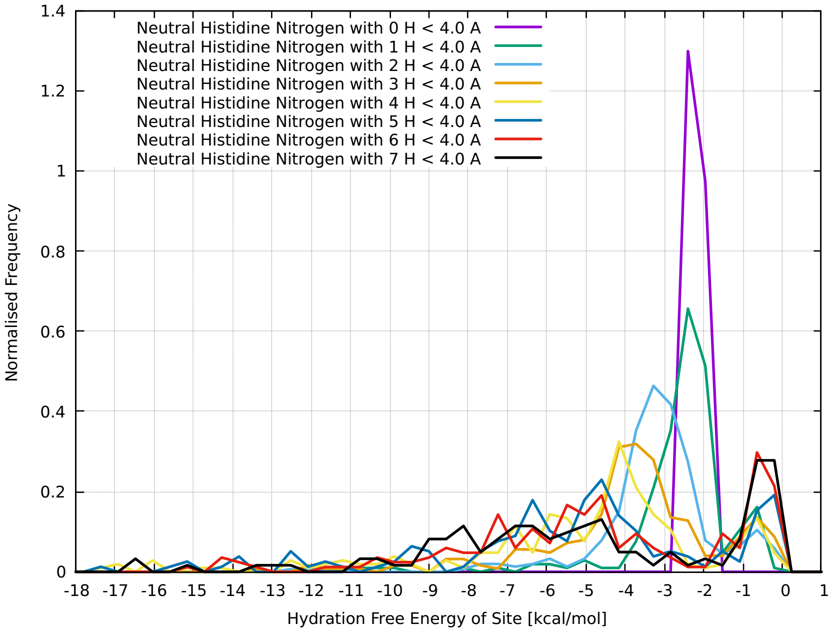


Fig. C.4 Histograms of the hydration free energy for sites near a histidine nitrogen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.

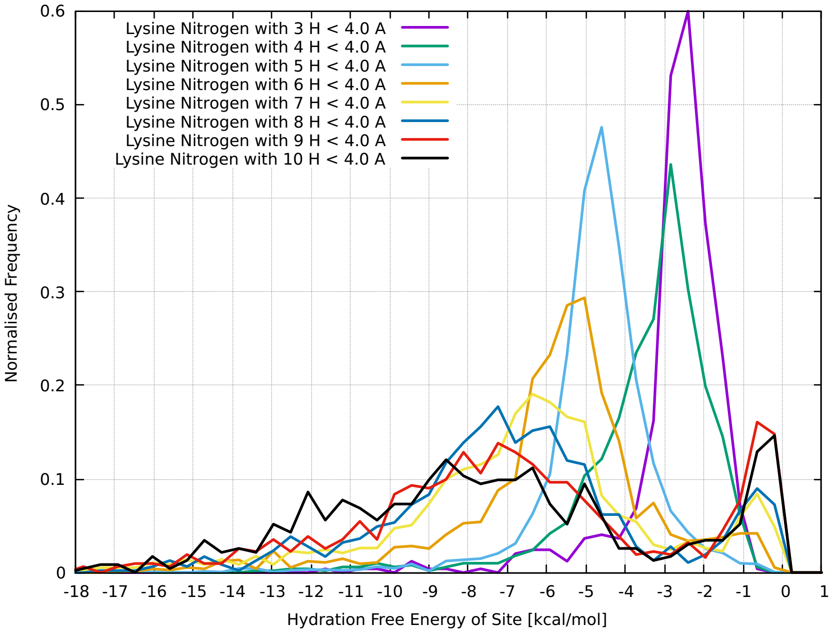


Fig. C.5 Histograms of the hydration free energy for sites near a lysine nitrogen with varying numbers of hydrogen atoms within 4.0 Å. The curves show broadening with increased numbers of hydrogen.