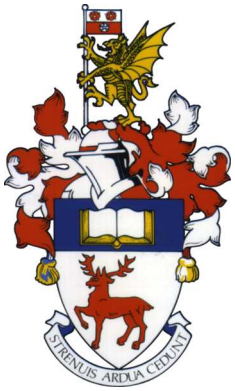
Development and application of free energy methods



Patrick Schöpf

Department of Chemical Biology

University of Southampton

A thesis submitted for the degree of

Philosophical Doctor (PhD)

2011 October

1. Reviewer: C. Edge

2. Reviewer: S. Khalid

Day of the defense: XX.XX.2011

Signature from head of PhD committee:

Abstract

The development of free energy simulation protocols for calculating relative binding free energies of ligands is presented in this thesis. To this end, the protein Dihydroorotate Dehydrogenase (DHODH), complexed to a highly congeneric series of compounds that show ambiguities in their binding modes, was studied in detail. To estimate the systematic error in force fields, relative free energies of hydration have been calculated using Replica-exchange Thermodynamic Integration (RETI) for sets of force field parameters and atomic partial charges in a classical molecular mechanics environment as well as a novel hybrid molecular mechanics/quantum mechanics model. The results demonstrated that all force fields and methods employed yield similar estimates of the relative free energies, while GAFF and OPLS-AA in conjunction with AM1BCC and AM1CM1A charges, respectively, performed best. To balance accuracy and ease of generating parameters, GAFF in conjunction with AM1BCC charges was selected to be the most valuable for describing the inhibitors in DHODH. To rigorously assess the thermodynamic end states for the ligands, crystal hydrates present in the binding site of DHODH have been investigated using the Just-Add-Water-Molecules (JAWS) algorithm, Grand-canonical Monte Carlo (GCMC) simulations and the double-decoupling approach (DDM). These findings clearly suggested a change in hydration networks for both the inhibitors and their different binding modes, while all three approaches essentially yield identical results. This allowed us to construct free energy cycles using the single and dual topology approach in order to calculate the free energies of binding of the ligands as well as the stability of their binding modes. The results obtained were precise within the error of the methods, but not accurate, and allowed to complement the crystallographic findings.

Acknowledgements

Looking back, I am surprised and at the same time very grateful for all I have received and all experiences I could make throughout the years of my PhD. It has certainly shaped me as a person. Pursuing a PhD project is both a painful and enjoyable experience and it would not have been possible without the help of many people.

I would like to gratefully and sincerely thank my supervisor Jon for his guidance, understanding, patience, and most importantly, his honesty during my time at Southampton University. His mentor-ship was paramount in providing a well rounded experience. He encouraged me not only to grow as a computational chemist but also as an instructor and an independent thinker and I am not sure many PhD students are given the opportunity to develop their own individuality and self-sufficiency by being allowed to work with such independence.

I thank my industrial supervisors Mike King and Richard Taylor for their enthusiasm and understanding, and the CompChem group at Celltech, Slough, for making my placement an unforgettable and rewarding experience, as well as for providing me with the necessary funding.

I cannot quantify the impact of Julien Michel not only on this project, but also on helping me to settle and integrate at Southampton University. I am deeply grateful for your input and friendship.

It is the nature of a Ph.D. project to exchange ideas in order to gain deeper understanding. Therefore I would like to give my special thanks to Christopher Woods not only for his helpful discussions on Sire but also for being a friend, and Michael Bodnarchuk, the master of water molecules, without whom some important details in this thesis would not have been discovered.

In my daily work I have been blessed with a friendly and cheerful group of people. Here, I would like to mention Nadia, George, Mario, Barbara and Mike.

When science takes over, good friends help you to escape for a little while. I am thankful for memorable moments spent with Amaury, Francesco, Nikos, Nico, Marie, Alexia, Dan, Nui, H, Fhiona and Rob.

Work such as this cannot be performed without adequate software. Today, we are fortunate enough to be experiencing a revolution in the way software is developed and distributed. This Open Source movement means that enterprise class software is available for anyone to use, develop, modify and learn from. This is all free from financial cost or intellectual restrictions. This research almost exclusively used open source software, and, with a few exceptions, this thesis has been written entirely using open source tools. Thank you Open Source.

Finally, and most importantly, I would like to thank my partner. Your support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past twelve years of my life have been built. Thank you so much! Finally, I would like to thank my parents, Irmgard and Lambert, and my little sister, Carina, for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on.

Contents

1

Introduction

The modern era of the pharmaceutical industry - of isolation and purification of compounds, chemical synthesis, and computer-aided drug design - has evolved after intuition and *trial and error*, led humans to believe that plants, animals and minerals contained medicinal properties[1]. The unification of research in the 20th century in fields such as chemistry and physiology increased the understanding of basic drug-discovery processes and led to a more rational approach that is outlined schematically in figure 1.1.

Structure-based drug design (SBDD) was born out of this development and became an essential component of modern drug discovery, where three-dimensional target structures are being exploited to design tightly binding small molecules to modulate their function, thus SBDD is exploring the microscopic world of a stereotypical biochemical process thought to cause a pathophysiological state[2]. Figure 1.2 gives an impression on how a SBDD approach might look. How molecules interact non-covalently with each other, i.e. molecular recognition, is of utmost importance for understanding the roles played by individual components in these processes[3].

Ligand binding affinities are determined via the equilibrium constant K_{eq} in bioassays, which are both expensive and time consuming[4,5], hence the efficient and accurate computation of binding affinity is one of the major challenges in SBDD and the main objective of this thesis. The driving force for a ligand binding to a macromolecule is the free energy of binding, which is related to K_{eq}. The physics of this process can, in principle, be understood at a microscopic level with statistical thermodynamics which defines free energy unambiguously. A subtle balance between

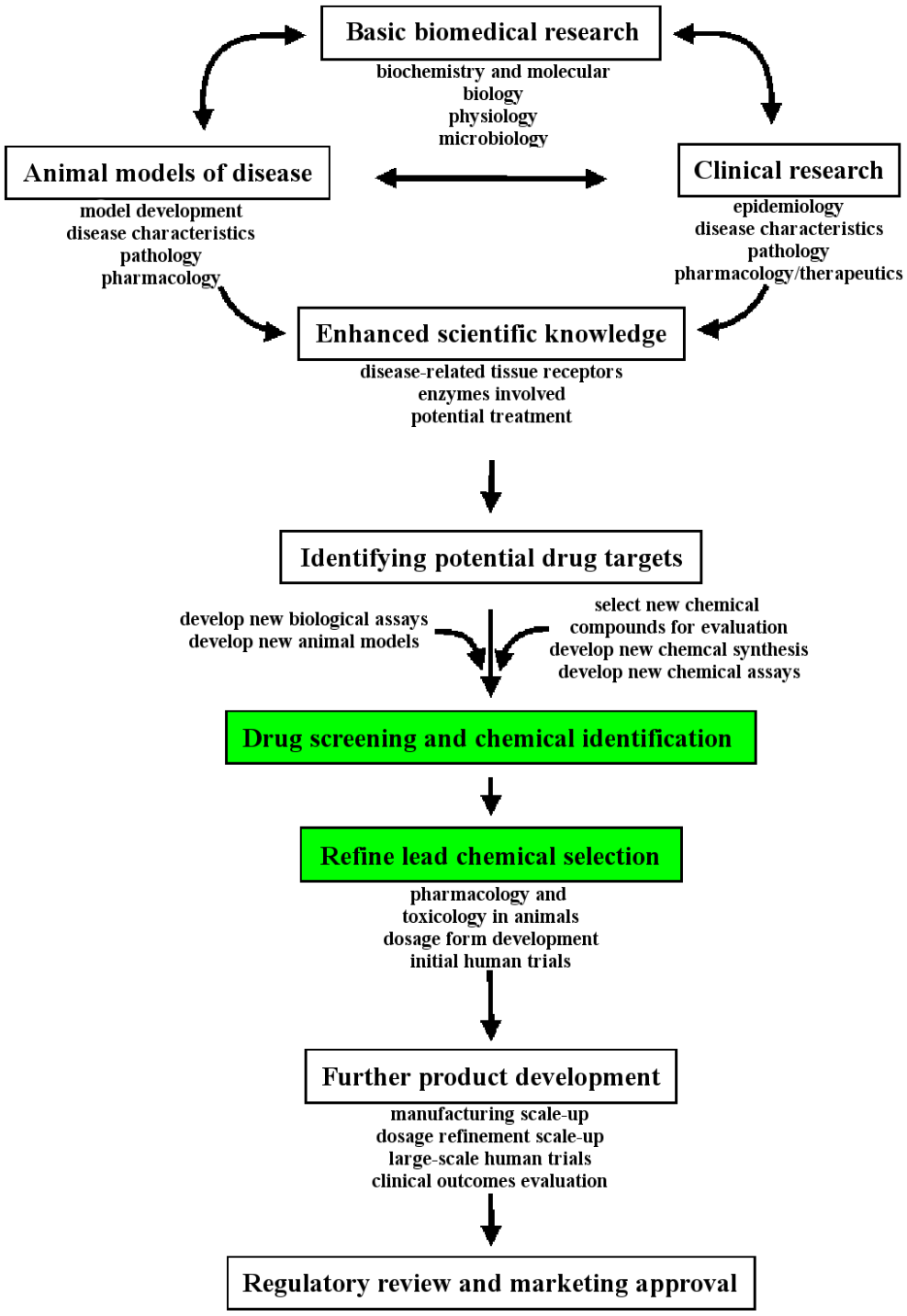


Figure 1.1: The modern drug design process requires the consultation of many scientific disciplines and approaches in order to escape the *trial and error* concept. This figure has only illustrative character and many more approaches can be thought of depending on therapeutic indication. The figure has been adapted from [1] and was created using DIA version 0-97-1. The green boxes indicate steps where structure-based drug design is most likely part of the key to success.

energetic and entropic effects, governed by direct ligand protein interactions, loss of conformational, orientational and translational degrees of freedom, and desolvation effects of both ligand and its biological counterpart make up this quantity. The resulting equations that govern the free energy of binding are complex and their solutions can be estimated numerically via computer simulations[6]. They rely on extensive conformational sampling of the relevant degrees of freedom, needed to describe in particular the entropy correctly, and an appropriate description of the energetics of the system in question. Chapter 2, *Biomolecular simulation in a nutshell*, covers the underlying theories to the methods that are called rigorous free energy methods and that are used in this thesis.

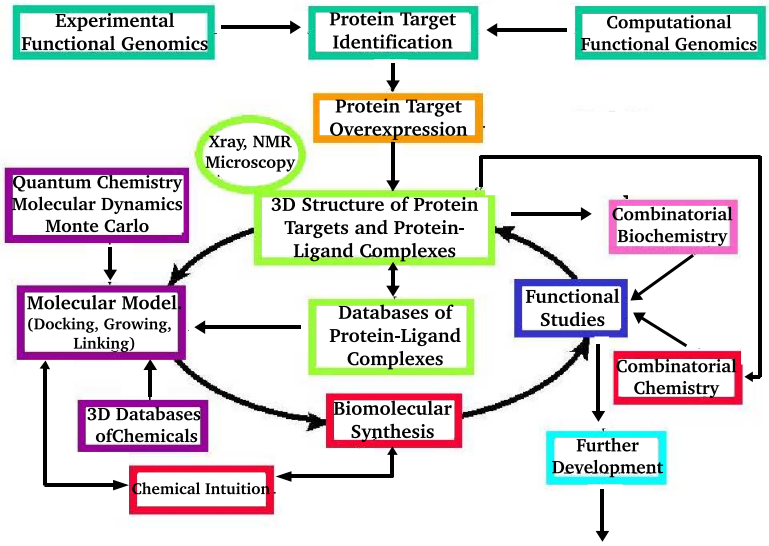


Figure 1.2: SBDD and its many functional roles in the drug design process. Within this process many different scientific disciplines have to meet and complement each other to maximize and rationalize the design of the most beneficial compound that is likely to have a maximum effect in vivo. The figure has only illustrative character and has been adapted from the group of Prof. Wim G.J. Hol at the University of Washington, Washington, USA.

Free energy is a *state function*[7], meaning the same estimate of a free energy difference can be obtained through different paths. This is commonly used in drug design projects by constructing a thermodynamic cycle, in which the free energies of interest are obtained by calculation of different, more readily available free energy

differences[8,9]. This general concept is demonstrated in Figure 1.3 for relative binding free energies, i.e. the free energy difference of binding of two different molecules. The two different molecules, S1 and S2, are complexed to a protein or free in solution. The specific manner in which a molecule is complexed - including potential water molecules - is called end state. Rigorous free energy methods provide us with a reliable estimate for the free energy difference between end states, assuming we have chosen a description of energetics appropriately and have met all configurations necessary to obtain an appropriate ensemble[6]. Therefore, the definition of end states is inevitably linked to the outcome of a free energy study.

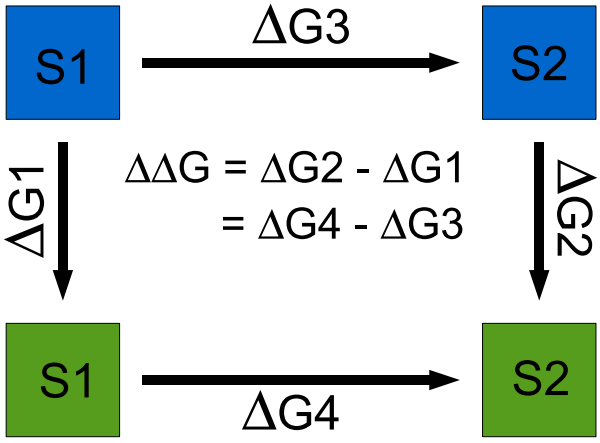


Figure 1.3: Thermodynamic cycle to calculate the relative free energy of binding. The cycle relates the difference in free energy between S1 and S2 in two media, indicated by the green and blue squares surrounding S1 and S2 in the figure. S1 and S2 could be two small molecules in two media, for instance blue for water and green for solvated protein, in which case the double free energy difference will correspond to the relative binding free energy of S2 with respect to S1. While the vertical processes, corresponding to ΔG_1 or ΔG_2, are often measured experimentally, the horizontal processes, corresponding to ΔG_3 or ΔG_4, are usually easier to calculate in a computer simulation.

Although the theoretical foundations for rigorous free energy calculations have been laid out as early as 1935 by J. Kirkwood, i.e. *Thermodynamic Integration (TI)* method[10], and 1954 by Robert W. Zwanzig, i.e. *Free Energy Perturbation (FEP)*[11] method, the idea of predicting binding free energies from computer simulations originates from the mid 1980s and was subsequently further developed[12,13]. Next to these formally rigorous methods, approaches have been proposed to compute free energies of binding that incorporate approximations and thus reduce computational cost,

i.e. approximate free energy methods such as the Linear Interaction Energy (LIE) method[14] and Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA)[15], to mention just a few.

However, rigorous free energy calculations are considered too demanding to apply to more than a handful of compounds, as calculations typically require a lot of computational power. They have also drawn criticism due to their difficult and time-consuming setup and the analysis of results obtained[16]. Scoring functions[17], rooted on empirical foundations, are a common practice solution and are routinely used to screen large datasets. Although often considered rather inaccurate and crude, they satisfy by generating a first rough dataset that in turn will be validated, refined and eventually bio-assayed. Scoring results usually lack information on system dynamics, which can be an integral part in ligand design. Free energy simulations provide, apart from their potentially much more reliable binding estimate, detailed information on system dynamics, as properties are being averaged while the system is sampled.

Accurate binding free energies and detailed understanding of system dynamics could answer many questions in SBDD: lead optimizations where small modifications to a given scaffold structure are attempted in order to maximize ligand binding affinity and drug-like properties; selectivity profiling for a compound targeting a series of related proteins; identifying targets for a compound with an unknown mode of action; reliably identifying the binding mode for a compound that has not been experimentally determined or where attempts to crystallise the complex failed[16].

The last years of rigorous free energy methods have widened and extensively validated free energy calculations by providing detailed information on dynamics and molecular recognition:

- Relative and absolute free energies have been calculated for different protein-ligand systems[18,19,20]; proteins have been investigated that pose mayor challenges towards free energy methods, i.e. conformational changes upon ligand binding, multiple binding modes of inhibitors[21].

- Relative and absolute free energies of solvation have been calculated on large datasets to assess not only the performance of currently used Force Fields but

also highlighted the importance of charge parametrization[22,23,24,25,26,27], investigations on hydration shell structures created by free energy simulation[28], and entropic effects caused by solute shape[29], to mention just a few.

- Hybrid models incorporating higher theory levels have been broadly published, as they complement current force field deficiencies, i.e. polarization, and provide ad-hoc parametrization added to the force field energies[30,31].

- Novel perturbation schemes that try to break the connection with topological dependencies, hence allow perturbations with greatly different topology, such as the dual topology paradigm[32]or the one-step perturbation approach[33].

- Novel sampling schemes have been developed to overcome sampling problems within a finite time frame, and that have been applied on protein conformational change[34] as well as inhibitors that are not sufficiently sampled due to high intra-molecular energy barriers[35].

- Approximate free energy methods have been adapted to answer yet more complex questions, a trend similar to the one observed in rigorous free energy techniques[36,37,38,39].

The work done in our lab has focused on widening the applicability of free energy methods and assessing their value on key questions in SBDD. Developing novel Hamiltonian replica-exchange methods[40], implications of system representation[41], considerations on charge[42] as well as quality assurance[43] when results are being analysed, were focused initially.

A more recent study performed in our lab[18] has compared the ability of free energy simulations and empirical scoring functions to rank-order by potency 10 inhibitors of Cyclooxygenase 2 (COX2), 10 inhibitors of neuraminidase (NA) and 18 inhibitors of Cyclin-dependent kinase 2 (CDK2). In these studies different simulation protocols have been applied that make use of a previously developed implicit solvent protocol, i.e. Generalized Born Surface Area (GBSA)[44,45]. Relative binding affinities for the COX2 and NA series excellently rank-ordered compounds but no meaningful correlation for the CDK2 inhibitors could be found. The focus of this study was on small topological changes between compounds within each dataset, typical for alterations

on a scaffold in SBDD. Results showed that empirical scoring functions did not have any productive effect[18].

Another study contrasted the ability of free energy simulation protocols to determine the mode of binding of 16 structurally diverse estrogen receptor α (ERα) inhibitors, and thus to identify binders and non-binders[46]. The simulation was enabled through the development of a novel scheme allowing major topological changes, i.e. *dual topology paradigm*[32], for the in-house software ProtoMS[47]. The most rigorous protocols could correctly identify 5 of the 6 known binders, but accuracy degraded when simpler protocols were used.

Within this project we aim to tackle yet more challenging protein-ligand systems. Chapter 6, *Prediction of binding modes and free energies of binding*, attempts to understand molecular recognition in a challenging protein, Dihydroorotate Dehydrogenase, where we again raise questions on the binding mode and the free energy of binding[48]. A detailed introduction to this system is given in chapter 3. To select an appropriate force field for the study of DHODH, we screen different charge sets[49,50,51] and combine Molecular Mechanics (MM) and Quantum Mechanics (QM)[52] to calculate the relative free energies of hydration on a small but representative set of compounds in chapter 4. To be able to define the thermodynamic end states in terms of hydration networks in this ambiguous system more rigorously, we use GCMC simulations[53] as well as the newly published JAWS algorithm[54]. The methods used together with the results obtained for the definition of these end states is presented in chapter 5. Finally, we try to close the thermodynamic cycles for calculating relative binding free energies and binding modes in DHODH. This leads to validating our findings against the experiment in chapter 6, and we conclude from the lessons we have learnt and finally close this thesis with the chapter 7, *Conclusions*.

2

Biomolecular simulation in a nutshell

Molecular recognition forms the basis for virtually all biological processes[55]. Understanding the interactions between proteins and their associates, i.e. ligands, cofactors or any other molecular entity, is key to rationalise molecular aspects of enzymatic processes and the mechanisms by which cellular systems integrate and respond to regulatory signals. From a medicinal perspective there is great interest in the development of computer models capable of predicting accurately the strength of protein-ligand association[56]. Structure-based drug discovery models seek to predict receptor-ligand binding free energies from the known or presumed structure of the corresponding complex[6,57]. Docking methods and empirical scoring approaches[58], which are useful in virtual screening applications[59], are now routinely employed in drug-discovery programs. In this chapter we introduce a class of computational methodologies that are rooted in the fundamental physical and chemical principles that govern molecular association equilibria[60,61,62], i.e. these methods are based on statistical thermodynamics[6].

Statistical thermodynamics is a branch of physics that applies probability theory to the study of the thermodynamic behaviour of systems composed of a large number of particles, by providing a framework to relate the microscopic properties of individual atoms and molecules to the macroscopic bulk properties of materials[6].

In the view of statistical mechanics, macroscopic properties, such as volume, compressibility, ..., of a system arise from the microscopic, i.e. atomistic, behaviour of

that system. If we consider a system as a collection of N particles in a box, then, at any instant, each particle has a given momentum and occupies a point in space. The set of all positions p^N and momenta r^N of each of the N particles defines uniquely a point $\Gamma = (p^N, r^N)$ in a $6N$ dimensional space called phase space[63]. Under a given set of conditions, for example, constant volume of the box and constant temperature, the collection of particles naturally adopt different sets of positions/momenta through time, i.e. the system follows a time trajectory in phase space. Instead of focusing on the evolution of this trajectory in time, it is possible to imagine that the collection of microstates of the system naturally forms an ensemble. At equilibrium, the microstates in that ensemble are distributed according to a probability density $\pi(\Gamma)$. Two important postulates in statistical mechanics are formulated[6]:

- *Postulate of equal a priori probabilities* states that two microstates i, j that have the same energy are equally probable and therefore $\pi_i = \pi_j$.
- *Postulate of ergodicity* states that the time evolution of a trajectory in phase space is such that one is guaranteed to visit eventually all the states that have a non-zero probability of existence, and thus the time average of a property will equal the ensemble average of that property at equilibrium.

2.1 The Boltzmann distribution

Under these conditions, Ludwig Boltzmann derived an expression for the probability density π for a particular ensemble that lies at the heart of statistical thermodynamics, i.e. the Boltzmann distribution. For the remainder of this chapter we will focus on the canonical ensemble, where N, the number of particles, V, the volume, and T, the temperature of the system are held constant. However, other ensembles exist, such the microcanonical ensemble[64], where N the number of particles, V the volume and E the energy of the system are held constant, i.e. NVE ensemble, and the grand canonical ensemble[6], where T the temperature, V the volume and μ the chemical potential are held constant.

The probability distribution for the NVT ensemble is[64]

$$\pi_{NVT}(i) = \frac{1}{Q_{NVT}} exp(-\beta E_i) \tag{2.1}$$

where E_i is the energy of state i and β is equal to $\frac{1}{k_B T}$, with T the temperature and k_B the Boltzmann constant. The exponential term is known as the Boltzmann factor and represents the weight of the state in that ensemble. Q_{NVT} is a normalisation constant called the partition function[64], and π_{NVT} is often refereed to as the Boltzmann distribution. For a system with a finite number of states, the partition function Q_{NVT} is then simply the sum of the Boltzmann factor of each state, i.e.

$$Q_{NVT} = \sum_i exp(-\beta E_i) \tag{2.2}$$

In the limit of a very large number of states, this equation may be replaced by an integral. In this case one may consider the phase space $\Gamma = (p^N, r^N)$ as a continuum and write under the conditions of the classical approximation[64]:

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \int\int dp^N dr^N exp\Big(-\beta E(p^N, r^N)\Big) \tag{2.3}$$

The term $\frac{1}{N!}$ is introduced for indistinguishable particles, as in this case two configurations differing only by the exchange of coordinates/momenta between two particles correspond to only one real configuration, and thus $\frac{1}{N!}$ must be adjusted if the system is a mixture of particles. The term involving Planck's constant h is of quantum mechanical origin and is introduced to define the volume of the system. The connection between a macroscopic observable A_{obs} and its microscopic value $A(\Gamma)$ can be made through the relationship

$$A_{obs} = \langle A_{ens} \rangle = \frac{1}{Q_{NVT}} \int\int dp^N dr^N A(p^N, r^N)\ exp\Big(-\beta E(p^N, r^N)\Big) \tag{2.4}$$

which states that the ensemble average $\langle A_{ens} \rangle$ is equal to the macroscopic observable A_{obs}. The ensemble average is calculated by integrating over all the positions and momenta that the set of N particles can adopt.

The coordinates and momenta of the system are independent, and thus the energy $E(p^N, r^N)$ of the partition function in equation 2.3 is separable into a kinetic part $K(p^N)$ and a potential part $U(r^N)$[64]. The kinetic part is called the ideal part, as a system where the only energy term is of kinetic origin would be an ideal gas. The potential part is called excess part by reference to thermodynamics where deviations from an ideal system are attributed to 'excess' terms.

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \int\int exp\Big(-\beta\Big(U(r^N) + K(p^N)\Big)\Big)\, dp^N dr^N \tag{2.5}$$

$$= \frac{1}{N!}\frac{1}{h^{3N}} \int\int exp\Big(-\beta U(r^N)\Big) exp\Big(-\beta K(p^N)\Big)\, dp^N dr^N \tag{2.6}$$

$$= \frac{1}{N!}\frac{1}{h^{3N}} \int exp\Big(-\beta K(p^N)\Big)\, dp^N \int exp\Big(-\beta U(r^N)\Big) dr^N \tag{2.7}$$

$$= Q_{NVT-ideal}\; Q_{NVT-excess} \tag{2.8}$$

The integral over momenta, i.e. the ideal part, can be solved analytically via quantum mechanics using the 'particle in a box' model[65]

$$Q_{NVT-ideal} = \frac{V^N}{N!\,\Lambda^{3N}} \tag{2.9}$$

where $\Lambda = (h^2/2\pi m k_B T)^{1/2}$, and Λ is the de Broglie wavelength, m the mass of each particle and V the volume of the system. The remaining excess part is often written as

$$Q_{NVT-excess} = \frac{1}{N!}\frac{1}{h^{3N}}\; Z_{N,NVT} \tag{2.10}$$

where $Z_{N,NVT}$ is called the configurational integral. The excess part cannot be solved analytically due to the large number of coupled energy terms that would need to be evaluated, and hence the integral over all possible configurations is solved numerically. It is common to omit the first two terms of equation 2.10 and focus on the configurational integral only. Hence, when we are interested in the ensemble average of a property that depends only on the coordinates, the momentum contributions can be safely ignored, and the calculation of $\langle A_{ens}\rangle$ simplifies to

$$\langle A_{ens}\rangle = \frac{\int dr^N\; A(r^N)\; exp\Big(-\beta U(r^N)\Big)}{Z_{N,NVT}} \tag{2.11}$$

which shows that the ensemble average of a property A is the ratio of two integrals over a space of r^N dimensions. Since under the postulate of ergodicity this ensemble average is equal to the macroscopic value of A, equation 2.11 provides us with the mean to derive thermodynamic properties ab initio.

2.2 Classical potentials

However, in order to calculate any thermodynamic property A, we need to evaluate the potential energy function $U(r^N)$ in equation 2.11. Undoubtedly, quantum

mechanics (QM) provides us, in principle, with the most accurate description of the potential energy function for any molecular species, while the computational expense of these methods for large biomolecular systems is currently limiting their application[66].

Molecular Mechanics (MM) is a method of modelling atoms and molecules as simple particles under the influence of classical physics[6]. It is a fundamental theory, where each atom may be described by a centre where the nucleus resides, containing the mass and partial charge of the atom. The charge is used to model electrostatic interactions using the Coulomb potential[7], and the size of the atom is described by the Lennard-Jones potential[7]. These atoms are then connected to create molecules by adding potentials describing bonds, angles, dihedrals as well as improper dihedrals.

A complete set of such potential energy terms for describing atoms and molecules and their interactions, with corresponding constants for all available atom types, is referred to as a force field[6,7,63]. Force fields rely on simple functional forms and sets of parameters to reproduce the either experimental properties, such as the OPLS force field[67,68], or quantum mechanical properties of molecules, such as the AMBER99[69] and CHARMM22[70] force fields.

Force fields may differ in their level of atomic description; there are all-atom force fields that explicitly account for all atoms in the system and there are united force fields which unite non-polar hydrogens with their neighbouring atom. For example, the force field OPLS may represent all atoms of a system individually, i.e. OPLS-all atom (OPLS-AA), or unite non-polar hydrogens with their neighbours, i.e. OPLS-united atoms (OPLS-UA). Since a united force field reduces the size of a system in terms of atoms, the advantage of using these force fields is an increase in speed of the calculations, as fewer potentials need to be evaluated, while a decrease in accuracy may become a disadvantage. Moreover, force fields that collapse molecular substructures to large 'atomic assemblies' to simplify the molecular description and yet again speed up calculations have been developed, i.e. coarse-grained force fields such as the MARTINI force field[71]. However, throughout this thesis we will be mainly employing the AMBER99 force field[69], hence we illustrate the functional forms that are associated with it.

The functional form of the total potential energy, i.e. U_{total}, in the AMBER99 force field can be written as the sum of all terms relating to bonds, angles, dihedrals and non-bonded interactions:

$$U_{total} = U_{bond} + U_{angle} + U_{dihedral} + U_{non-bonded} \tag{2.12}$$

where the bond and angle terms, U_{bond} and U_{angle} respectively, are described by harmonic potentials to account for all directly bonded, i.e. 1-2 interacting, or directly angled, i.e. 1-3 interacting, atoms:

$$U_{bond} = \sum_{bonds} K_b(r - r_{eq})^2 \tag{2.13}$$

$$U_{angle} = \sum_{angles} K_\theta(\theta - \theta_{eq})^2 \tag{2.14}$$

where r and θ correspond to the bond length and valence angles, respectively, while r_{eq} and θ_{eq} are the associated equilibrium values, and K_b and K_θ are the force constants.

The harmonic bond potential, and similarly the harmonic angle potential, can be viewed as a spring connecting the atoms. The functional form of the harmonic potential gives rise to steep curves far from the bond equilibrium, which are not physically realistic. The harmonic description yields, however, potentials that are fast to evaluate and gives a good approximation close to the equilibrium bond distance, while the loosening of an atomic bond could not be described accurately.

Dihedrals describe the potential energy barriers around a bond during a full rotation and are calculated by a Fourier sum of several periodic functions. Each of the terms in the Fourier series is given by

$$U_{dihedral} = A_n \Big(1 + cos(n\ \phi - \delta)\Big) \tag{2.15}$$

where n is the period describing the number of minimum point of the function as the dihedral angle changes from 0 to 2π. δ is the phase angle, ϕ is the dihedral angle, and A_n is the force constant.

Finally, the non-bonded terms are calculated for every pair of atoms in the system. Non-bonded terms describe the interaction between atoms in separate molecules and atoms connected by more than three bonds. These terms are the driving force of the interactions between a protein and a ligand, and are divided into a Lennard-Jones

potential part and an electrostatic part. In the AMBER99 force field, non-bonded interactions are calculated via

$$U_{non-bonded} = \sum_i \sum_{j>i} \left\{ \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \tag{2.16}$$

where the sum is over all atom pairs i,j. The q_i and q_j are the atomic partial charges, ϵ_{ij} and σ_{ij} are the Lennard-Jones well-depth energy and collision diameter parameters respectively; ϵ_0 is the permittivity of free space and r_{ij} is the inter-atomic distance. Since the non-bonded term is also applied to atoms that are separated by more than three bonds, the magnitude of these intramolecular interactions is reduced by a scaling factor, i.e. 0.83 and 0.5 for the intramolecular Coulombic and Lennard-Jones interactions respectively, in the AMBER force field.

The evaluation for non-bonded interactions constitutes the majority of interactions which need to be calculated during a simulation. The reason for this is that the sum over all atoms needs to be calculated in the system which makes the number of non-bonded interactions scale quadratically with the number of atoms in the system, and thus united-atom force fields reduce computational time. In comparison, the bonded terms scale linearly as there is only a limited number of bonded terms per atom. To increase the speed of the calculation, intermolecular terms may be truncated such that interactions of atoms separated by more than a cutoff distance are ignored[7]. The cutoff distance may be applied between pairs of atoms or it may be based on the distance between pairs of groups, e.g. if the closest distance between two residues of molecules is greater than the defined cutoff distance, then all pair-pair interactions between the groups are ignored. However, it has been shown that the truncation of non-bonded terms can cause discontinuities in the potential energies and forces associated with the interaction[63]. In the current work, the intermolecular non-bonded terms have been scaled by a switching function, $S(r)$, to overcome this problem.

$$U^{scaled}_{non-bonded}(r) = S(r)\ U_{non-bonded}(r) \tag{2.17}$$

where r is the distance between atoms. To preserve the nature of the non-bonded interactions at low r, the interactions are gradually switched to zero by the cutoff

distance, and may be applied over a range of distances, such as:

$$U^{scaled}_{non-bonded} = U_{non-bonded}(r) \; for \; r < r_{feather}, S(r) = 1 \tag{2.18}$$

$$U^{scaled}_{non-bonded} = S(r) \; U_{non-bonded}(r) \; for \; r_{feather} \leq r \leq r_{cut} \tag{2.19}$$

$$U^{scaled}_{non-bonded} = 0 \; for \; r > r_{cut}, S(r) = 0 \tag{2.20}$$

where r_{cut} is the cutoff distance, and $r_{feather}$ is the distance beyond which the switching function feathers the non-bonded interactions gradually to zero.

2.3 Sampling methods

Having defined a force field that allows the evaluation of $U(r)$ in equation 2.11, we now introduce methods to generate ensembles and thus allow the estimation of the configurational integral in the same equation.

2.3.1 Metropolis Monte Carlo

In 1953, Metropolis and coworkers reported the *Metropolis Monte Carlo* method (Metropolis MC method)[72], which forms the basis for Monte Carlo statistical mechanics simulations of atomic and molecular systems[73]. The application of the Metropolis MC method for a biomolecular system, in essence, consists of the following steps:

1. Start in state i
2. Attempt a move to state j with probability p_{ij}
3. Accept this move with probability $\alpha_{ij} = min(1, \chi)$ where $\chi = \frac{\pi_j}{\pi_i}$
4. If the move is accepted set i = j, otherwise set i = i (i.e. reject the move)
5. Accumulate any property of interest $A(i)$
6. Return to step 1 or terminate after a certain number of moves have been attempted

An important property that the Metropolis MC algorithm must obey is the principle of detailed balance or microscopic reversibility[7]:

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{2.21}$$

Now we let Q_{ij} be the probability that a move i to j is accepted and assume $\pi_j < \pi_i$.

$$\pi_i Q_{ij} = \pi_j Q_{ji} \tag{2.22}$$

$$\pi_i p_{ij} \alpha_{ij} = \pi_j p_{ji} \alpha_{ji} \tag{2.23}$$

$$\pi_i p_{ij} \frac{\pi_j}{\pi_i} = \pi_j p_{ji} \tag{2.24}$$

$$p_{ij} = p_{ji} \tag{2.25}$$

and we see that detailed balance is not violated when the unmodified transition matrix is symmetric, i.e. the probability of moving from i to j, before weighting by π_i and π_j, is the same as the probability of moving from j to i.

Accordingly, if we want to use Metropolis MC sampling to generate an ensemble appropriate for the Boltzmann distribution, then the acceptance test is[7]

$$\frac{\pi_{j,NVT}}{\pi_{i,NVT}} = \frac{exp(-\beta U_j)/Z_{N,NVT}}{exp(-\beta U_i)/Z_{N,NVT}} \tag{2.26}$$

$$= exp\Big(-\beta(U_j - U_i)\Big) \tag{2.27}$$

Equation 2.27 shows that we do not need to know the normalisation factor $Z_{N,NVT}$. This is fortunate, as it is usually not possible to determine this parameter[64], as the numerical integration over all configurations converges very slowly for all but the simplest systems. Therefore, in a Metropolis MC simulation, the ratio of the Boltzmann factor of states i and j is calculated and compared to a random number u drawn uniformly between [0,1]. If $u < (\pi_{j,NVT}/\pi_{i,NVT})$ the move is accepted, otherwise it is rejected[7].

2.3.2 Monte Carlo moves

To satisfy detailed balance, i.e. to assure that the transition matrix probabilities p_{ij} and p_{ji} are equal, Monte Carlo methods often select a trial state j randomly. As the majority of all possible configurations of a molecular system have very high energies, and we have no a priori knowledge of the interesting regions of phase space, the trial state j is usually formed by attempting a small alteration to state i. This is because if state i is a member of the ensemble p, then a state j that is similar to state i has a reasonable probability to be part of the same ensemble. In practice this is often done by picking randomly one particle in the system and performing a random translational/rotational displacement of that particle, and if the particle

has any internal degrees of freedom, then these degrees of freedom can be randomly modified as well. Here, the rule of thumb suggests to set these modifications to allow an overall acceptance rate of 40 % throughout the simulation. Moreover, once the maximum ranges for random displacements and alterations of internal degrees of freedom have been set, they should be kept constant throughout the simulations, as detailed balance would be violated if they were changed during the simulation.

2.3.3 Generalized ensembles

The potential energy surface of biomolecular systems of interest is often found to be frustrated, owing to the large number of degrees of freedom usually encountered in these systems. Thus, many minima may exist that are separated by substantial energy barriers, which poses a serious challenge to standard Metropolis MC. Parallel Tempering (PT)[74] tries to overcome these problems by increased sampling of the entire system through the formation a generalised ensemble over temperature. This is achieved by running a set of simulations of a given system at different temperatures. In this situation, each system is referred to as a replica and the method is called *replica exchange*. Periodically, moves are attempted between different replicas, and if the move is accepted, then the replicas exchange their temperature/coordinates and the simulations proceed normally until the next attempted move. The acceptance test for this parallel tempering moves is designed to ensure that each simulation is forming a correct NVT or NPT ensemble[7]. For instance, in a NVT simulation a replica i at temperature β_A should exchange with replica j at temperature β_B with probability[6]

$$exp\Big[\Big(\beta_B - \beta_A\Big)\Big(U_B(j) - U_A(j)\Big)\Big] \geq rand\ (0,1) \tag{2.28}$$

Hence, parallel tempering aids in configurational sampling of the ensemble, as low temperature configurations can be taken into a high temperature simulation, where larger configurational changes are more likely, and then cool down back to their original temperature, in which case enhanced configurational sampling has been achieved. A difficulty with this method is that two different replicas must be simultaneously exchanged and the high temperature replica is less likely to be a representative member of the low temperature ensemble. It is therefore necessary to keep a small temperature interval between two different replicas, and hence depending on the temperature many more simulations may be required. Another drawback

is the necessity to run several simulations at high temperature when only one ensemble at room temperature may be of practical interest.

2.3.4 Molecular dynamics

Apart from Metropolis MC methods, the most commonly used method to generate ensembles of thermally relevant states is Molecular Dynamics (MD)[7]. Although MD simulation has not been subject of this thesis, we briefly introduce the method. MD takes the most obvious route available to generate an ensemble of configurations, namely that of evolving the time trajectory.

Here the system is placed in a starting configuration, at a particular point on the energy surface. The gradient at that point on the surface can be evaluated, and the negative of this gradient is a force that can be converted into an acceleration via Newton's law of motion. However, the direct solution to Newton's law of motion requires the solution of 3N coupled, second order differential equations, where N is the number of atoms in the system. This is too difficult to solve analytically, and thus numerical approaches relying on finite difference methods are used. These methods integrate the forces over time to yield a trajectory.

Newton's law of motion conserves energy, and the resulting trajectory from an MD simulation samples naturally from the microcanonical ensemble, i.e. the NVE ensemble. However, algorithms that connect the system to a thermostat or barostat have been developed and allow the sampling of the NVT or NPT ensemble.

Compared to MC simulation methods, where random displacements are attempted, MD simulations are of deterministic nature as they follow Newton's laws of motion. Therefore, in an MD simulation, a starting configuration of a biomolecular system is sufficient to generate a trajectory of the system over time. Because of the postulate of ergodicity, the ensemble of states visited in a MD simulation should be identical to those generated by a MC simulation (in the limit of sufficiently long sampling).

While MC and MD approaches should theoretically give the same answer, in practice one method may outperform the other on a particular system. MC is algorithmically simpler to implement than MD, particularly for simulations in the NPT ensemble. Because MD follows the time evolution of a system, dynamical properties can be studied, which is not feasible in typical MC simulations although the Kinetic Monte Carlo method can partially overcome this difficulty[75]. In a MD simulation, all the

degrees of freedom of the system are subject to forces and hence move, and it is often necessary to constrain many degrees of freedom using algorithms such as the SHAKE method[76]. In a MC simulation, no degree of freedom is sampled unless it has been chosen and the implementation of constraints is therefore trivial. In principle, MC is not required to climb an energy barrier to sample two connected minima although it can be difficult to design a move that efficiently explores unrelated minima.

2.4 Rigorous free energy calculation methods

The free energy governs many important thermodynamic phenomena, as it points in the direction of spontaneous change. Being able to predict the free energy allows one to predict solvation, stability, phase transitions and many other properties[65]. The free energy of binding is the change in free energy associated with the binding of a guest to a host, and is a direct measure of the strength of that binding. In the canonical ensemble, the free energy of a system, i.e. the so-called Helmholtz free energy, can be related directly to its partition function[6]:

$$A = -k_B T \; lnQ_{NVT} \tag{2.29}$$

If the partition function was calculated over the isothermal-isobaric ensemble, i.e. the NPT ensemble, then the quantity on the left hand side of equation 2.29 is the Gibbs free energy, usually denoted G. This is because the partition function for any ensemble, Q_{ens}, can be related to the thermodynamic potential for the ensemble, ψ_{ens}[64]

$$\psi_{ens} = -k_B T \; lnQ_{ens} \tag{2.30}$$

The thermodynamic potential for an ensemble has a minimum value at thermodynamic equilibrium[64], and hence it is possible to determine the direction of a process by measuring the absolute free energy of two comparable systems.

2.4.1 Absolute free energy calculations

As demonstrated in section 2.1, we ignore the contribution of the ideal part to the partition function, which may be analytically evaluated[63], and write the following for

the excess free energy:

$$A = -\frac{1}{\beta} ln Q_{NVT} \tag{2.31}$$

$$= \frac{1}{\beta} ln\Big(\frac{1}{Q_{NVT}}\Big) \tag{2.32}$$

$$= \frac{1}{\beta} ln \frac{N! h^{3N}}{\int exp\Big(-\beta U(r^N)\Big) dr^N} \tag{2.33}$$

Now we can write

$$1 = \frac{1}{(8\pi^2 V)^N} \int exp\Big(+\beta U(r^N)\Big) exp\Big(-\beta U(r^N)\Big) dr^N \tag{2.34}$$

where the constant factor in equation 2.34 comes from the integration of 1 over phase space. Inserting equation 2.34 into equation 2.33 yields

$$A = \frac{1}{\beta} ln \frac{N! h^{3N}}{(8\pi^2 V)^N} \frac{\int exp\Big(+\beta U(r^N)\Big) exp\Big(-\beta U(r^N)\Big)}{\int exp\Big(-\beta U(r^N)\Big)} \tag{2.35}$$

$$= \frac{1}{\beta} ln \frac{N! h^{3N}}{(8\pi^2 V)^N} \int exp\Big(+\beta U(r^N)\Big) \pi(r^N) dr^N \tag{2.36}$$

$$= \frac{1}{\beta} ln \frac{N! h^{3N}}{(8\pi^2 V)^N} \Big\langle exp\Big(+\beta U(r^N)\Big) \Big\rangle \tag{2.37}$$

Thus equation 2.37 shows that the free energy of a system can be calculated as an ensemble average. Calculation of the constant factor would require the definition of the volume of phase space V which can be difficult to calculate. However, if we were interested in the difference in absolute free energy between two systems, then this term may be safely ignored as it only acts to shift the value of the absolute free energy by a constant offset.

However, equation 2.37 for calculating the absolute free energy of binding for a protein-ligand system for an ensemble generated using Metropolis MC sampling, shows poor convergence behaviour. This is because Metropolis MC samples states according to the Boltzmann distribution, which generates mostly states of low energy, while states of high energy are rarely encountered. This is unfortunate, as high energy states contribute largely to the ensemble average because of the sign of the exponential.

2.4.2 Free energy perturbation

Although a direct approach to the calculation of free energies is impractical, the calculation of the free energy difference between two systems is possible, i.e. the calculation of the relative free energy. This was reported by Robert Zwanzig in 1954[11]. According to Zwanzig, the relative free energy between two different systems A and B can be calculated as follows:

$$\begin{aligned}
\Delta G_{A\to B} &= G_B - G_A && (2.38)\\
&= \Big(-\frac{1}{\beta} ln Q_B\Big) - \Big(-\frac{1}{\beta} ln Q_A\Big) && (2.39)\\
&= -\frac{1}{\beta} ln\Big[\frac{Q_B}{Q_A}\Big] && (2.40)\\
&= -\frac{1}{\beta} ln\left[\frac{\int exp\Big(-\beta U_B(r^N)\Big) dr^N}{\int exp\Big(-\beta U_A(r^N)\Big) dr^N}\right] && (2.41)
\end{aligned}$$

multiply by $1 = exp\Big(-\beta U_A(r^N)\Big) exp\Big(\beta U_A(r^N)\Big)$ gives,

$$\begin{aligned}
&= -\frac{1}{\beta} ln\left[\frac{\int exp\Big(-\beta U_A(r^N)\Big) exp(-\beta\Big(U_B(r^N) - U_A(r^N)\Big) dr^N}{\int exp\Big(-\beta U_A(r^N)\Big) dr^N}\right] && (2.42)\\
&= -\frac{1}{\beta} ln\left[\int \frac{exp\Big(-\beta U_A(r^N)\Big)}{Q_A} exp\Big(-\beta \Delta U_{AB}(r^N)\Big) dr^N\right] && (2.43)\\
&= -\frac{1}{\beta} ln\left[\int \pi_A(r^N) exp\Big(-\beta \Delta U_{AB}(r^N)\Big) dr^N\right] && (2.44)\\
&= -\frac{1}{\beta} ln\left\langle exp\Big(-\beta \Delta U_{AB}(r^N)\Big)\right\rangle_A && (2.45)
\end{aligned}$$

where π_A is the Boltzmann probability of configuration A in the ensemble of state A, and ΔU_{AB} is the difference in energy between system A and B. The derivation of Zwanzig shows that the relative free energy is the logarithm in the ensemble average of the exponential of the Boltzmann weighted energy difference between the two states.

In computer simulations, the Zwanzig equation is implemented using the Free Energy Perturbation (FEP) methodology[7]. Here, a simulation is performed with the potential U_A and at each iteration/move the quantity $exp(\Delta U_{AB}(i)/k_B T)$ is accumulated.

Now we imagine that we use equation 2.45 to calculate the free energy difference

between two systems, S_1 and S_2, of very different potential energy functions. If the low energy regions of S1 are part of the phase space that corresponds to high energy regions of S_2, then a simulation run with potential U_{S1} will rarely generate the significant configurations of potential U_{S2}. As a result, the free energy change $\Delta G_{S1 \rightarrow S2}$ is likely to be overestimated. Likewise, if the potentials are switched, $\Delta G_{S2 \rightarrow S1}$ will be overestimated. Any difference between these two quantities is known as hysteresis, and if hysteresis is large, the calculated free energies will be a poor approximation of the actual quantity. However, a simple solution is to multi-stage the calculation with the introduction of the reaction coordinate λ.

2.4.3 The λ-coordinate

The λ-coordinate helps in connecting the phase space of two systems S_1 and S_2, and thus increases the chances for the Zwanzig equation to converge. More specifically, λ is used to gradually morph one system into the other, such that at $\lambda = 0.0$, the system represented by the potential energy function is system S_1, and at $\lambda = 1.0$ it is system S_2. Any intermediate λ-values define the system as a non-physical hybrid of S_1 and S_2. The Zwanzig equation can therefore be rewritten

$$G_{S_2} - G_{S_1} = \Delta G = \sum_{\lambda=0}^{1} -k_B T \; ln \Big\langle exp\Big(- \Delta U^{'} / k_B T \Big) \Big\rangle_{\lambda_k} \tag{2.46}$$

where $\Delta U^{'} = U_{P(\lambda)_{k+1}} \; - \; U_{P(\lambda)_k}$.

Two concepts may be applied to calculate the free energy differences of the two systems using the Metropolis MC method. These are the more traditional *single topology* approach[6], and the *dual topology* method that was previously published from our group[32]. Both methods are described in the following two sections.

2.4.3.1 Single topology method

The implementation of relative free energy calculations in a computer program requires the definition of a scheme to transform the potential energy function of a system S1 into the potential energy function of system S2. The calculated single free energy differences are not directly comparable to experiment if they are not related before to a reference state. This is usually accomplished through the construction of a thermodynamic cycle. A representative cycle is illustrated in figure 2.1.

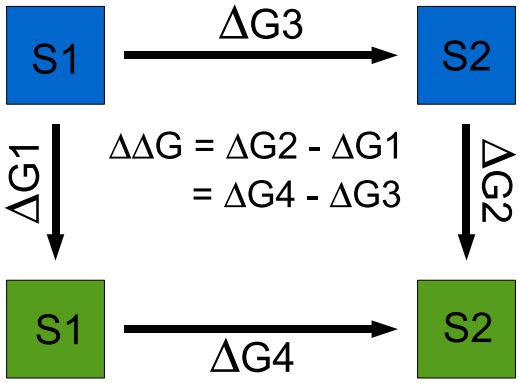


Figure 2.1: A general thermodynamic cycle that relates the difference in free energy between S1 and S2 in two media, indicated by the green or blue squares surrounding S1 and S2 in the figure. S1 and S2 could be two small molecules in two media, for instance blue for vacuum and green for solvent, in which case the double free energy difference will correspond to the relative hydration free energy of S2 with respect to S1. If the media would represent a solvated protein and pure water, i.e. green and blue respectively, then the double free energy difference will correspond to the relative binding free energy of S2 with respect to S1. While the vertical processes, corresponding to ΔG_1 or ΔG_2, are often measured experimentally, the horizontal processes, corresponding to ΔG_3 or ΔG_4, are usually easier to calculate in a computer simulation.

To calculate the free energy difference according to figure 2.1 by for example FEP we need to define the potential energy functions representing the interactions of S1 and S2 with their environments. A suitable functional form is

$$U(\lambda) = U_0 \; + \; \Delta U(\lambda) \tag{2.47}$$

where U_0 corresponds to any energy terms that are not related to either S1 or S2. The second term in equation 2.47 depends on the lambda coordinate and often varies between 0 and 1 to represent S1 at $\lambda = 0.0$ and S2 at $\lambda = 1.0$. In a single topology calculation each force field term $\Delta U(\lambda)$ is defined as a linear combination of the values of the force field terms of both systems S1 and S2. For example, the angle stretching term may be expressed as

$$U_{ang}(\lambda) = K_\theta(\lambda)[\theta - \theta_{eq}(\lambda)]^2 \tag{2.48}$$

$$K_\theta(\lambda) = \lambda K_\theta(S2) + (1-\lambda)K_\theta(S1) \tag{2.49}$$

$$\theta_{eq}(\lambda) = \lambda\theta_{eq}(S2) + (1-\lambda)\theta_{eq}(S1) \tag{2.50}$$

The coulombic energy for an atom i belonging to the perturbed system and an atom j belonging to the surrounding medium would be

$$U_{coul}(\lambda) = \frac{q_i(\lambda)q_j}{4\pi\epsilon_0 r_{ij}(\lambda)} \quad (2.51)$$

$$q_i(\lambda) = \lambda q_i(S2) + (1-\lambda)q_i(S1) \quad (2.52)$$

$$r_{ij}(\lambda) = \lambda r_{ij}(S2) + (1-\lambda)r_{ij}(S1) \quad (2.53)$$

Inspection of equation 2.51 shows that geometric terms and force field terms vary in the coupling of S1 and S2. Hence, the single topology method encounters difficulties when the number of atoms between S1 and S2 changes, and traditionally dummy atoms are introduced to balance the number of atoms between the two systems and thus overcome this problem. In the particular end state where the dummy atom should not exist, it should also not contribute to the intermolecular energy. However, a complete lack of, for example, bond or angle terms associated with the dummy atom could result in the dummy atom dissociating from the molecule it is attached to, which would invariably lead to a divergence of the calculated free energy differences. Therefore, bond or angle terms are usually associated with the dummy atom throughout the perturbation, and, as shown in figure 2.1, its contribution should normally cancel out in the double free energy difference. Thus, the single topology method requires the topologies of either system to be identified by a single transition matrix, using the same number of atoms. Consequently, if for a particular perturbation this matrix cannot be defined, then the perturbation cannot be carried out using the single topology approach.

2.4.3.2 Dual topology approach

The exact nature of the coupling while going from $\lambda = 0.0$ to $\lambda = 1.0$ is arbitrary, as long as the systems S1 and S2 correspond rigorously to the end states. Hence, it has been proposed to define simultaneously S1 and S2 in the medium[6]. In this case the potential energy function becomes

$$U(\lambda) = U_0 \; + \; \lambda U(S_2) \; + \; (1-\lambda)U(S_1) \quad (2.54)$$

And for example the coulombic energy of atoms i from S2, i‘ from S1 with an atom j from the medium would be:

$$U_{coul}(\lambda) = \lambda \, \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \; + \; (1-\lambda)\frac{q_i^{'} q_j}{4\pi\epsilon_0 r_{i'j}} \quad (2.55)$$

Here, the coupling of λ is associated with the scaling of the interaction energy terms, instead of scaling the force field parameters. As described by Michel and coworkers[32], ignoring any pair-pair interaction energy terms involving systems S1 and S2 while the intramolecular non-bonded energy terms for the systems are not scaled with λ, results in a fully decoupled system at either end state. This is equivalent to transferring S1 or S2 to the ideal gas phase. Because both systems are represented individually, and they do not experience any interaction with each other, the method is called dual topology. A representative free energy cycle for the dual topology approach is shown in figure 2.2.

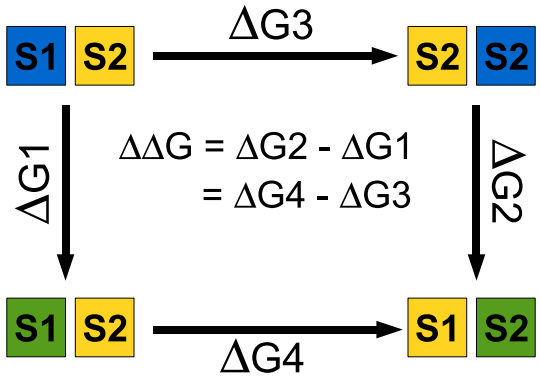


Figure 2.2: A thermodynamic cycle constructed with the dual topology method that relates the difference in free energy between S1 and S2 between two media shown in green and blue, while the system in their decoupled states are shown by yellow squares, i.e. the ideal state where S1 or S2 are in the gas phase.

At intermediate values of λ both systems are interacting with their media, while at the end states only one system is interacting. This means that depending on the value of λ the system may overlap or show close contact with regions of the media. The functional form of the dual topology scaling for the Lennard-Jones terms would be

$$U(\lambda) = U_0(LJ) \; + \; \lambda U(S2)_{LJ} \; + \; (1-\lambda)U(S1)_{LJ} \tag{2.56}$$

where the Lennard-Jones terms involving interactions with S1 or S2 and the medium are scaled by λ or $(1-\lambda)$ respectively. Very high energies for non-bonded interactions will be experienced for systems that approach closely, because of the high exponent

on the repulsive term of the Lennard-Jones interaction energy. Therefore, simulations using dual topology with a standard Lennard-Jones potential almost invariably yield divergent free energy profiles close to the end states, where one of the systems must be turned off completely. This is also true, in principle, for single topology calculations involving dummy atoms. However, in practice this problem is overcome by retracting the dummy atoms into the van der Waals radius of a nearby non-dummy atom, and thus overlaps can be overcome in the single topology approach.

In the dual topology method described by Michel and coworkers, this so called "Lennard-Jones end point singularity" is overcome by soft-scaling the interactions by using the approach of Beutler et al.[77] and Zacharias et al.[78] using the following equation

$$U_{LJ,soft,\lambda} = (1-\lambda)4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{(\lambda\alpha_{soft}+r_{ij}^2)^6}\right) - \left(\frac{\sigma_{ij}^{6}}{(\lambda\alpha_{soft}+r_{ij}^2)^3}\right)\right] \tag{2.57}$$

Equation 2.57 is equivalent to the standard LJ equation when λ is set to 0 and as the coupling parameter increases, the Lennard Jones interactions are gradually softened such that when λ is close to unity, atomic overlaps are permitted. This allows for the smooth annihilation of an atom i belonging to S1. For the atoms i of S2, the parameter λ is simply substituted by $(1-\lambda)$. The degree of softness of the potential depends on the parameter α_{soft} that is chosen to yield a smooth free energy gradient and several simulations may be needed to define the optimum value.

The softening of the Lennard-Jones interactions may result in atomic overlaps between non-bonded particles. Therefore, it could be possible for two atoms of opposite charge to adopt the same coordinates and experience an infinitely attractive coulomb energy. Thus, the coulombic interactions may be scaled as well.

$$U_{coul} = \frac{(1-\lambda)q_i q_j}{4\pi\epsilon_0\sqrt{(\lambda+r_{ij}^2)}} \tag{2.58}$$

2.4.4 Thermodynamic integration

While FEP directly employs the Zwanzig equation to calculate the difference in free energy along the λ-coordinate, thermodynamic integration (TI) takes a different approach[10]. TI still looks at discrete λ-values along the coordinate, and generates an MC or MD trajectory at each λ-value. However, instead of calculating the difference

in energy between neighbouring λ-values, it calculates the rate of change of free energy, with respect to λ at each point, i.e. it estimates the free energy gradient $(\frac{\delta G}{\delta \lambda})_\lambda$. Thus, TI avoids the problems of bad overlap potentially encountered in FEP, because the free energy gradient is a property of the system at each value of λ only. Having collected all of the free energy gradients during the simulation, they may then be integrated to yield the free energy change along the λ coordinate.

$$G_{\lambda=1} - G_{\lambda=0} = \int_0^1 \left(\frac{\delta G}{\delta \lambda}\right)_\lambda d\lambda \tag{2.59}$$

The integral can be evaluated numerically, e.g. via the trapezium rule or simpson rule[6], while the free energy gradients themselves may be obtained analytically or numerically. The analytical approach uses a modified force field to calculate the gradient of each force field term directly with respect to λ. The ensemble average of the gradient of the force field, $\langle \frac{\delta U}{\delta \lambda} \rangle_\lambda$, is equal to the free energy gradient[6]

$$\int_0^1 \left(\frac{\delta G}{\delta \lambda}\right)_\lambda d\lambda = \int_0^1 \left\langle \frac{\delta U}{\delta \lambda} \right\rangle_\lambda d\lambda \tag{2.60}$$

The numerical route approximates the gradient, $(\frac{\delta G}{\delta \lambda})_\lambda$, via the finite difference, $(\frac{\Delta G}{\Delta \lambda})_\lambda$. This free energy difference may be calculated using the Zwanzig equation, with a state at λ, i.e. the reference state, and the state at $\lambda + \Delta\lambda$, i.e. the perturbed state. This results in a forward estimate of the free energy gradient, and a perturbed state of $\lambda - \Delta\lambda$ yields the backwards estimate. Both estimates should be equal if $\Delta\lambda$ was sufficiently small, and the trajectory ran until the Zwanzig equation had converged. This method is normally referred to as *Finite Difference Thermodynamic Integration* (FDTI).

Using this scheme, FDTI is very similar to FEP, and the perturbed states are the neighbouring windows, while for FDTI, the perturbed states are $\Delta\lambda$ above and below each window.

2.4.5 Replica exchange thermodynamic integration

Work previously published from our group resulted in a novel free energy method, i.e. *replica exchange thermodynamic integration* (RETI)[40]. The idea of RETI is based on the idea of generalized ensemble methods[74]. To conduct a RETI simulation, a set of replicas of the system that cover the range of the coupling parameter λ are run.

Periodically, moves between replicas i and j of Hamiltonian H_A and H_B are attempted, subject to the acceptance test

$$exp\Big[\beta\Big(E_B(j) - E_B(i) - E_A(j) + E_A(i)\Big)\Big] \;\geq\; rand\;(0,1) \tag{2.61}$$

A RETI simulation can be performed at no extra cost since all simulations are already required for TI or FDTI simulations, and neighbouring replicas tend to exchange with reasonably high probabilities as the systems tends to be more similar over a change of λ. Hence, RETI provides enhanced sampling as it allows individual trajectories to exchange to neighbouring λ-windows, i.e. exchange with related configurations in phase space. In favourable cases, it can allow some replicas to overcome barriers, as a replica at λ_i exchanges with another replica running at λ_j value which does not experience this barrier. There the replica performs some local sampling and then exchanges back into the original λ_i value in a region that lies beyond the barrier. However, if at every λ value a similarly high barrier is present, then the quality of the sampling will not be improved much over standard methods. However, the evaluation of the RETI method has been reported to perform better than established free energy methods[40], and thus our simulations made use of the RETI methodology.

2.4.6 Calculating errors in free energy simulations

The errors in free energy simulations are due to many numerical approximations in a computer simulation. These approximations can be understood as arising from two major sources: the representation of the system and finite sampling.

Examples of errors that may arise from the representation of the system include the choice of the empirical force field, the choice of interaction cutoff distances, the treatment of electrostatic effects, the choice of the system size, and the choice of the model representation, e.g. coarse-graining or the use of an implicit solvent protocol. These choices can lead to systematic errors, creating a computer model with an unrealistic representation of the system of interest that will not converge to the correct free energy. As such, these errors will not be reduced by adopting improved free energy calculations or by increased sampling[6].

Imperfect sampling contributes the other major source of error found in free energy calculations. Although statistical mechanics provides us with a formally exact

solution to compute free energies - with the knowledge of the entire phase space of a system - a computer simulation generates a finite number of configurations. Therefore, sampling errors are an intrinsic part of free energy simulations and may be even more critical part of the analysis due to rough energy landscapes. However, unlike the systematic error that may occur due to the unfortunate choice of a system representation, sampling errors will be reduced when more sampling is conducted and better free energy algorithms or sampling algorithms are used[6].

The ensemble average of a property $\langle A \rangle$ is said to have converged if it does not change significantly when the number of configurations used to determine it is increased. The block-average method can be used to assess the convergence behaviour[79]. Here, a simulation of N configurations is subdivided into K blocks of N/K configurations. $\langle A \rangle_K$ is then calculated for each block followed by calculating the standard deviation from that distribution of values. In principle, if all the portions of phase space that contribute significantly to $\langle A \rangle$ have been visited with the right probability in each block, all the values will be similar and the standard deviation low. However, it is essential to make sure that the blocks must be long enough to be completely statistically uncorrelated with each other. Monte Carlo or molecular dynamics simulations generate successively highly correlated states and the number of steps that are necessary before a configuration is uncorrelated to its starting configuration is system dependent and can not be easily determined. Furthermore, a low standard deviation by no means guarantees that simulation results have converged to the right answer. If we imagine, that the system we are simulating is unable to climb local barriers, then the entire simulation may explore thoroughly one local minimum only, while missing out other important regions of phase space. Here, a block-average analysis will suggest the results are converged, but they are actually not.

Rather than relying on block averaging to obtain error estimates, one could run several independent simulations, using different starting points that may have been obtained previously by annealing or that are initiated using different random number seeds. This method has the obvious drawback that one has now to run several simulations instead of one. Depending on the actual context in which the simulations are employed one may consider running multiple simulations to generate more accurate results.

Moreover, when one is interested in the free energy difference of several related systems, it is possible to assess to some degree the convergence of the simulation results by running a few additional simulations that help to close thermodynamic cycles. Figure 2.3 highlights the general principle. Because free energy is a state function, the sum of the changes in free energy along a pathway that start from state A and eventually returns to that state, should be equal to zero. This allows the user to specify arbitrary states that seem convenient to prove thermodynamic cycle closure within a reasonable error estimate. The extent by which the cycle closure deviates from zero is a measure of the lack of convergence.

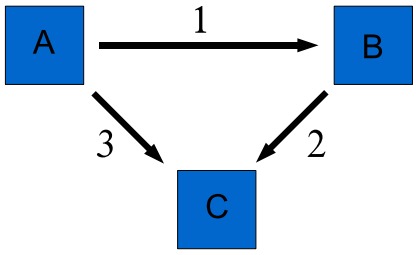


Figure 2.3: The closure of a thermodynamic cycle illustrated on a general system. While only two simulations may be necessary for the calculation of the relative free energy of system B or C with respect to A, a third simulation that calculates the relative free energy of system C with respect to B allows the closure of a cycle involving A, B, and C. Any deviation from zero while walking along the cycle, e.g. 1 + 2 - 3 = 0, is an error estimate.

Unfortunately, none of the approaches discussed here guarantees that the simulation results are truly converged. Once again, the formally exact equations from statistical thermodynamics to calculate free energies assume that all important regions of phase space have been visited with the appropriate probability. Therefore, it appears in the realistic context that without an *a priori* knowledge of the potential energy surface, it is impossible to assert rigorously whether or not the results of a simulation are truly converged.

3

Defining a problem for structure based drug design

A realistic model for simulating protein-ligand binding would ideally consider the binding and unbinding processes directly, leading to a simulation that covers the lifetime of the resulting complex of protein and ligand many times. With current computer resources this is not a feasible approach. Rigorous free energy methods transform one chemical species into another or a non-interacting dummy particle by introducing nonphysical intermediate states. Because free energy is a state function, the choice of intermediate states is arbitrary, but in practice has great impact on the efficiency of such calculations[80]. Following an initial wave of enthusiasm around 10 to 15 years ago, it became evident that initial results may have been mainly successful due to simply luck, although the methods, in principle, allow an exact theoretical definition of the ligand binding problem[81]. As with any scientific discipline, more refined theoretical developments led to yet deeper understanding and recent reviews can help to visualize the long and winding roads from which free energy simulations have come[61,82,83,84]. Important developments and insights are discussed in this section and a new and elaborate problem in rational drug design is presented, aiming to capture the most important and influential aspects one has to consider for a real biological target in order to develop a more reliable simulation protocol. The challenges in obtaining accurate estimates for binding affinities are varied and as such are discussed separately in the next sections of this chapter.

3.1 Generating an atomic model system

Before a free energy calculation can be performed, an atomistic model for the system under investigation must be constructed. The simplest form of a biomolecular complex consists of the protein and a ligand together with solvent molecules, The modelling of real biological targets usually contains the inclusion of ions, often catalytically active metal ions in protein binding sites, or crystallographic water molecules. The initial structure will inevitably influence the outcome of a free energy study. The nature of alchemical methods is to connect two thermodynamic states, and the more realistic and accurate these states are defined the more likely will a free energy calculation be accurate.

3.1.1 Crystallographic data sources

The most common form of experimental data available for biomolecules are x-ray crystallographic structures, although the nubers of NMR resolved structures are constantly growing too[85]. These x-ray structures are important for the accuracy of free energy calculations. First, hydrogen atoms are usually not resolved by crystallographers and need to be added, and a variety of algorithms exist to aid this task. While this might not be a problem for standard amino acids it can become difficult to unambiguously define a protonation or tautomeric state for charged species, and hence a careful visual inspection of critical residues is highly advised. Moreover, at a resolution of 2 Å the relative position of the δN and δO atoms of asparagine and the γN and γO atoms of glutamine side chains cannot be determined directly, as their electron densities are isoelectronic. The decision on whether each density should be assigned to the N or O atoms should be based on the local hydrogen bonding network. However, if this decision has been made by the crystallographer before solvent molecules have been added to the model, then the local environment has not been accounted for, and hence it is based on incomplete hydrogen-bonding networks. More light on this matter is shed via a large-scale validation study performed with the software WHAT-IF, as listed in on the PDBREPORT website[86]. According to this study as many as one in six of all histidine, asparagine and glutamine residues in the PDB may have to be modelled in a 'flipped' orientation, supporting the fact that, during

the step-wise refinement process, waters have not been included in the assignment of amino acid side-chain orientations.

Water, sodium and ammonium ions are the the most abundant entities in the crystallographic process, as they act as constituents in crystallization media[87]. Unfortunately, these entities provide excellent means to improve model statistics in crystallography, and cannot always be distinguished based on their density alone, because they are almost isoelectronic. For water molecules this becomes particularly problematic, as with standard atomic resolutions of around 2 Å it becomes a subjective matter for the crystallographer whether a feature in the density should be ignored as noise or modelled as a water molecule. Here we have to be aware that the uncritical addition of solvent molecules (each of which introduces four adjustable parameters: x, y, z coordinates and an isotropic temperature factor into the model) provides the crystallographer with excellent means of absorbing problems into both the atomic model and the experimental data. Addition of water molecules is then simply used to artificially reduce the differences between observed and calculated structure-factor amplitudes[87]. Many examples exist where crystallographers in independent studies on the same biomolecule try to assess the impact of refining water molecules. The structure of transforming growth factor $\beta 2$ was determined by two independent laboratories at similar resolutions, 1.8 Å (1TGI)[88] and 1.95 Å (1TGF)[89]. There are 58 water molecules in 1TGI with an average temperature factor of 31.8Å^2 and 84 water molecules with an average temperature factor of 43.3Å^2 in 1TGF. In 1TGF the 54 molecules common to 1TGI have much lower temperature factors (average 34Å^2) than the 30 extra water molecules (average 60Å^2), which suggests that the latter have a much lower level of reliability[90]. Finally, statistics from the PDB verification tool WHAT-IF[91], found at the PDBREPORT website[86], identify 99793 water molecules in 10857 structures deposited in the PDB that have no hydrogen bonds to any other atom in the structure.

Even when the protonation state and orientation of key residues are known, they may change upon complexing with an inhibitor. Enthalpies of complexation measured by isothermal titration calorimetry in aqueous buffers with different enthalpies for ionization established that the Roche thrombin inhibitor napsagatran binds to thrombin using an additional proton, while an inhibitor with similar structure from Behring, i.e. compound CRC220, binds to thrombin without an additional proton[92].

For the inhibitor molecules similar uncertainties arise, although they are generally even more difficult to capture by experiment. While high-quality dictionaries of acceptable bond lengths, angles and torsions are available for amino and nucleic acids in model refinement, the same is not true for complexed ligands, because of their huge diversity. Thus, when setting up an atomic model for a biomolecular simulation, it is rather common to find multiply resolved binding modes or orientations for the ligand. For instance, the position of the carboxylate group of oleic acid in mutant rat intestinal fatty acid binding protein (1FABP) was ambiguous when the electron density was examined[93]. Indeed, the crystallographers report three positions for the carboxylate group in the deposited PDB file (1ICN), with occupancies of approximately 0.3 for each. The computational chemist often interprets these findings as if multiple binding poses for the inhibitor, or natural substrate in this particular case, exists. As such this complex, despite its uncertain structure from a crystallographic point of view, was selected as a validation structure when the docking program Flex-X was tested with the scoring function DOCKSCORE[94] and GOLD[95]. It may not come as a surprise that both Flex-X and GOLD failed to reproduce any of the observed conformations of oleic acid. Instead, Flex-X calculated the best pose to be rotated by 180°, leaving us with a total of four different binding poses. Further experiments, however, using different types of IFABP confirmed that in fact none of the suggested poses was accurate, as the carboxylate group points towards the solvent and does not interact with the any protein residues[96]. Such a binding mode has already been seen in another fatty-acid-binding protein resolved by crystallographers[97], indicating the importance for the computational chemist to look at the family of targets and not a single PDB structure alone.

3.1.2 Implications for free energy methods

How does this now relate to free energy calculation? As developments in free energy methods progress, a more detailed picture is being revealed with regards to the sensitivity of results in connection with initial input structures of both ligand as well as protein, and together with further crystallographic experiments they reflect unexpected protein and ligand dynamics, and in combination result in a deeper understanding of biomolecular targets[62]. Rigorous free energy methods can be used to compute either absolute binding affinities, i.e. for an individual ligand and receptor,

or relative binding affinities, i.e. for the difference between two or more related ligands[62]. For relative and absolute free energy studies the most popular model system to study new developments is the hydrophobic cavity mutant of T4 lysozyme[62], as well as a mutant of this enzyme aiming to introduce polarity to the ligand binding site, i.e. M102Q T4 lysozyme. It has been shown that despite the simplicity of its small, apolar, and relatively rigid binding site, it exhibits these problematic crystallographic characteristics a free energy study needs to address in order to produce accurate results[98]. In fact, T4 lysozyme poses unexpected challenges not only towards rigorous free energy methods, but also virtual screening approaches[99]. Initially, free energy methods have failed to quantitatively predict affinity[100] for some of the compounds. This is most surprising, as the typical ligands complexed to T4 lysozyme are small substituted benzene molecules, such as toluene, and would most likely be described as a congeneric set of compounds. To date, it has been assumed that ligands in a congeneric series will bind to the protein in an almost identical manner, and hence once a crystal structure of a complex has been obtained, their 'congeneric' members of the series have been modelled in the same binding mode, while tremendous efforts would have been required to resolve all complexes for a series. This typical modification scheme to a formerly developed scaffold should reveal whether small modifications to the ligand scaffold would actually result in increased affinity and selectivity. Typically, relative free energy studies have been carried out for congeneric sets of compounds. This is for technical reasons, i.e. the setup of a transition matrix while transforming one ligand into another, and the error cancellation postulate in relative free energy calculations. If all ligands share the same binding mode and no protein conformational changes occur that would alter the interactions of ligand and protein in the series, then relative free energies benefit from fortuitous cancellation of errors, hence facilitating the computation of accurate binding affinities while computer times are reduced, compared to the calculation of absolute free energies of binding[62].

In the case of the T4 lysozyme model system, free energy simulations combined with further crystallographic studies have shown that, even though the ligands are of congeneric nature, they are also reminiscent of fragment screening sets, and as such may possess multiple nearly degenerate binding orientations, separated by substantial kinetic barriers[101], which makes it also difficult to refine the binding modes

unambiguously through experiment[46,102] and frustrates affinity estimation[101] in a free energy study if conformational sampling has not accounted for this fact. Hence, dominant ligand binding modes can be far from obvious, even given the structure of a closely related ligand. In some cases, multiple binding modes may be relevant, which has been observed in calculations and experiments in which multiple binding mods are clearly resolved; in others they may simply be an artefact of crystallographic refinement. Moreover, it has been shown that minute changes to a ligand scaffold can dramatically alter the binding mode[102], further complicating the setup of a free energy study for a congeneric series of compounds. Therefore, when multiple binding modes or important rotameric states of protein side chains have been addressed in a simulation, current force fields did allow errors of 1 to 2 kcal mol^{-1} [19,101,103], while prediction errors against experiment of up to 6 kcal mol^{-1} have been observed when these issues have been left untouched. Finally, ligands as well as proteins may change protonation or tautomeric states upon complexation, or there may be significant populations of multiple states during some part of the binding process, a problem recently termed *multiple state problem*[104]. Semigrand canonical simulation methods may assist to incorporate changes in protonation or tautomerism, but practical applications are rare[105].

Current literature clearly demonstrates that protein conformational change, even on a level of a single rotamer sidechain, has been shown to be potentially too slow to be sampled during a simulation[100,106], and relative free energy calculations only avoid this problem if this change affects all the ligands under consideration in exactly the same way, unless the conformational change is accounted for[106]. Ligand binding modes, even though generally considered predictable by docking methods, are far from trivial to predict, and assuming a dominant mode based on the bound structure of a closely related ligand may be problematic. As a result of the lack in sampling of those modes, dramatically different relative binding free energy estimates may be calculated, and dependence on the starting structure[98] may occur. Hence, treating a PDB entry as a simply array of atomic coordinates at perfect resolution is a gross oversimplification and can easily lead to false assumptions concerning the model. Also, assumptions made by crystallographers in modelling the electron density may appear minor when one considers the correctness of an entire protein-ligand

complex, but they will have profound effects when the structure is used as the basis for structure-based drug design and free energy calculations. Situations where the 'cancellation of errors' assumption breaks down are almost impossible to predict ahead of time, and can lead to erroneous free energies that make failure to agree with experiment difficult to interpret[98].

3.2 Biological affinities

A common practice to assess the performance of free energy methods is to look at the correlation of predicted free energies versus experiment, which often measures affinities indirectly, i.e. IC_{50} or the apparent inhibition constant K_{app}. Care has to be taken when converting those experimental proxies to free energies, as they only relate to the free energy under very specific mechanistic conditions[107]. Having met these specific mechanistic criteria, experimental measurements are still invariably contaminated with error, which again could limit the maximum possible correlation of predicted and measured affinities and also hide errors within the calculations. Howeve, an assessment of the rigorous free energy protocols by comparing to experiment is a viable route.

3.3 Sampling

Protein-ligand binding is regulated by the conformational preferences of both the ligand and the protein in ways that are often not fully elucidated by the analysis of crystallographic data alone, or requires substantially more diffraction experiments than are often feasible in a drug design project. However, rigorous free energy simulations, based on statistical mechanics are only limited, in principle, by the accuracy of the force field used and the extent of conformational sampling achieved during a simulation[6]. While sampling is discussed in this section, concerns regarding the use of force fields as well as the impact of biomolecular hydration are discussed in the next sections. Conformational sampling algorithms applicable to binding free energy calculations must be able to not only find the relevant conformations of the system but also visit them with the correct probability[62]. For example, if as outlined above, a situation arises where a minute change to a ligand leads to a dramatically

different binding mode, or even subtle conformational changes on the level of a single amino acid side-chain, are not being accounted for, then the free energy estimate may not be accurate. Also, it has been shown that strongly bound, but rarely visited conformations of the complex are not necessarily the most relevant for the binding equilibrium, a situation that could not be understood via crystallographic experimentation, but simulation methods only. Given that alchemical methods require the simulation of the free as well as the bound states further complicates the analysis of free energy methods[62].

Binding free energy simulation protocols based on traditional Metropolis MC sampling usually only consider a very limited number of conformations. However, recently, enhanced sampling algorithms capable of equilibrating distinct conformational macrostates have been applied to the ligand binding problem and assisted in obtaining accurate estimates for the observed binding affinities[106]. To simplify the discussion on sampling, it may be useful to categorize the degrees of sampling that are relevant to the association process. First, we can differentiate intermolecular motions of the ligand relative to the protein from the intramolecular motions of either ligand and protein. Second, since we are focusing on systems of biological origin, we can look at the degrees of freedom for the water molecules, implicated in every biological process, and therefore of importance.

As outlined in section 2, alchemical protocols using MC or MD sampling usually compute relative free energies of binding by FEP and TI, as well as variations thereof, while for example the double decoupling method, explained in chapter 5 are applicable to calculate absolute free energies of binding. All these methods are based on a reaction coordinate λ used to connect any two thermodynamic states by modulating the interaction of the ligand with its environment. To allow any potential energy function to be assessed in terms of accuracy it is essential to ensure that all conformational space has been visited appropriately. However, most studies to date are characterized by an uncertain coverage of conformational space, most likely because in many systems the subtle structural requirements that have to be met are simply unknown. This may be reflected in poor convergence rates, hysteresis effects, or simply highly inaccurate free energy estimates without apparent reason.

A promising remedy to improve sampling of the intermolecular degrees of freedom are schemes that are based on generalized ensembles or replica-exchange protocols,

such as RETI[40], λ-dynamics[108], FEP/REMD[109] and BEDAM[103], all of which have shown to yield superior conformational sampling and more rapid convergence rates through the introducing of λ-hopping, which allows an exchange of different simulation threads with one another, or by propagating those across the entire lambda coordinate. Alternatively, a soft-core potential can be defined for a reference state to help sampling of alternative binding modes[33,35]. When energetic barriers are too high to be overcome by these schemes, it is also possible to perform complete sampling within a well defined, local macrostate, as it it easier to achieve sampling than equilibrating distinct binding modes. From an appropriate combination of contributions of each binding mode, the absolute binding free energy can then be obtained. However, the challenge here is to identify the highest contributing mode, and failure to do so can introduce major errors, as can neglecting important secondary modes, although effects will be less dramatic in terms of accuracy of the results obtained[101,103]. When the position of a ligand in a protein binding site is restricted to a single macrostate, improved sampling can also be achieved, if the imposition of conformational restraints is matched by their release at a later stage of the simulation[61]. However, the effects of multiple binding modes are shifted towards the restraining free energy component which may be difficult to fully converge unless accelerated sampling methods are being used to sample all important conformations, as the restraining free energy component is actually computed with full ligand-protein interactions. A very recent development is the Mining Minima technique introduced by Gilson and coworkers[110], as it is not relying on MD or MC sampling methods, and hence does not suffer from typical slow transition rates, but a complete enumeration of all important stable minima of the protein-ligand complex is required, rendering the method quite exhaustive.

The rotation of a sidechain upon binding of an inhibitor from a congeneric series, is a small but important example for the underlying, and often much more substantial, reorganization of the natural ensembles of ligands and receptors. This phenomenon, often referred to as conformational reorganization, induced fit or conformational selection[111], corresponds to the free energy associated with restraining ligand and receptor in their bound states. In case of a single side-chain reorganization, more traditional approaches have been applied, such as the work published

by Mobley and coworkers using an umbrella sampling potential together with restraints, allowing the reorganisational effect of a single amino acid side chain to be quantitatively captured, i.e. the *confine and release* approach[106]. It has been shown that predictions improved markedly for the T4 lysozyme system, but for the case of huge conformational changes, such as opening and closing of receptors, novel methods needs to be developed as the conformational search becomes significantly larger. Some models for protein-ligand binding free energies include reorganisation or reorganization energies via

$$\Delta G_{bind} = \Delta G_{inter} + \Delta G_{reorg} \tag{3.1}$$

where ΔG_{inter} is the intermolecular component of degrees of freedom discussed in the former paragraph, and ΔG_{reorg} represents the reorganization energy. If we were to assume that a series of compounds complexed to a protein is strictly congeneric, hence neither ligand nor protein reorganization will occur, then reorganization energies would not influence a relative free energy study. This situation, however, cannot be known *a priori*. Most novel sampling approaches have so far focused on the energy associated with the intermolecular degrees of freedom, although it is now widely recognized that reorganization may play important roles. A recent example successfully used reorganization energies as a design principle for the optimization of the presentation of HIV epitopes for vaccine development[112]. Since the nature of this binding interface is biologically constrained, preorganisation of the ligand to the bound conformation is the only viable route for optimizing affinity. However, the development and implementation of reorganization for alchemical free energy methods is not advanced, mainly because the effect is inherently dynamic[113], requiring the knowledge of both a range of conformational states and their probability of occurrence in solution. Nevertheless, computer simulation is to date the only feasible route to model these phenomena and provides us with more insight into this problem. Clearly, the reorganization of a receptor is much more difficult to capture than for a ligand, hence, apart for Mobley's work[106] and an approach using a two-dimensional Hamiltonian replica-exchange free energy perturbation approach to soften side-chain torsional barriers[114], current approaches mainly focus on ligand reorganization effects[115]. The Mining Minima method, where the reorganization free energy is assembled by directly computing the configurational partition functions of a set of low

energy conformational states is one of them[110,116], and there is evidence that MD sampling aided by temperature replica-exchange can be used to compute conformational populations of ligand conformational states[117]. In fact, Generalized Ensemble MD sampling methods establish in principle a particularly elegant approach, due to their generality and scaling properties, while no exhaustive conformational enumeration is required. However, future developments will eventually provide us with a clearer picture of the performance of these new methods.

3.4 Water in free energy simulations

Water is fundamental to any biomolecular association, and hence often mediates ligand-protein interactions and allows a rationalization of the physics involved upon ligand binding[118,119]. Modelling has revealed the complex thermodynamics of water in protein cavities[120,121] and recent studies have investigated the free energy gain or loss for displacing water molecules to accommodate the ligand[122], and to use this knowledge in rational drug design. Michel and coworkers have employed an algorithm[123] based on the DDM method for displacing water molecules[124] and concluded that water molecules are ambiguous partners, and their contribution to binding very much depends on the exact details of the structural and energetic properties of ligand and protein[125]. Their approach allows for a hydration pattern to be defined more rigorously. Moreover, Grand Canonical Monte Carlo (GCMC) algorithms, promoting exchange of waters between the binding site and bulk water, have provided another route to assess hydration patterns[126,127]. Given the ease of adding water molecules by crystallographers, it is essential to properly evaluate the biomolecular hydration pattern before a reliable free energy simulation may be attempted. Not only may waters constitute an essential part to define thermodynamic end states, i.e. the estimated free energy of binding would miss the contribution of deleting or creating a certain water molecule that mediates important ligand-protein interactions, but may also slow diffusion and the kinetic trapping of waters can lead to systematic errors and slow convergence of binding free energy calculations, prohibiting sampling to some degree. Similar situations have been observed for alchemical transformations in water only. Although solvent degrees of freedom generally relax on a much shorter

timescale than protein and ligand conformational changes, thereby aiding convergence, slow torsional transitions have been reported, leading to considerable errors even in the calculation of free energies of hydration[25].

3.5 Force Fields

Several steps have been taken to improve force fields[128,129] but given their nature of parametrization, i.e. the condensed phase, quantum mechanical optimization in the vacuum phase or reproduction of experimental properties, they may be inherently dependent on these environments . As such, it may not be expected for force field parameters to perform well in the gas phase when initially they were parametrised for the condensed phase, unless significant corrections are applied[130]. Improved sampling has often led to an improved understanding of the limits of force field accuracy. For example, a direct comparison of force fields was published by Shirts and coworkers where large-scale distributed computing was used to calculate the free energies of hydration of amino acid to high precision[131]. A study of Mobley and coworkers has led to the identification of problematic Lennard-Jones parameters for alkynes[132]. Other groups have addressed force field effects on the kinetics and thermodynamics of α-helices[133].

Despite their obvious failings and advantages, it is more important for force fields to be transferable, as protein environments are very different in dielectric, polarization and density from aqueous environments, and hence protein-ligand binding affinities may also accumulate inaccuracies due to the neglect of this environmental dependency. For this reason, several groups have been working on polarisable force fields, such as AMOEBA[134] and CHARMM[135], but extensive validation and testing is still outstanding. Another popular approach, that in principle aims not only to incorporate polarization but also generate force field parameters - especially helpful for exotic chemistries as a full molecular mechanics based parameterization is an extensive task - are methods that combine the traditional molecular mechanics force fields with quantum mechanical approaches, i.e. QM/MM[66]. This approach, together with a validation of molecular mechanics force field parameters, are the subject of chapter 4 and as such will not be discussed in great detail here.

Most commonly used force fields for biomolecular systems, such as GAFF[136], AMBER[69,137], GROMOS[138], OPLS[67,68] and CHARMM[70], have been validated and extensively tested as essentially all groups working on biomolecular simulation and that make use of a molecular mechanics description of the energetics involved, have to rely on one of these force fields. The majority of the results show that indeed current force fiels are useful and applicable for the study of biomolecular system, while minor modifications undoubtedly lead to yet better descriptions of energetics and properties[132]. Finally, the much more varied nature of ligand molecules as opposed to proteins and nucleic acids, makes clear that special attention has to be paid to the setup of these compounds. Numerous software tools exist that aid in generating force field parameters for these molecules and their careful use makes them valuable tools. Very exotic chemistries require the parametrization of new force field parameters, which clearly is an elaborate task.

3.6 A realistic problem in rational drug design

In this section we would like to present a new system that we believe establishes a very realistic biological target for rational drug design. Our aim was to identify a system that alleviates as many issues discussed above as possible, while providing us with the most thorough experimental data at the same time, in order to assess the impact for free energy simulations. Work previously published from our group by Michel et al.[18] has demonstrated the application of novel free energy simulation protocols to calculate free energies of binding to COX2, NA and CDK2. While all these targets are certainly of great pharmaceutical interest, their impact for validating newly developed protocols is still limited, mainly because they were carried out according to best practices for congeneric series, and as such did not address any of the issues raised by the community recently[62]. Additionally, for systems that had not been the subject of successful free energy studies that can be found in the literature, such as the CDK2 system, the protocol failed to be accurate. In another study published previously from our group, the prediction of binding modes of a diverse set of compounds complexed to ER-α was attempted with free energy simulations[46]. While the performance of this free energy study was outstanding compared to simpler docking protocols, the majority of compounds investigated did not offer any experimental

data that would allow the assessment of the developed protocols in terms of accuracy. For this project, we found a particular set of compounds complexed to the protein Dihydroorotate Dehydrogenase (DHODH) that seems to match all the requirements we are keen to address with free energy simulations for validation purposes:

- Five crystal structures for five congeneric inhibitor molecules
- Presence of prosthetic groups, substrates and inhibitors
- Large protein conformational changes are not observed
- Changes in protonation/tautomerism of either protein or inhibitors are unlikely
- Minute changes to the inhibitors dramatically alter the binding modes
- Significant changes in hydration patterns are proposed for the different inhibitors

3.6.1 Dihydrooorotate Dehydrogenase

Human Dihydroorotate-Dehydrogenase (DHODH) is an oxido-reductase group enzyme, anchored at the inner mitochondrial leaflet, where it aids the catalysis in the rate-limiting step in the *de-novo* biosynthesis of pyrimidine nucleotides[48,139], i.e the conversion of 4,5-Dihydroorotic acid (DHO) into Orotic acid (ORO) as shown in figure 3.1.

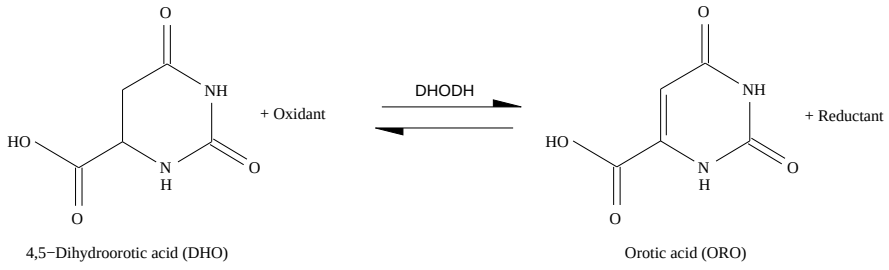


Figure 3.1: The rate-limiting step in the de-novo synthesis of pyrimidine nucleotides utilizes DHODH to convert DHO to ORO. Dependent on biological species, different entities are being used for their oxidation potential.

When increased levels of nucleotides are required, such as during the course of inflammation, mammal pyrimidine nucleotide synthesis is elevated mainly through

this *de-novo* recruitment, while the *salvage pathway* synthesis is the main source of pyrimidine production during non-inflammatory states[139].

The group of DHODH enzymes can be differentiated based on sequence alignments[140], and each group differs in its selectivity for the oxidizing substrate. The family 1A enzymes are water soluble enzymes containing flavin-mono-nucleotide as the only prosthetic group and use fumarate as the oxidizing substrate[141]. Family 1B contains a second protein subunit harbouring an iron-sulfur centre and a flavin-adenine di-nucleotide (FAD)[142,143,144]; this electron transport chain reduces the physiological oxidant, nicotine amide-adenine di-nucleotide (NAD). The family of class 2 enzymes, found in the majority of all eukaryotes and gram-negative bacteria, contain flavin mono-nucleotide (FMN) and are oxidized by ubiquinone instead; hence human DHODH is member of class 2[48,139].

In all DHODH enzymes, DHO is oxidized by the enzyme-bound FMN prosthetic group to form the α,β-unsaturated carbonyl moiety of the product ORO (see figure 3.1). The oxidations that form α,β-unsaturated carbonyl compounds are catalysed by a number of flavin-containing enzymes. More specifically, the somewhat acidic proton α to the carbonyl is removed by an enzymatic base and the β-hydrogen is transferred as a hydride to the isoalloxazine system of the flavin[145]. The two hydrogens may then be transferred in a single step (a concerted reaction), or substrate deprotonation may precede hydride transfer (a stepwise reaction)[145]. Isotope effects on steady-state kinetics suggest for family 1A DHODH from *Crithidia fasiculata* an abstraction of the proton and transfer of hydride to FMN in a stepwise manner, while similar studies for families 1B and 2 point towards a concerted mechanism for DHO oxidation. It appears, that the reaction mechanism is imposed by the protein rather than by the requirements of FMN and DHO, and the active site where DHO binds and is oxidized is nearly identical throughout all structures available for all classes of DHODH enzymes[146]. For human DHODH, i.e. class 2, the concerted mechanism - also referred to as *ping-pong mechanism*[48] - may be thought of as illustrated in figure 3.2.

3.6.2 DHODH as a drug target

DHODH is a well characterized target for small molecular weight (D)isease-(m)odifying-(a)ntirheumatic (d)rug(s), i.e. DMARDs[147]. The term DMARDs was originally intro-

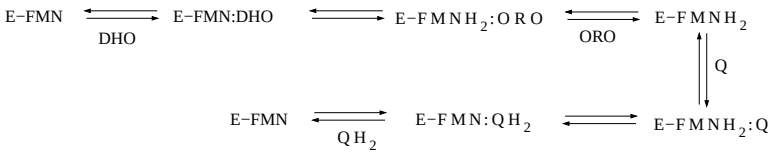


Figure 3.2: Schematic of the concerted reaction suggested for class 2 DHODH. Picture taken from[48].

duced as a group of heterogeneous agents that have the capacity to alter the course of rheumatoid arthritis (RA), with a typical inflammatory pathophysiology , i.e. marked by the infiltration of immunocompetent cells, i.e. CD4+ cells, and monocytes, proliferation of synovial lining cells and fibroblasts, as well as neovascularization mediated by pro-inflammatory cytokines, e.g. Interleukin-1 (IL-1) and tumor-necrosis factor α (TNFα), produced by those cells[148]. To give a short overview of the inflammatory processes and joint destructive mechanisms involved during the pathophysiological progression of arthritis, a schematic representation can be found in figure 3.3. From a drug discovery point of view, it is interesting to see that the pro-inflammatory cytokines involved in these cascades are major targets for current drug discovery projects, often in the form of biologicals[149], aiming to get hold of not only inflammatory but also autoimmune diseases, while DHODH inhibitors, in principle, offer a more traditional route to treatment using small molecular weight compounds.

Although the use of DMARDs was first propagated in RA, the term has nowadays come to pertain to many other diseases. Leflunomide is a DMARD that has become a base-line therapy not only for RA but also received orphan drug status[151]. The drug's activated form, the metabolite A771726, also known under the name Teriflunomide, shown in figure 3.4 and a product of the first-pass metabolism[148], is a potent inhibitor of DHODH. Even though Leflunomide outperforms other base-line therapies in this indication, current DHODH inhibitors seem to cause severe side effects, i.e. hepato- and nephrotoxicities[152], and hence the search for new DHODH inhibitors could be of pharmaceutical value.

3.6.3 Baumgartner's series of DHODH inhibitors

Inhibition of human DHODH, prevents the recovery of FMNH2 via ubiquinone, recruited via the inner mitochondrial membrane while anchoring the enzyme to the

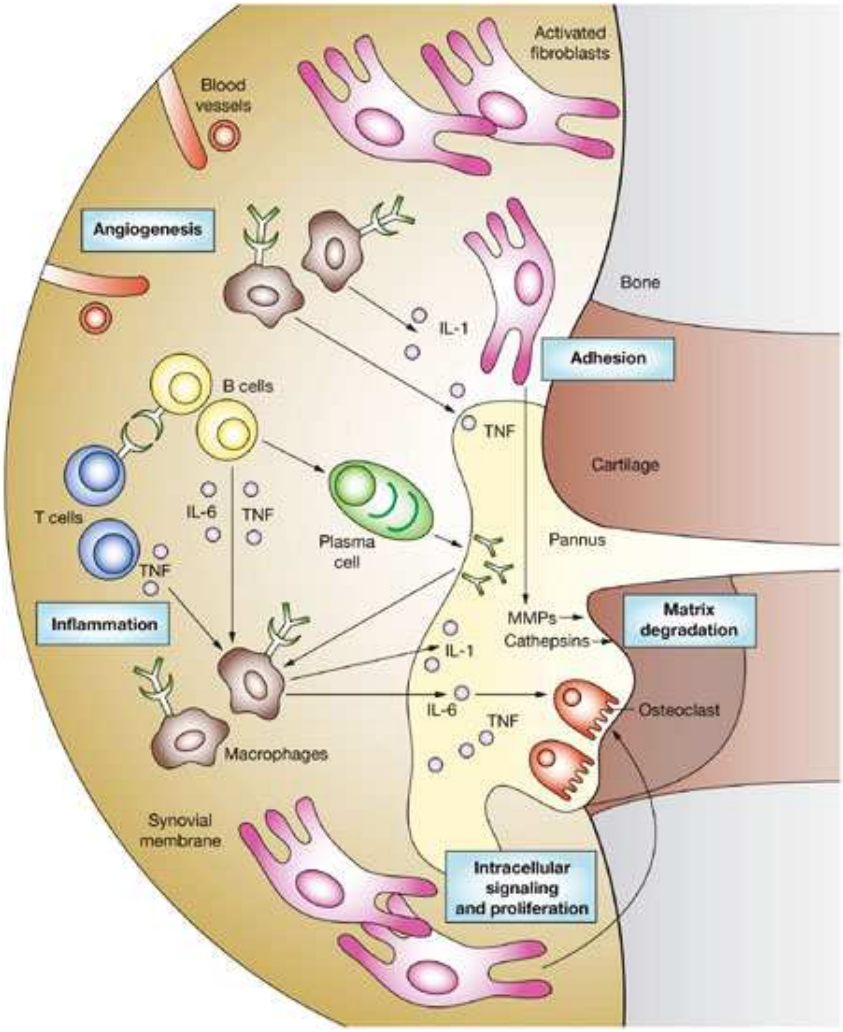


Figure 3.3: Pathways involved in inflammation and destruction in the rheumatoid joint. The five key factors of intracellular signalling and proliferation, adhesion, inflammation, angiogenesis, and matrix degradation are linked by various inflammatory effector cells, such as TNF, IL-1 and interleukin-6 (IL-6), and matrix-degrading enzymes, including matrix metalloproteinases (MMPs) and cathepsins, finally resulting in a persisting viscious circle. The figure has been taken from [150].

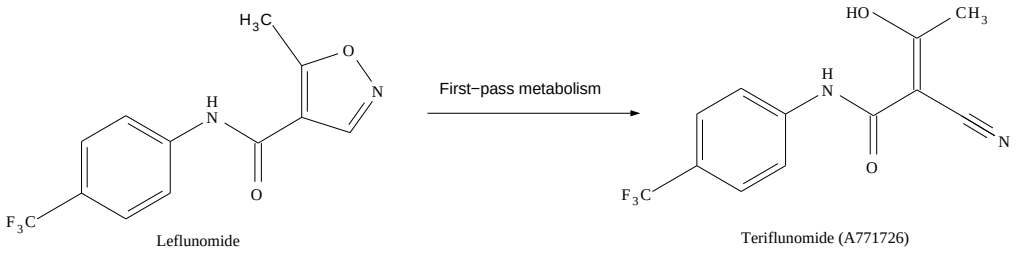


Figure 3.4: The active form of Leflunomide is a product of first-pass metabolism where the ring-opening of the isoxazole ring leads to the current product.

membrane, hence aborting *de-novo* pyrimidine nucleotide synthesis, a major pathway during inflammatory states[48]. Kinetic as well as structural studies suggest two distinct binding sites for DHO and ubiquinone, and it is the ubiquinone binding site that is commonly targeted by DHODH inhibitors[48].

The pharmacophoric features of the drug Brequinar, the most prominent inhibitor of DHODH, together with other design approaches, i.e. QSAR and docking, have been used as a starting point to develop a novel series of molecular entities with DHODH inhibitory activity by Baumgartner and coworkers[48]. Figure 3.5 shows Brequinar complexed to human DHODH and the particular binding pose adopted by the compound is termed brequinar-like binding mode.

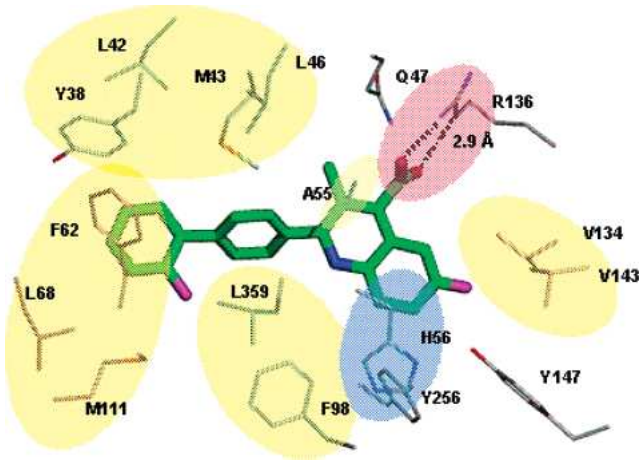


Figure 3.5: A depicted view of Brequinar in the binding site of human DHODH. Figure has been taken from[48].

3.6.3.1 Crystallographic data source

Here, we focus in particular on a set of 5 congeneric compounds of extraordinary structure activity relationship that have been resolved in high-resolution quality. The crystallographic data allows, according to the authors, one to analyse prevailing interactions and modes of binding of the inhibitors in detail[48]. A summary of crystallographic parameters can be found in table 3.1 and on the PDB[85] with deposit tags 2BXV, 2FPT, 2FQI, 2FPV and 2FPY for compounds 3, 4, 5, 6 and 7 respectively. Unfortunately, only the refined coordinates were made available by the authors. It is suggested by the PDB[85] to deposit the experimental data, i.e. electron densities, on the Uppsala electron density server[153], as the raw data allows researchers to gain deeper understanding of the refinement process. Unfortunately, Baumgartner and coworkers did not use this feature.

In summary[48], co-crystallization was performed with the hanging-drop method with a truncated version of human DHODH, following the protocol described by Liu et al.[154], the inhibitors described in this section and brequinar at 20 °C. The drops consisted of a mixture of equal amounts of 20 mg/mL protein in 50 mM HEPES, pH 7.7, 400 mM NaCl, 30% glycerol, 1 mM EDTA, and 10 mM N,N-dimethylundecylamin-N-oxide (C11DAO) with a precipitant solution of 0.1 M acetate pH 4.6-5.0, 40 mM C11DAO, 20.8 mM N,N-dimethyldecylamine-N-oxide (DDAO), 2 mM dihydroorotate (DHO), 1.8-2.4 M ammonium sulfate, and 1 mM of the inhibitor compound. The drops were incubated against 0.5 mL of reservoir of 0.1 M acetate pH 4.8, 2.4-2.6 M ammonium sulfate and 30% glycerol and crystals usually appeared as small cubes within 3 days and reached a full size of 0.2 x 0.2 x 0.2 mm^3 within 3-4 weeks.

3.6.3.2 Overall structure of human DHODH

The overall structure of DHODH is composed of a large C-terminal domain (MET78-ARG396), that can best be described as an α-β barrel fold with a central barrel of eight parallel beta strands surrounded by 8 helices. This domain harbours the substrate binding site as well as the prosthetic group, thus reduction of substrate is located in this domain. The smaller N-terminal domain (MET30-LEU68) contains the binding site for the cofactor ubiquinone and comprises two helices, i.e. helices α-1 and α-2, that span a slot of about 10 x 20 $Å^2$ in the so called hydrophobic patch and are

	Compound 3	Compound 4	Compound 5	Compound 6	Compound 7
Crystal data					
Space group	$P3_221$	$P3_221$	$P3_221$	$P3_221$	$P3_221$
Resolution	2.15 Å	2.4 Å	1.95 Å	1.8 Å	2.0 Å
molecules/AU[a]	1	1	1	1	1
Data collection					
X-ray source	DESY BW6	DESY BW6	DESY BW6	DESY BW6	DESY BW6
Completeness[b]	99.2/99.0	95.8/97.1	99.8/99.9	99.9/100.0	98.6/99.0
R-factor (%)[c]	20.1/19.1	17.6/19.4	18.5/20.6	19.5/20.5	18.1/19.7
R-free[d]	22.1/23.2	21.1/23.2	20.2/23.6	20.5/22.7	20.0/22.0
Water molecules	153	250	264	227	291
Bond length (Å)	0.005	0.005	0.005	0.005	0.005
Angles (°)	1.2	1.2	1.2	1.2	1.2
Torsions (°)	21.5	21.3	21.2	21.9	21.2
Torsions_{imp} (°)	0.8	0.8	0.8	0.8	0.8

Table 3.1: Data collection and refinement statistics for the 5 high-resolution structures of inhibitors complexed to human DHODH. [a] AU, asymmetric unit. [b] and [c] The second values refer to the highest resolution shell. [d] R-free (test-set), identical to R-factor, but calculated for 5 % of the refractions omitted from the refinement for cross validation.

connected by a short loop, whereas an extended loop connects C- and N-terminal domain. It is also this domain that has been proposed to assemble with the inner mitochondrial leaflet upon recruiting ubiquinone to recover FMNH2[48].

As commonly seen in most high-resolution structures, some parts of the protein appear disordered[48]. This is particularly true for residues 69 to 71, part of the extended loop that appears disordered in all structures, and residues 216 to 221 in some of the structures. Also, the HIS-tag preceding the N-terminal MET30 is disordered in all structures. The preceding amino acid sequence is thought to act as a signalling sequence upon recovering FMNH2. However, subsequent bioassay conditions used by Baumgartner and coworkers did use a truncated version of the protein, and hence the determined affinities are not influenced[48]. In figures 3.6, showing a front view of the protein, and 3.7, showing the protein in a side view thus allowing for a better recognition of the α-β barrel fold, the smaller N-terminal domain is colored navy blue, while the large C-terminal domain is composed of the remainder.

The slot, zoomed in with the protein viewed in a bottom orientation and displayed

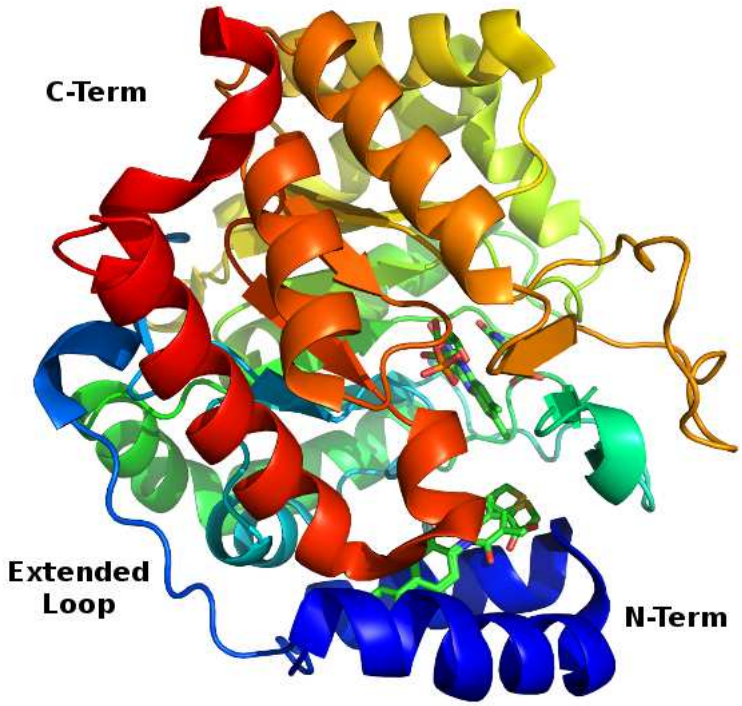


Figure 3.6: Front view of DHODH showing the dual binding mode inhibitors, the cofactor and the substrate in stick representation, and protein in cartoon. Pictures have been created using Pymol [155].

in figure 3.8, forms the entrance to the tunnel that ends at the FMN moiety. It is characterized by the presence of a number of charged or polar side-chains, i.e. Gln47, Tyr356, Thr360 and Arg136, and ends near the loop segment connecting the two helices. This slot contains the binding site of ubiquinone and thus it is also the binding site for the type of inhibitors that are the subject of this study. The distribution of amino acids forming this binding site reflects its biochemical function. The entrance to the tunnel almost exclusively consists of hydrophobic amino acids, which agrees with the fact that helices $\alpha 1$ and $\alpha 2$ are involved in anchoring the protein to the membrane for the purpose of recovering FMNH2, while polar side chains at the end of the tunnel harbour the natural substrate ubiquinone.

3.6.3.3 Biological affinities

The biological affinities for the compounds have been determined in a consistent manner. The reaction was followed spectrophotometrically by measuring the de-

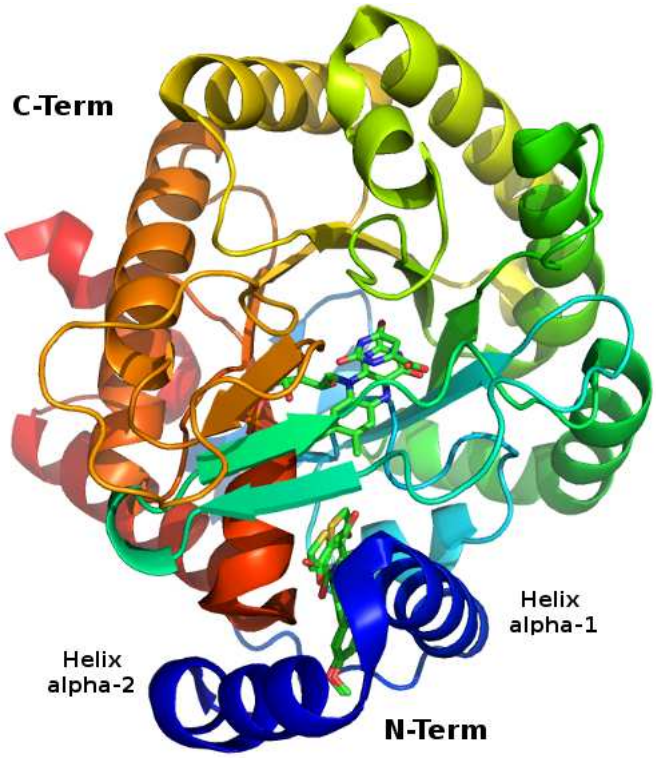


Figure 3.7: Sideview of DHODH showing the dual binding mode inhibitors, the cofactor and the substrate in stick representation, and the protein in cartoon. Pictures have been created using Pymol [155].

crease in absorption at 600 nm for 2 min. The assay was linear in time and enzyme concentration. Inhibitory studies were conducted in a standard assay with additional variable amounts of inhibitor. For the determination of the IC50 values (concentration of inhibitor required for 50% inhibition) eight different inhibitor concentrations were applied. Each data point was recorded in triplicate on a single measurement day and resulted in IC_{50}s at 280, 33, 7, 44 and 2 nM for compounds 3, 4, 5, 6 and 7 respectively (with the structures displayed in figure 3.9.

3.6.3.4 Crystallographic details for compound 3

Compound 3, shown in figure 3.10, was refined to a resolution of 2.15 Å and biological activity was determined with an IC_{50} of 280 nM by assay conditions described in 3.6.3.3. Conserved key residues for ligand binding to DHODH, i.e ARG136 and GLN47, were causing unexpected issues. ARG136 seemed to display two well-

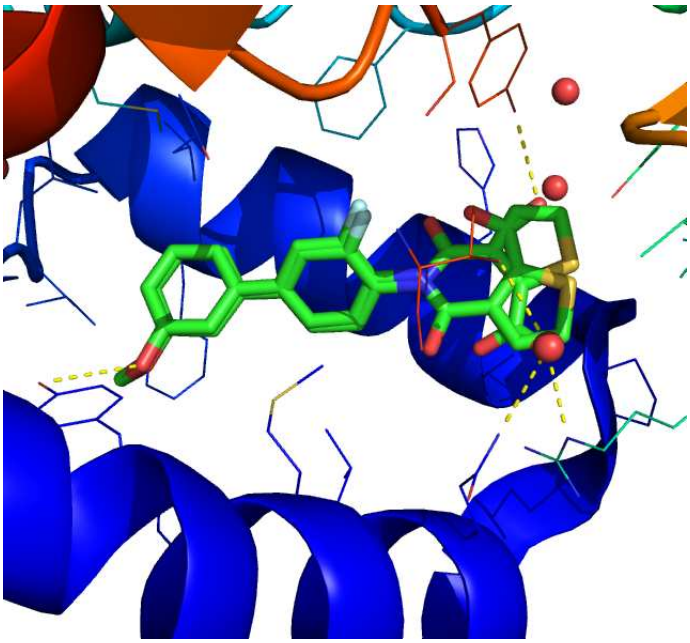


Figure 3.8: The slot formed by the two N-terminal helices α 1 and 2 complexed with compound 6 in both proposed modes of binding.

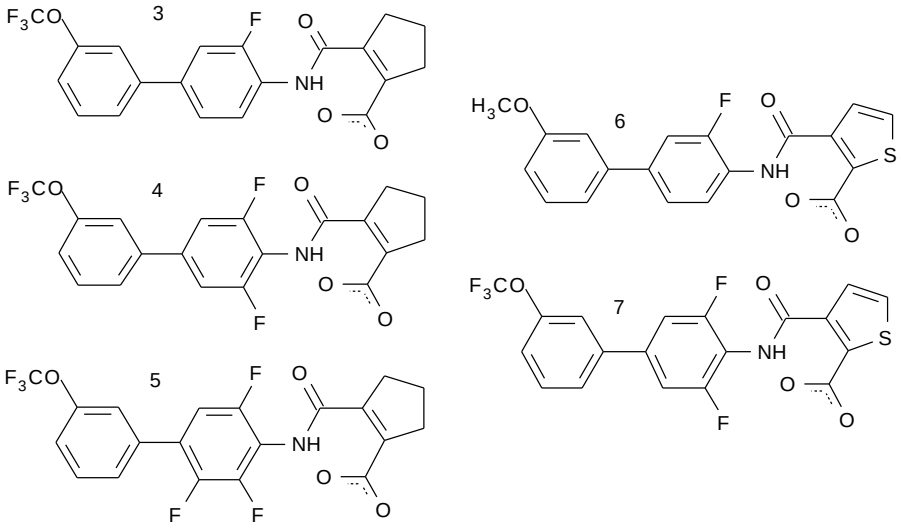


Figure 3.9: The structures of the inhibitors of the Baumgartner series.

resolved side chain conformations, an extended conformation and a solvent exposed conformation, with B-factors for both not significantly above the mean value (mean, 22.5 Å^2; ARG136-extended, 25.6 Å^2; ARG136-solvent exposed, 26.1 Å^2). For protein conformational change this would result in a conformation where ARG136 covers the

top of the binding site, i.e. extended conformation that modulates ligand interaction, while it could also protrude into the solvent and as such not mediate any interactions with the ligand. GLN47, according to the authors, can thus only be interpreted in the context of ARG136, and no 'excellent' density exists for this residue. Two possible scenarios are being proposed, one in which ARG136 is rotated towards the solvent, hence GLN47 can be modelled into the remaining density, and another where ARG136 is built into the extended conformation, and thus would clash with GLN47 which would subsequently have to be rotated away towards the solvent, although no density is visible for this area[48].

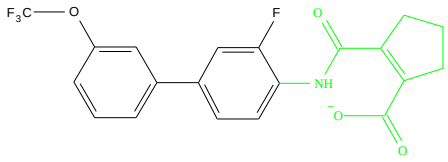


Figure 3.10: Structure of compound 3.

This might complicate the analysis of the binding mode of compound 3. According to the authors the inhibitor revealed a rather unexpected binding mode. In contrast to the brequinar (BQ) binding mode, the carboxy moiety is positioned in the opposite direction protruding towards the interior of the protein. This binding mode has been termed 'nonbrequinar' (NBQ) in the following and can be thought of as a 180° rotation around the F-C-C-N dihedral, i.e. the entire moiety that is colored green in figure 3.10 is rotated. The authors provide the final Fo-Fc electron density map, i.e. the density resulting from the difference of observed minus calculated densities, refined with compound 3 omitted at 2.15 Å resolution as a representative example. This is shown in figure 3.11.

More specifically, the carboxylate moiety at the five-membered ring protrudes into the protein's interior and interacts with residues TYR256 and TYR147 by the formation of hydrogen bonds. The hydrogen bond to TYR256 is formed directly, whereas the hydrogen bond to TYR147 is mediated by a water molecule. The five-membered ring is located in a plane with the adjacent amide bond, and an intramolecular hydrogen bond between the amide and the carboxy moiety is formed. Residue HIS56 is interacting with the fluoroquinolone ring of Brequinar according to Liu et al., while for this compound the imidazole ring of HIS56 is rotated away by almost 70° from

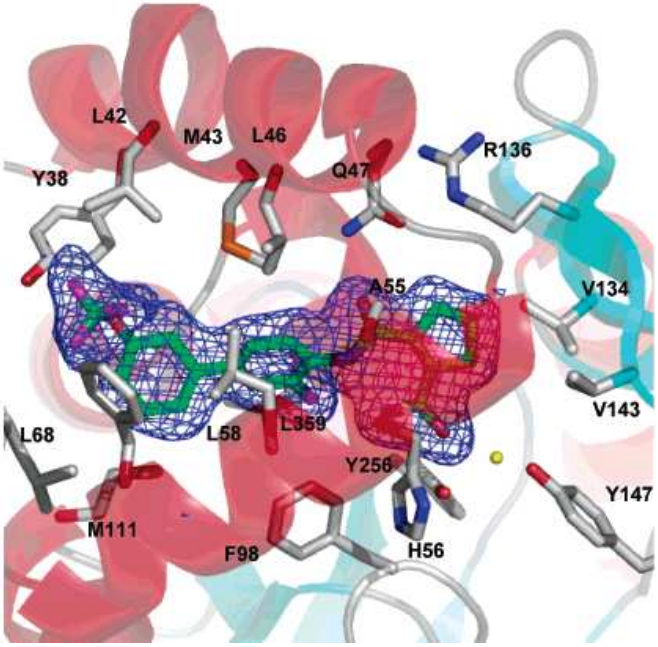


Figure 3.11: Electron density map for compound 3. The solute, and key residues are shown in stick together with the crystallographic water in stick. Picture has been taken from[48].

its brequinar-like position pointing toward the carboxylate group of the inhibitor molecule. The imidazole ring can be placed either with the Nδ1 atom or the Cδ1 atom toward the inhibitor. Since a water molecule on the opposite side of the imidazole ring is within hydrogen-bonding distance to Nδ1 (distance 2.6 Å), the latter orientation was suggested by the crystallographer. The bi-phenyl ring system occupies most of the hydrophobic pocket in a manner similar to that of Brequinar. The two aromatic rings are inclined to each other by approximately 70°, a value closely resembling that of brequinar. The suggested binding mode of compound 3 is represented in figure 3.12 and an overlay of both binding poses, i.e. compound 3 and Brequinar, are shown in figure 3.13.

3.6.3.5 Crystallographic details for compound 4 and 5

The biological affinities for compounds 4 and 5, shown in figure 3.14(a) and 3.14(b) respectively, were determined at 33 nM and 7 nM, pointing towards the preferred lipophilicity for ligands to be accommodated in the DHODH ubiquinone binding pocket. In the structures of both compounds ARG136 and GLN47 do not appear

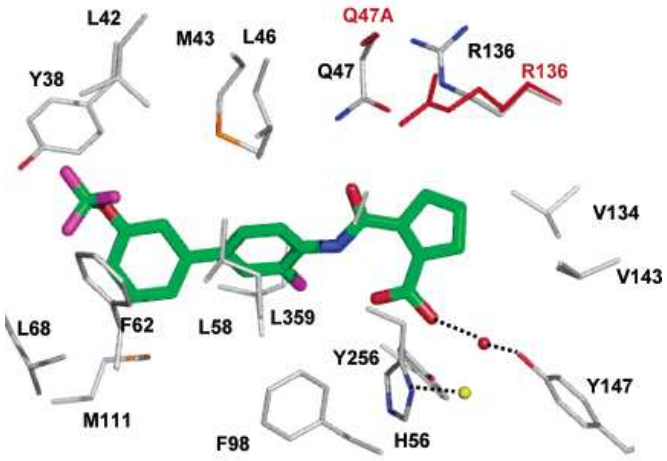


Figure 3.12: Suggested mode of binding for compound 3 by Baumgartner and coworkers. Here compound 3 is shown in stick representation and key residues, including crystallographic waters are shown in line representation. The figure has been taken from [48].

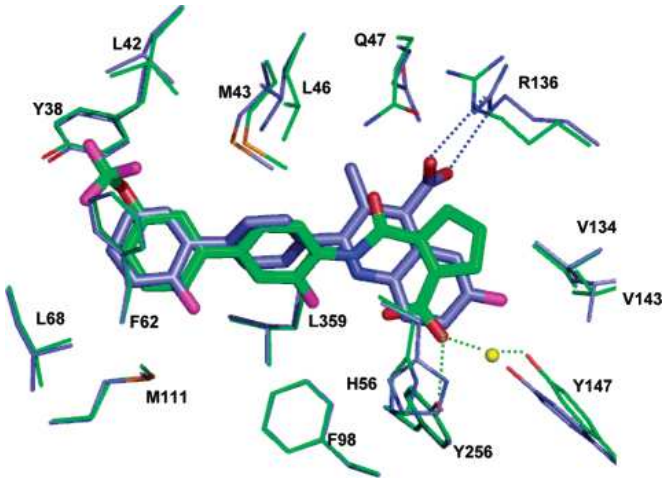


Figure 3.13: Superimposition of the suggested binding mode of compound 3 (in green) with the mode adopted by Brequinar (in purple). This figure has been taken from [48].

disordered or present in an alternate conformation.

However, for both compounds, electron densities were less detailed and more extended, in particular for the 5-membered ring of the inhibitors, than one would expect for the level of resolution, i.e. 1.95 and 2.4 Å respectively. In fact, residual density in the difference maps of observed and calculated electron densities, i.e. Fo-Fc maps,

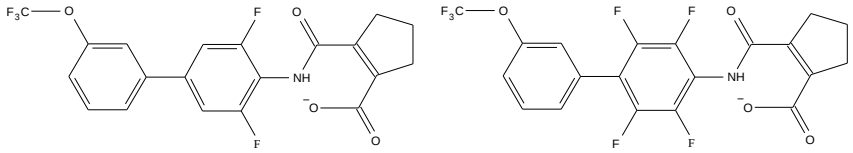


Figure 3.14: Compounds 4 (a) and 5 (b) of the DHODH series developed by Baumgartner et al[48].

indicated missing model input. Figure 3.15 taken from[48] highlights the electron densities that cannot be explained if either the brequinar or nonbrequinar binding mode was assumed for model refinement. However, the density appeared big enough to accommodate both compounds 4 and 5 in both conformations, and seems to reappear, after refinement, where atoms would normally appear for the brequinar mode and vice versa. This is illustrated in figure 3.15 and led the authors to postulate the possible existence of both binding modes for inhibitors 4 and 5. As for the case of the structure of compound 3, water molecules were found to mediate interactions between ligand and protein residues, i.e. the TYR147 interaction is bridged by a water molecule in the non-brequinar binding mode, while HIS56 adopts a similar orientation as observed for compound 3.

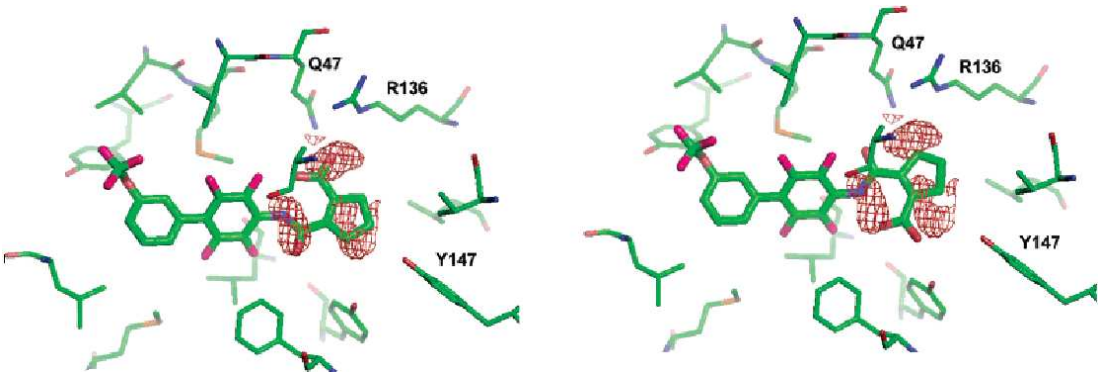
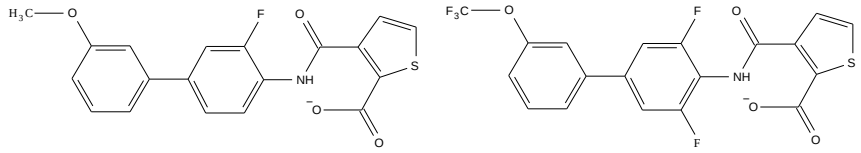


Figure 3.15: Missing model input for compounds 4 and 5 after refinement (both in green). The red areas indicate the unexplained electron densities, when a brequinar mode, i.e. left picture, or the non-brequinar mode, i.e. right, was assumed and the model was redefined. The figure has been taken from[48].

3.6.3.6 Crystallographic details for compounds 6 and 7

The biological activities for compounds 6 and 7 were determined at 44 and 2 nM respectively. Both compounds differ from compounds 3, 4 and 5 by the replacement of the cyclopentene ring for a thiophene ring and are shown in figures 3.16(a) and 3.16(b).



(a) (b)

Figure 3.16: Compounds 6 (a) and 7 (b) of the DHODH series developed by Baumgartner et al.[48].

The crystal structures were solved at 1.8 Å and 2.0 Å for compounds 6 and 7 respectively. While the electron densities of compound 6 appeared less ordered and more extended, this was not observed for compound 7. Again, the authors postulate a mixed type binding mode for compound 6, while compound 7 seemed to be more brequinar like according to its electron density and refinement. This is highlighted in figure 3.17. Neither compound 6 or 7 show alternate conformations of ARG136 and GLN47, and hence a considerable amount of structural similarity has been observed for the structures for the entire series, but not for compound 3. The water molecule bridging the interaction with TYR147 has been resolved by the crystallographer for compound 6 but not for compound 7. Thus, the interaction pattern does not significantly deviate from compounds 3, 4 and 5, apart for an additional site attracted via the introduction of the thiophene ring. The sulfur atom of the thiophene is in close contact to a small hydrophobic pocket formed by the side chains of VAL134 and VAL143, thus contributing an additional interaction to the binding affinity that is not present in the other molecules.

3.6.3.7 Consensus of binding motifs using all compounds and binding modes

Finally, according to the specific nature of the protein-ligand interactions one can identify a number of subsites. If we were to assume that all binding modes proposed by Baumgartner and coworkers exist, then based on the crystallographic data alone, a consensus binding site can be imagined. According to the work of Baumgartner

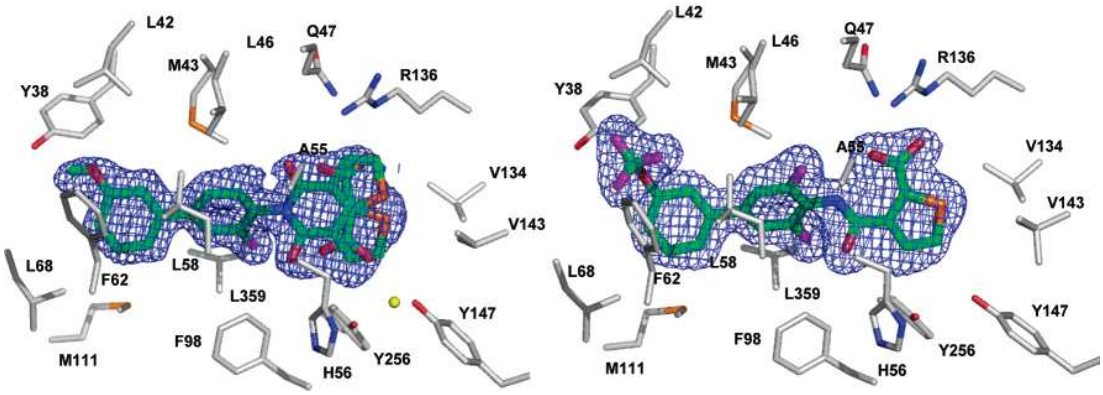


Figure 3.17: The picture on the left shows the final Fo-Fc electron density in blue, calculated with compound 6 omitted, and contoured at 3σ. Two molecules of 6 are built into the density (in green), one in a brequinar conformation and one in a nonbrequinar conformation. The image on the right is the drawing of compound 7 bound into the DHODH inhibitor binding site, modelled in its brequinar conformation. The figure has been taken from[48].

this consensus is illustrated in figure 3.18 and highlights the different subsites encountered in DHODH using an overlay of both binding modes. The biological data presented by Baumgartner would suggest, that compounds adopting the brequinar conformation exhibit higher affinity than compounds adopting the non-brequinar mode.

It is evident that each subsite that can be identified includes functional groups capable of forming stabilizing interactions with complementary functional groups of a possible inhibitor molecule. The distribution of amino acids forming the binding site is quite clearly demarcated. The entrance to the tunnel almost exclusively consists of hydrophobic amino acids. This agrees with the fact that helices α-1 and α-2 are involved in membrane association. The narrow end of the tunnel forms a rather polar environment capped by a small hydrophobic pocket formed by side chains of Val134 Val143.

The potential lead-like inhibitors presented here clearly match this amphipathic character of the binding site and show IC50 values between 7 and 280 nM in vitro - competitive assay with ubiquinone. The majority of the binding site is made up by a hydrophobic site, i.e. subsite 1 and 2 in 3.18, which is adequately occupied by the bi-phenyl ring of the inhibitors. Subsites 3 and 4 in 3.18 can be addressed by functional groups capable of forming hydrogen bonds. For the inhibitors in question this is enabled by the carboxy group attached to the 5-membered ring and by the

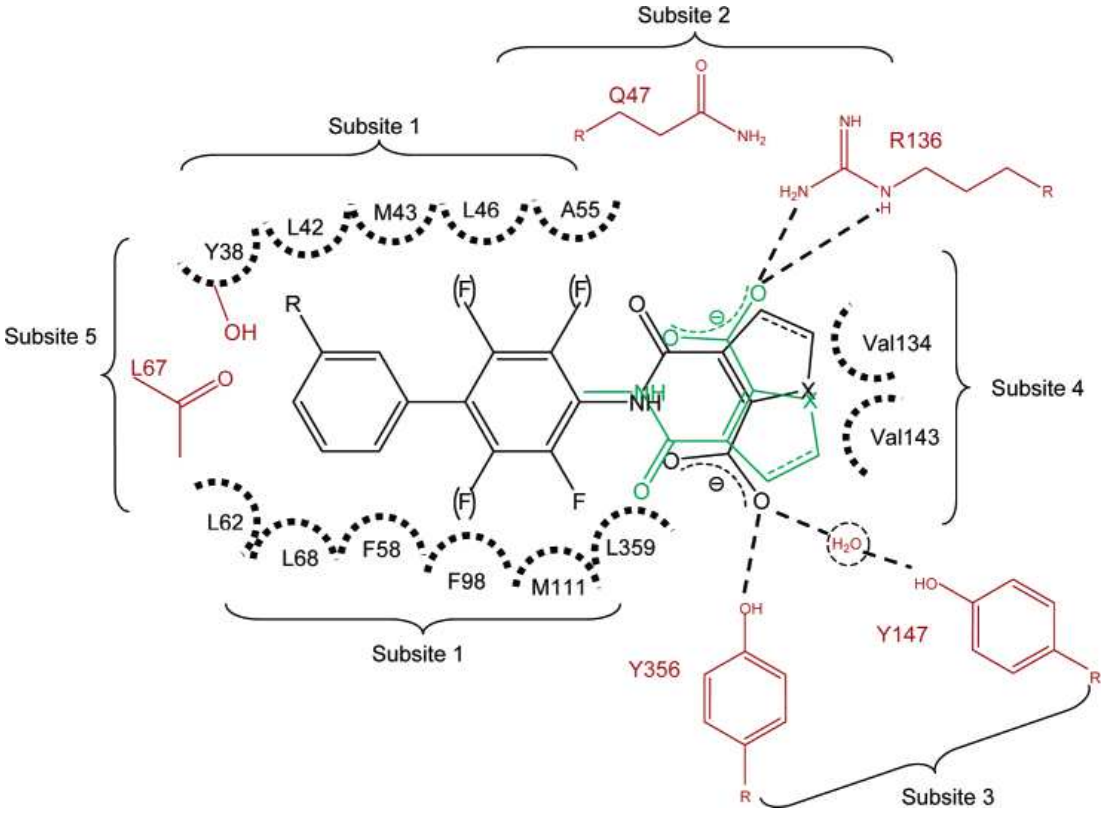


Figure 3.18: Schematic diagram of the DHODH inhibitor binding site representing several subsites, labeled 1-5, suitable for interacting with an inhibitor molecule. Black dotted circles depict residues contributing to hydrophobic interactions; amino acids contributing to electrostatic interactions are highlighted in red. The drawing of the inhibitor molecule represents a consensus molecule, with differing degrees of fluorine substitution, indicated by (F), a variable substituent R (-OCF3 or -OCH3), and a heteroatom X (C or S) at the five-membered ring. The alternative conformation of the inhibitor is depicted in green. Possible hydrogen bonding is indicated by black dashed lines; amino acids are given in a one-letter code plus sequence number. This image has been taken from[48].

proposed ability to form a dual binding mode. Additionally, the binding pocket is capped by the more remote hydrophobic site 4 and compounds 6 and 7 utilize subsite 4 for hydrophobic interaction via the sulfur of the thiophene ring. Baumgartner also proposed a fifth subsite for binding, predominately polar in nature, consisting of the hydroxy group of TYR38 and the backbone carbonyl of LEU62. This subsite is in close proximity to the trifluoromethoxy (or methoxy group in compound 6) of the inhibitors. Although interactions by these groups have not been observed, it appears reasonable to design future inhibitors containing functional groups able to utilize subsite 5.

3.6.4 How to study the Baumgartner series using rigorous free energy methodologies

The series of inhibitors developed by Baumgartner and coworkers together with the generation of their high-resolution structures is a striking and outstanding example of exceptional beauty for the development of novel small molecules in structure-based drug design. We are used to looking at congeneric series of compounds as simple homologous series, while aiming to reach stages in the development process where we 'hop' from one ligand scaffold to another, and have biological affinities readily available. One cannot deny that this 'scaffold-hopping' is and will be a major goal for the development of new drug. However, the findings published by Baumgartner, together with the crystallographic studies published by Shoichet on fragment-like ligands on T4 lysozyme, are just two examples of the very peculiar ways proteins recognize ligands and vice versa[19]. Surely, we are much less likely to assume that two inhibitors of substantially different structure bind to a protein in the same way, than we were to expect a dramatic change in binding modes by the simple addition of a single fluorine atom that is 'recognised' by non-directed hydrophobic interactions with the protein. So should we abandon the concept of congenerity altogether?

Are we focusing on artificial and misleading experimental evidence? The study on DHODH clearly demonstrates the current limit of crystallographic refinement. The missing model input presented by Baumgartner is the best solution that could be found using a certain refinement procedure. However, can experiment tell with certainty that actually one certain binding mode exists, or probably even both binding modes for ligands 4, 5 and 6? Are the structures resolved with a single binding mode more accurate than the ones were a dual binding mode was detected? The authors in fact do not go further than saying there 'might' be dual binding modes. We believe it is a reasonable question to address the issue of hydration of the binding site in this context. For 5 essentially identical protein structures, i.e. all-protein-atom RMSDs of less than 0.5 Å have been recorded, 153, 250, 264, 227 and 291 water molecules have been 'resolved'. Additionally, can the refinement procedure account for structural water molecules mediating protein-ligand interactions and clearly distinguish diffraction of oxygens that are part of carboxylate or carbonyl moieties from those of waters? So what is the best way to proceed with the Baumgartner series, if one

wants to attempt to elucidate the binding mode and structure activity relationships in DHODH quantitatively?

We believe the answers to these questions are clearly rigorous free energy simulations. If the system setup, sampling and force field criteria have been met, then they not only provide us with a rigorous and accurate physical basis, but they also allow us to study the dynamics in more detail than any other computational approach, and hence in principle even allow us to refine crystal structures and answer these final questions on structure activity relationships in DHODH, while at the same time provide an excellent testcase for the methods used. Indeed, crystal structure refinement has already been subject of high level QM/MM approaches published by Ryde and coworkers [156].

To work out a suitable free energy protocol for this target we are aiming for a stepwise approach, trying to capture all problematic aspects one could potentially encounter. The actual perturbations to be covered are essentially based on the growth of fluorine atoms, unless binding modes are to be perturbed. Halides, the dray-horse for typical ligand alterations in drug design due to their convenient atomic properties, should not pose serious problems as far as force field parameters are concerned. However, to identify the most suitable and accurate force field for our perturbations we will perform hydration free energy studies using different sets of force field parameters as well as using higher levels of theory to describe the energetics, i.e. QM/MM. These studies will be presented in the next chapter, i.e. chapter 4.

Another important aspect for defining structure activity relationships in DHODH are structural water molecules. A clear change in hydration pattern has been proposed by the crystallographer. However, it is not clear as to whether the waters are an artefact of refinement, or indeed mediating ligand binding. Recent developments applying a form of λ-dynamics have been recently proposed by Michel and coworkers, i.e. the JAWS algorithm, to assess hydration patterns prior to simulations, and shown to improve subsequent free energy simulation results [125]. This is not surprising as in principle waters may be an essential element of thermodynamic end states. Another route to answer this question lies in methods that attempt to insert and delete particles using an appropriate ensemble. This type of method is currently being developed in our lab by M. Bodnarchuk; hence we make use of his work and

show in chapter how hydration pattern can be defined more rigorously while we confirm the findings for both approaches for critical water molecules using the rigorous double-decoupling method, proposed by McCammon and coworkers and also successfully applied on a variety of protein-ligand systems by members of our group in the past.

These preliminary studies help us to define our system more reliably and are subject of the follwoing two chapters, while the final chapter concludes on the usefulness of our approach in understanding structure activity relationships in DHODH.

4

Free energy of hydration

‘

Aqueous solvation, or hydration, is of critical importance in all biochemical processes. Thus, proper accounting for hydration is an active area of research for the prediction of the binding of small molecules to proteins. In principle, this process involves the desolvation of part or all of both the protein and binding ligand. In modern docking and scoring algorithms desolvation is now included in terms of correction factors, aiming to estimate this effect. As an example, it has been shown that the inclusion of correction factors improve the correlation of docking scores with experimental binding affinities and aid in eliminating molecules with inappropriate charge states from a set of potentially high-affinity inhibitors[157,158], while without these corrections, highly-charged ligands do not experience a desolvation penalty when they are transferred from solvent to the potentially low-dielectric of a protein binding site.

As explained in chapter 2, relative binding free energies between two solutes, S1 and S2, for the same protein can be calculated by morphing the first solute into the second. This is illustrated in figure 4.1, a representative thermodynamic cycle for the calculation of relative free energies of binding with computer simulations. It is clear from this illustration that the alchemical transformation must be performed while the solutes, S1 and S2, are bound to the protein (the bound leg, complexed to the protein shown by green squares in figure 4.1), and while the ligands are free in solvent (the free leg, shown in blue). Hence, solute binding can be seen as a competition of two solutes between the protein binding site and the solvent. In other words, when we ask the question of which of a pair of solutes binds best to a protein, we are often ask

which solute shows greater affinity for the protein and lower affinity for the solvent, because lower free energies of hydration often favour binding.

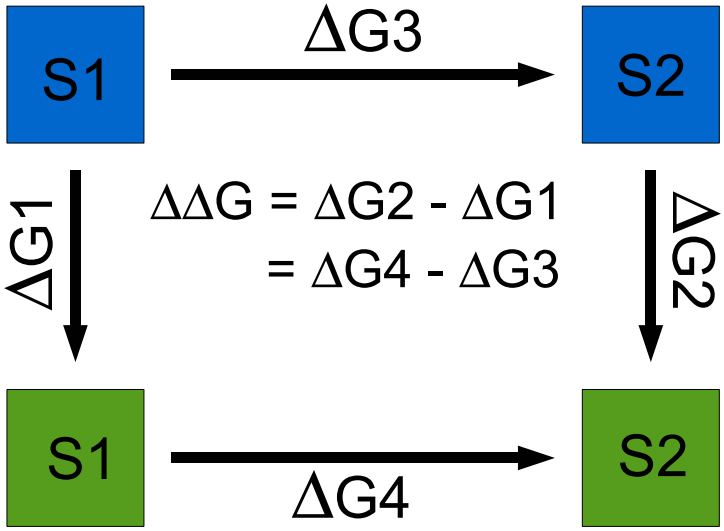


Figure 4.1: Thermodynamic cycle used to calculate relative binding free energies using computer simulation and alchemical free energy methods.

Equally, one can calculate free energies of hydration using this cycle. If the green and blue squares in figure 4.1 correspond to the water and gas phase respectively then the double free energy difference is the relative free energy of hydration. Theoretically, the free energy of hydration corresponds to the reversible work associated with transferring a solute from the gas phase to a water phase. This process can be understood as a natural consequence of the versatility of water to respond to molecules of varying size and polarity and one may envision this process by further subdividing the hydration process into a series of reversible coupling steps, i.e. sequentially introducing a repulsive core (cavity), van der Waals (dispersion), and electrostatic (multipolar/ionic) interactions[159]. If we were to alchemically transform a solute into nothing, the absolute hydration free energy would be obtained. Thus, hydration free energies are an essential component in binding free energy studies, and are related to physical properties of interest in drug discovery, such as the solubility and water-octanol partition coefficient, and their study by computer simulation can provide fundamental insight into how water organises around solute molecules.

4.1 Accuracy and precision in classical hydration free energy studies

The study of hydration free energies is not only important for capturing the desolvation penalty associated with ligand binding - although it will of course have a direct impact on the calculated free energies of binding - but it is also evident that hydration free energy studies can offer greater insight into the method itself that was employed to calculate this property. It allows one to assess accuracy and precision, as the changes in the degrees of freedom in a hydration free energy study will be easier to capture than those of binding free energy studies, and hence their noise to signal ratio will be better. A lack of accuracy, i.e. the discrepancy between the model used for the simulation and the experimentally measured reality, and a lack of precision, i.e. whether all thermally relevant contributions to the ensemble average of the observable of interest have been sampled sufficiently in a simulation, may occur and will give further insight as to why a hydration free energy study has failed to agree well with experiment. Precision issues often result in calculated free energies dependent on random number seed, input structure as well as the amount of sampling achieved. Accuracy issues often result in precise calculated free energies but they fail to agree with experiment. Standard errors for rigorous free energy studies, however, usually stay below the 1 kcal mol^{-1} level.

While a study of Mobley and coworkers has shown that precision can become a serious issue for compounds with carboxylic acids[25], attributed to the slow sampling of conformational changes, it is generally assumed and has been shown that free energies of hydration can be computed to high precision[160]. However, their accuracy is sensitive to a number of force field parameters, and as such, hydration free energies have been used not only to predict experimentally measured free energies of hydration on large datasets, but to take up the challenge in large scale validation studies, where experimental values have not been known prior to simulation, i.e. the SAMPL test. These extensive studies have provided a good benchmark for the performance of force fields to predict hydration free energies, and allow an estimate of their performance in a binding free energy study.

4.2 Current generation force fields

Current generation biomolecular force fields such as AMBER99[69], CHARMM22[70] and OPLS-AA[67,68], are successful in many cases for capturing protein structure and dynamics, although constant improvements are being presented, such as the correction for dihedrals for GLY in the AMBER99 force field and resulting in the AMBER99SB force field[129]. Force fields for small molecular weight compounds need to address a much bigger chemical space due to their diversity. A number of general force fields for these type of molecules exist, including GAFF[136], CHARMM22[70] and OPLS[67,68], but there are no accepted standards for their generation, particularly for atomic charges. For example, AMBER94[161] and AMBER96 i[162] have charges derived based on fitting of the electrostatic potential from self-consistent field (SCF) HF/6-31G* calculations[65], after which Lennard-Jones parameters were fitted, while CHARMM22[70] charges come from fitting solute-water dimer energetics from SCF HF/6-31G* calculations, after which Lennard-Jones parameters have been fitted. OPLS-AA[67,68] has traditionally derived charges and Lennard-Jones parameters from fitting to pure liquid properties, such as transfer free energies, rather than quantum mechanical calculations, although more recently charges based on the semi-empirical CM1 method that are scaled by 1.14[163] for neutral molecules were proposed. Finally, charges for the widely used and popular GAFF[136] force field for small molecules, are usually derived using the ANTECHAMBER package[164], which allows no less than seven different charge models to be used.

It is important to remember that when performing a simulation, to simply specify a set of parameters is not enough to define a protocol. In fact, anything that affects the Hamiltonian changes the model used and will affect the results obtained. For example, a large number of parameters need to be defined, such as the definition of boundary conditions, the truncation and detailed handling of non-bonded interactions, as well as constraints on bonds or other degrees of freedom, which will all result in a modified Hamiltonian, while each force field has been parametrised using very certain assumptions and simulation procedures. A violation of these could potentially render force field parameters not transferable, resulting in inaccuracies that would be falsely attributed to the performance of the force field.

Whatever force field is being used, there is likely to be a fundamental limit to the accuracy due to the neglect of explicit polarisation and the adoption of simple functional forms used to describe atomic interactions. On the other hand, force fields are being developed that incorporate a polarisation term, e.g. the AMOEBA force field[134], but examples are still rare and too few studies have been published to allow a thorough validation of their usefulness and applicability[165]. Additionally, in case a force field parameter for a certain molecule cannot be found, then an elaborate procedure is required to strictly follow the initial force field parametrisation, in order to produce a new set of parameters for the particular molecule in question. Hence, approaches that allow the generation of force field parameters on the fly, or that could correct potential flaws inherent in force fields, i.e. lack of polarisation, would be of great value. One way of achieving this goal is to combine classical mechanical simulations with higher levels of theory, i.e. quantum mechanics[66].

4.3 Quantum Mechanics Molecular Mechanics (QMMM) methods

The foundations of the QMMM method have been laid out as early as 1976 by Warshel and Levitt[166], and have become a valuable tool not only for modelling biomolecular systems, but also for inorganic/organo-metallic[167,168] and solid state systems[169], and for studying processes in explicit water[170,171]. The general idea of a QMMM scheme is shown in figure 4.2.

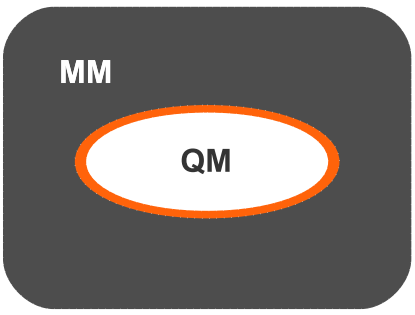


Figure 4.2: Partitioning of a system into an inner subsystem, shown as a white ellipsoid, and treated with a QM method, and the outer subsystem, the black area, treated in a classical way, with the boundary regions shown in red.

4.3.1 Technical dissection of the principles of QMMM methods

According to figure 4.2, a system is divided into subsystems, and each is treated using a different level of theory, i.e. QM or MM. It is important to highlight that any QM method can be coupled to any MM procedure. Between these subsystems, a boundary may be defined, where MM and QM procedures are being augmented in some way[66]. Depending on the QMMM scheme used, this boundary region may be defined and augmented in different ways, i.e. one may define link atoms that are neither part of the MM nor the QM system, or it may be defined in terms of atoms that appear in both the QM and the MM calculation. Although less common, the boundary region can also be of dynamical nature, hence change during the course of a simulation[172].

The calculation of a QMMM energy using a *subtractive scheme* requires an MM calculation of the entire system, a QM calculation of the inner subsystem, as well as a MM calculation of the outer subsystem[66]. I, II, S and L shall denote the inner and outer subsystem, the entire system as well as the linker atom(s) respectively. The entire QMMM energy is then calculated according to

$$E_{QM/MM}(S) = E_{MM}(S) + E_{QM}(I + L) - E_{MM}(II + L) \tag{4.1}$$

To avoid double counting of the inner subsystem I, $E_{MM}(II + L)$ is being subtracted, giving this scheme its name. This subtraction corrects for artefacts caused by the link atoms, L, as long as the MM force field terms involved in the link atom reproduce the QM potential reasonably well. In other words, a certain region of a system is cut out and treated quantum mechanically. The advantages of this approach lies in its simplicity: no explicit QM-MM coupling terms are necessary, and standard MM and QM procedures can be applied, making room for an easier implementation. However, if force field parameters are missing for the inner subsystem, or even the outer subsystem, then no parameters can be generated on the fly, and one has to derived them according to the original protocols depending on the force field used. Additionally, the coupling between the MM and the QM system is treated entirely on an MM level, which may be particularly problematic for electrostatic interactions, as they will be represented by the fixed atomic charges in the QM and MM regions. Examples of subtractive QMMM schemes are the IMOMM method[173], i.e. integrated molecular orbital/molecular mechanics method by Morokuma and co-workers, and

derivatives thereof such as the extended IMOMM method[174] that allows for the combination of two QM methods and further generalized to n layers (typically n=3), each of which can be treated at an arbitrary QM or MM level, i.e. ONIOM[175].

On the contrary to the subtractive scheme, an *additive QMMM scheme* may be applied, and QMMM energies calculated via[66]:

$$E_{QM/MM}(S) = E_{MM}(II) + E_{QM}(II + L) + E_{QM-MM}(II, I) \tag{4.2}$$

In contrast to the subtractive scheme, here the MM calculation is performed on the outer subsystem only. In addition, an explicit coupling term, $E_{QM-MM}(II, I)$, collecting the interaction terms between the two subsystems is introduced, and the capped inner subsystem is (I+L) is treated at the QM level as in the subtractive scheme. The majority of QMMM schemes make use of this type. It is the exact form of the QM-MM coupling term E_{QM-MM} that defines a particular QMMM method. In accordance with the interactions considered in the force field, it usually includes bonded, van der Waals, and electrostatic interactions between QM and MM atoms:

$$E_{QM-MM}(II, I) = E^{bond}_{QM-MM} + E^{vdW}_{QM-MM} + E^{ELE}_{QM-MM} \tag{4.3}$$

4.3.1.1 Electrostatic QM-MM Interaction

Different levels of sophistication can be used to account for the electrostatic coupling between the QM charge density and the charge model used to represent the MM region. The extent of their mutual polarisation classifies these methods into mechanical, electrostatic and polarised embedding[176,177]. For the case of mechanical embedding, the QM-MM electrostatic interaction is treated identically to the MM electrostatics, i.e. the charge model of the MM method is simply applied to the QM region as well. Although conceptually straightforward and computationally efficient, this method comes with several disadvantages. First, the charges of the outer region do not interact with the QM density, and hence no direct polarisation of the outer region occurs. Second, if the charge distribution in the QM region changes, the MM charges therein should ideally be changed, which would cause discontinuities in the potential energy surface. Third, the generation of MM charges for the QM region may not be trivial. Finally, individual MM charges need not be physically meaningful, as long as the force field provides an overall balanced description.

However, major shortcomings of this mechanical embedding can in principle be overcome by an electrostatic embedding. In an electrostatic embedding, the QM calculation is performed in the presence of the MM charge model[66], i.e. by incorporating the MM point charges as one-electron terms in the QM Hamiltonian:

$$H^{el}_{QM-M} = -\sum_{i}^{N}\sum_{j\epsilon\Phi}^{L}\frac{q_j}{|r_i - R_j|} + \sum_{\alpha\epsilon II+L}^{M}\sum_{j\epsilon\Phi}^{L}\frac{q_j Q_\alpha}{|R_\alpha - R_j|} \tag{4.4}$$

The symbols q_j are the MM point charges located at R_j; Q_α are the nuclear charges of the QM atoms at R_α; and r_i designate electron positions. The indices i, j, and α run over the N electrons, L point charges, and M QM nuclei, respectively. In electrostatic embedding schemes the electronic structure of the inner region can adapt to changes in the charge distribution of the environment and is automatically polarized by it. It is also advantageous that no charge model needs to be derived for the inner region, as the QM-MM electrostatic interaction is treated at the QM level, thus providing a more advanced and accurate description than that found for a mechanical embedding scheme. Obviously, electrostatic embedding also increases the computational demands, especially for the calculation of the Coulomb forces as a result of the QM density acting on the (many) MM point charges. The QM-MM boundary can however be problematic, where the MM charges are placed in immediate proximity to the QM electron density and can thus cause over-polarization, in particular when the boundary runs through a covalent bond.

On the downside of electrostatic embedding, remains the general issue of compatability between the MM charge model and the QM electron density. As the electrostatic MM parameters are not primarily designed to provide a reliable and accurate representation of the real charge distribution, it is not strictly legitimate to stitch a true charge distribution from a QM calculation into the carefully parametrized MM charge model. Nevertheless, this has become common practice, and reports in current literature show that results are generally reasonable, at least for the combination of a QM density with one of the widely used biomolecular force fields[66]. The fact that MM atomic partial charges are readily available and that their inclusion in the QM Hamiltonian is efficient makes electrostatic embedding the most popular embedding scheme in use today, certainly for biomolecular applications.

As electrostatic embedding accounts for the interaction of the polarizable QM density with fixed MM charges, the next logical step is to introduce a flexible MM charge

model that is polarized by the QM charge distribution. These polarised embedding schemes can be divided into approaches where the polarizable charge model in the MM region is polarized by the QM electric field but does not itself act back on the QM density, or fully self-consistent formulations that include the polarizable MM model into the QM Hamiltonian and therefore allow for mutual polarization. As outlined above, there exist few models for treating polarization in classical simulations, but there are as yet no generally established polarizable biomolecular force fields, certainly not that would allow the description of arbitrary molecules for drug design purposes, and hence biomolecular applications have remained scarce; a notable exception is a QMMM study of excited states in the photosynthetic reaction center in which the MM polarization was included self-consistently[178]. Apart from this, polarised embedding QMMM calculations have essentially been restricted to explicit solvation (in particular, hydration), where the solute is treated at the QM level and the solvent by a polarizable force field[179,180]. An interesting approach has recently been proposed by Zhang and coworkers, aiming to polarise only the MM atoms at the boundary[181].

It is clear that an accurate description of the electrostatic forces on the QM subsystem arising from the environment is essential for a realistic modeling of biomolecules, while the inclusion of all electrostatic interactions explicitly is computationally challenging. Additionally, simple QMMM electrostatic cutoffs may be problematic because of the long-range nature of Coulomb interactions. Although the reliable and efficient treatment of these long-range electrostatic interactions is well-established in classical simulations, it has only recently attracted attention in the context of QMMM methods, i.e. linear-scaling particle-mesh Ewald schemes for QMMM simulations under periodic boundary conditions[182] as well as a charge-scaling procedure using continuum electrostatics where only a limited number of explicit solvent molecules is considered and the charges are scaled to mimic the shielding effect of the solvent[183].

4.3.1.2 van der Waals QM-MM interactions

Apart from the electrostatic interactions that are clearly of major importance in QMMM methods, there are also van der Waals and bonded contributions to the QM-MM coupling term. However, their treatment is considerably simpler as they are usually handled purely at the MM level. The van der Waals interaction is typically

described by a LennardJones potential so that suitable parameters are needed for the QM atoms in the inner region. Often they are adopted from similar atom types. Furthermore, even if suitable LennardJones parameters exist for a given configuration, QM atoms can change their character, for example during a reaction, which then raises the question of whether the parameter set should be switched from a reactant description to a product description somewhere along the reaction path. Practice shows that complications arising from an inconsistent treatment of van der Waals interactions in the QM-MM coupling term are mainly alleviated by the short-range nature of the van der Waals interaction, and hence only atoms close to the boundary contribute significantly. Thus, possible errors due to unreasonable Lennard-Jones parameters may be minimized by moving the QM-MM boundary further away from the incriminated QM atoms, unless this is not appreciated. For example, Friesner and coworkers have re-optimised the QM Lennard-Jones parameters against QM data for hydrogen-bonded pairs of small amino acid models[184]. The Lennard-Jones radii thus obtained using their approach are 5-10 % larger than those of the underlying force field, i.e. OPLS-AA in this case, while Lennard-Jones well depths were left unchanged. The increased repulsion induced by the modified parameters, compensates for the too strong QM-MM electrostatic attraction which arises from over-polarisation at the boundary. Additionally, a set of Lennard-Jones parameters optimized for B3LYP/AMBER was presented by another group[185]. However, studies to date seem to indicate that thermodynamic quantities in the condensed phase, such as free energies, calculated from QMMM simulations, are rather insensitive towards these QM-MM van der Waals parameters, while, as one might have expected, some influence on the detailed structure around the QM region was observed.

4.3.1.3 Bonded QM-MM interactions

For intramolecular degrees of freedom, such as bonded terms, i.e. bond-stretching, angle-bending and torsional degrees of freedom, similar reservations against using standard MM parameters to describe QM-MM interactions apply. Again, the solution is entirely pragmatic: usually the standard MM parameter set is retained and is complemented as necessary with additional bonded terms not covered by the default assignment rules of the force field[66].

4.3.2 QMMM methods for the calculation of free energies

It it obvious that the use of higher levels of theory on parts of a biomolecular system is advantageous: an improved and more realistic physical description can better represent important physical phenomena, such as polarisation, charge transfer, and bond breaking and formation, hence enabling the study of enzymatic mechanisms. It is also obvious that an improved physical description requires substantially more computer time, and a more complex implementation to make QM calculations available. However, QMMM studies in the context of predicting free energies and binding affinities have proven particularly useful, for example in docking, scoring as well as in combination with classical rigorous free energy simulations[39,51,66,186,187].

The substantially higher computational demand is particularly challenging for free energy simulations, as it prevents the evaluation of large ensembles of conformations typically required to successfully converge the averages in a free energy simulation. However, a variety of methods have been developed within the free energy simulation framework to overcome this barrier. Highly promising methods involve the use of fast and approximate Hamiltonians to sample phase space, with the full QMMM Hamiltonian being used sparingly to calculate or estimate free energies using only a limited subset of the generated configurations[66].

4.3.2.1 QMMM using fast and approximate Hamiltonians

Two main classes of these methods are of particular relevance for free energy simulations: The first group of methods use a fast Hamiltonian to estimate the relative free energy, and then use efficient algorithms to approximate the difference in free energy between the QMMM representation and the fast Hamiltonian[188]. The second group of methods use the fast Hamiltonian as a means of enhancing sampling of the QMMM Hamiltonian, thereby producing ensembles correct for the QMMM Hamiltonian that can be fed directly into free energy methods such as free energy perturbation (FEP)[189,190].

The first class of methods has been pioneered by Warshel and coworkers[188] and applies the free energy cycle shown in figure 4.3. Here, an approximate reference Hamiltonian, i.e. either a pure MM Hamiltonian or an empirical valence bond, is used

to estimate the free change, and the estimate is subsequently corrected by calculating the energy necessary to change from the approximate Hamiltonian to a QMMM Hamiltonian. To make the method more efficient all of the sampling is performed using only the reference Hamiltonian, and the free energy difference between the reference and QMMM Hamiltonians is calculated from a single-step FEP perturbation from the ensemble generated at the reference state. Theoretically, this will return the exact free energy change, but in practice may suffer from poor convergence owing to the large fluctuations of the difference between reference and QMMM Hamiltonians. Hence, these methods focus on the derivation of approximate reference Hamiltonians that are a good match to the target QMMM Hamiltonian, and the use of the linear response approximation to improve convergence by running on both the approximate reference Hamiltonian and on the QMMM Hamiltonian.

Similarly, the quantum mechanical thermodynamic cycle perturbation (QTCP) has been developed by Rod and Ryde[187,191]. A reaction pathway using QMMM and a selected number of configurations for the QM region along the reaction pathway are defined. Based on calculated point charges for the QM region, classical MD sampling is performed and classical MM-QM interaction free energy changes between subsequent fixed QM configurations along the reaction pathway are determined. Similar to the approach by Warshel and coworkers, the MM-QM free energy change for each QM configuration along the reaction pathway is calculated, and, in this way, a high-level QMMM PMF may be obtained. This was demonstrated for the methyl transfer reaction in catechol O-methyltransferase (COMT), showing converged PMF behaviour. As the QM region was fixed in the implementation of the QTCP method, well converged free energy barriers were observed.

Instead of estimating the relative free energy, the approximate Hamiltonian may also be used to accelerate phase space sampling described by a QM or QMMM Hamiltonian. This second class of methods produces ensembles that are correct for the QM or QMMM Hamiltonians used, while the ensembles can be used directly with free energy methods such as FEP. This approach has been pioneered by Schofield and coworkers[189]. The sampling is performed by using Monte Carlo methods with a Metropolis-Hastings algorithm[192]. The algorithm, shown in figure 4.4, applies an approximate Hamiltonian to guide the Monte Carlo sampling of the QM or QMMM Hamiltonian and a speed-up is achieved by sampling the majority of Monte Carlo

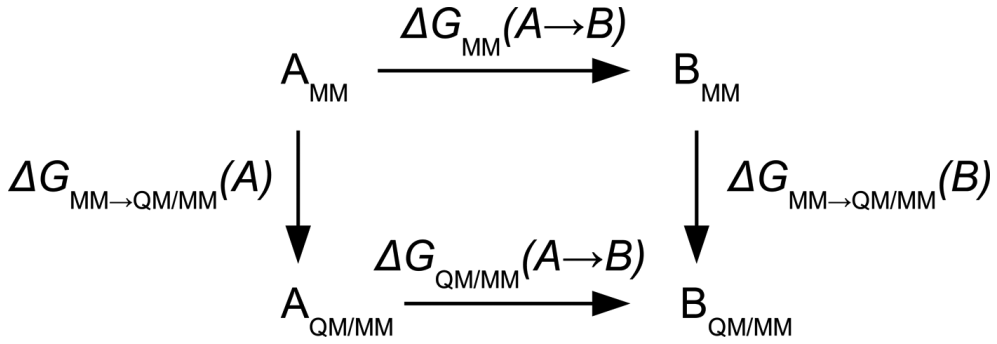


Figure 4.3: The free energy cycle used to calculate the QMMM free energy difference between systems A and B, $\Delta G_{QM/MM}(AB)$. The free energy difference between A and B is first estimated using an approximate potential (e.g., MM potential), giving $\Delta G_{MM}(AB)$. This is then corrected to the QMMM value by calculating the free energy necessary to perturb system A from MM to QMMM $[\Delta G_{MM\rightarrow QM/MM}(A)]$ and the free energy to perturb system B from MM to QMMM $[\Delta G_{MM\rightarrow QM/MM}(B)]$. This figure has been taken from[52].

moves using only the approximate Hamiltonian, while the form of the Monte Carlo acceptance test provides an ensemble that is rigorously correct for the QM or QMMM Hamiltonian. The method was popularized for application to Monte Carlo sampling of QM and QMMM Hamiltonians by Iftimie and coworkers who call this method molecular mechanics based importance function (MMBIF)[189].

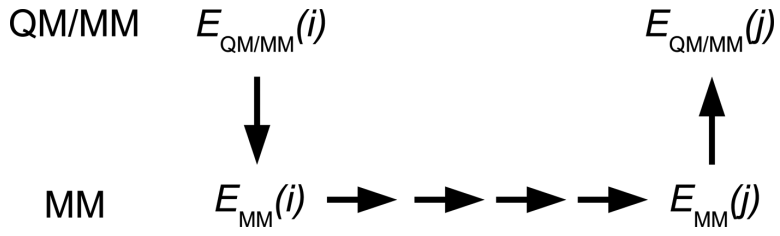


Figure 4.4: Application of the Metropolis-Hastings algorithm to accelerate sampling of a system represented using a QMMM Hamiltonian. The Monte Carlo move starts at configuration i. The energy of this configuration is evaluated using the target QMMM Hamiltonian [giving $E_{QM/MM}(i)$] and on an approximate (MM) Hamiltonian [giving $E_{MM}(i)$]. Standard Metropolis Monte Carlo moves are then attempted from configuration i using only the approximate MM Hamiltonian, until after a set number of moves, the system is in configuration j. The energy of configuration j is evaluated using both the QMMM and MM Hamiltonians [giving $E_{QM/MM}(j)$ and $E_{MM}(j)$]. Configuration j is then accepted into the QMMM ensemble according to the probability $min1, exp(-\Delta\Delta E/kBT)$, where $\Delta\Delta E$ = $[E_{QM/MM}(j) - E_{MM}(j)] - [E_{QM/MM}(i) - E_{MM}(i)]$. This figure has been taken from[52].

One of the drawbacks of the methods of class one are sampling problems, due to large fluctuations in the difference between approximate and QMMM Hamiltonian

resulting in poor convergence. This is similar for class two methods. The efficiency of the MMBIF method critically depends on the level of phase-space overlap between the QMMM and MM Hamiltonians. If the overlap is poor, then the probability of accepting each Metropolis-Hastings Monte Carlo move will be low, and convergence of the free energy averages will be poor. As in the case for the first class of methods, effort may therefore address the optimisation of the approximate MM Hamiltonian so that it is a better match for the desired QMMM Hamiltonian. Alternatively, Iftimie and coworkers have demonstrated how parallel tempering, i.e. replica exchange methods, can be used by adding an additional degree of freedom to the system that enhances sampling[189].

4.3.2.2 A novel QMMM free energy simulation protocol

The Metropolis-Hastings algorithm is generally applicable and can be applied to any situation where an expensive potential can be approximated by a cheap potential. Hence, it is continually rediscovered[193], and it has been applied to accelerate Monte Carlo sampling of implicit solvent force fields[45] as well as for the Ewald sum for long range electrostatics[194]. Woods and coworkers have used the power of the Metropolis-Hastings algorithm together with the QMMM thermodynamic cycle method proposed by Warshel, see figure 4.3, to form an efficient and in principle exact QMMM free energy cycle[52]. Here, they calculate the free energy difference between two systems, A and B, using a pure MM Hamiltonian. Rather than estimating the correction free energies using a single-step FEP perturbation, as proposed by Warshel[188] , they instead make use of the Metropolis-Hastings algorithm, as described as part of the MMBIF method in the previous section, see figure 4.4. Correction free energies are calculated using the Hamiltonian

$$H = (1 - \lambda)H_{QM/MM} + \lambda H_{MM} \tag{4.5}$$

and simulations are then run at different values of λ, while the correction free energy is calculated using TI,

$$\Delta G_{QM/MM\rightarrow MM} = \int_0^1 \left(\frac{\delta G}{\delta \lambda}\right)_\lambda d\lambda \tag{4.6}$$

$$= \int_0^1 \left\langle \frac{\delta H}{\delta \lambda} \right\rangle_\lambda d\lambda \tag{4.7}$$

$$= \int_0^1 \left\langle H_{MM} - H_{QM/MM} \right\rangle_\lambda d\lambda \tag{4.8}$$

Hence, the λ coordinate is used to scale the QMMM Hamiltonian into the MM Hamiltonian. As the MM Hamiltonian is also used as the approximate potential for the Metropolis-Hastings algorithm, so as λ is increased, H becomes closer to H_{MM}. The acceptance probability of the Metropolis-Hastings Monte-Carlo moves is subsequently increased with an increase of λ as a logical consequence. Additionally, replica exchange moves are being applied across the λ coordinate and will therefore enhance QMMM sampling, i.e. in a manner identical to the simulated tempering moves used by Iftimie and coworkers[189], but since λ is provided naturally by TI, no additional exchange coordinate is required. The use of replica exchange moves to enhance sampling along the λ coordinate provided by TI was first described in the replica exchange thermodynamic integration (RETI) free energy method and has shown to enhance sampling and reduce statistical error in MM free energy calculations[40]. For QMMM corrections RETI may bring similar advantages.

4.4 Selecting an optimal set of force field parameters

The discussion so far makes clear that one can in principle select from a considerable number of force field parameters, or, when polarisation is expected to have a big impact on the binding event, may also switch from a classical description towards methods that include quantum mechanics. The study of current literature also demonstrates that most biomolecular and widely accessible force fields, such as AMBER[69], GAFF[136], OPLS-AA[67,68] and CHARMM[70] 'usually' perform considerably well and it seems that this choice is almost arbitrary. It is important, however, to keep in mind that the performance of force field parameters is also very much dependent on the system studied, as well as the initial simulation conditions used to generate the respective parameters.

The alchemical transformations encountered for the Baumgartner series are indeed incredibly simple, and as such may prove rather simple when one is to select an appropriate set of force field parameters, i.e. the growth of fluorine atoms on an aromatic moiety. Additionally, the chosen force field should allow the description for the prosthetic group FMN as well as for the endogenous substrate ORO. For this reason, we have aimed to test some of the most widely used force fields, i.e. GAFF and OPLS-AA, for the prediction of free energies of hydration. The novel QMMM method proposed by Woods and coworkers, see section 4.3.2.2, seems to be an attractive tool to improve free energy estimates[52]. However, more studies are necessary to validate this method, and hence part of this chapter is focused on finding out how much benefit a hydration free energy study could potentially gain. Therefore, we have selected a small but representative series of small molecules to select an appropriate force field to best capture our perturbations on the DHODH system. This series consists of fluoro-, chloro- and bromobenzene, as well as aniline, toluene and nitrobenzene.

4.4.1 Generalised Amber Force Field (GAFF)

GAFF[136] uses 33 basic atom types and 22 special atom types to cover almost all the chemical space composed of H, C, N, O, S, P, F, Cl, Br, and I and is based on quantum mechanical calculations, mainly MP2/6-31G*, together with crystallographic data. The basic atom types, all the bond length, bond angle, and torsional angle parameters are available or may be calculated with empirical rules, and special atom types were introduced to describe certain chemical environments accurately, such as conjugated single and double bonds. GAFF determines parameters for all combinations of atom types algorithmically for each molecule based on the bonding topology and the geometry provided by the user. Although this procedure is fully automated in the ANTECHAMBER suite[164] of the AMBER programs, and allows one to specify a complete set of parameters, i.e. atom types, charges as well as force field parameters, it is not without errors and the convenience of its availability should be combined with a thorough evaluation of the accuracy of results generated. GAFF is an extension of the AMBER force fields, particularly parametrized for most of the organic molecules one may encounter, and it is a complete force field by itself.

To accurately fit conformational and non-bonded energies in a transferable fashion a consistent charge should be considered. Even though a total of seven different

methods of generating atomic charges are currently available using ANTECHAMBER, they have not all been parametrised with GAFF. The restrained electrostatic potential (RESP) at HF/6-31G* is the default charge approach[50] applied in the Amber protein force fields. As RESP requires the input of quantum mechanical calculations, thus being computationally expensive, it has been of limited use in the community. However, the alternative AM1BCC charge model[49] is much cheaper and has shown to essentially reproduce RESP charges, and hence AM1BCC is one of the most widely used charge models for biomolecular simulations. The main idea of AM1BCC is to carry out a semi-empirical AM1 calculation, providing Mulliken charges, followed by a bond charge correction scheme, i.e. BCC, to obtain charges comparable to RESP, as the correction is designed to make AM1BCC charges match the electrostatic potential at the HF/6-31G* level.

To answer the question of whether these methods of charge generation in conjunction with the GAFF parameters do result in different free energies of hydration, several studies have been conducted by Mobley and others[27,132]. The charge sets considered were either based on semi-empirical calculations, i.e. AM1/CM2 and AM1BCC, as well as RESP fitting applying different levels of theory, such as HF/6-31G*, B3LYP/6-31G*, B3LYP/cc-pVTZ and MP2/cc-pVTZ, both with and without a reaction field treatment[63] of the solvent and using two different water models, i.e. TIP4P-Ew. and TIP3P[195]. Interestingly, the water model and the treatment of electrostatics did not yield any systematic difference in the predicted free energies of hydration, nor did a higher level of theory in generating the charges result in more accurately predicted free energies of hydration. Indeed, AM1BCC performed as well as RESP procedures using higher theory levels, and results indicate that the accuracies achievable are within 1 kcal mol^{-1} of the experimental value. Hence one could find errors of similar magnitude for the prediction of the free energies of binding. A large scale study on 504 small molecules published by Mobley[132], and using GAFF in combination with AM1BCC charges, gave rise to the identification of systematic errors in the Lennard-Jones terms for alkynes in GAFF, which have subsequently been modified, but overall underline the accuracy and reliability in predicting free energies of hydration for small and fragment-like molecules using free energy simulations.

4.4.2 Optimised Potentials for Liquid Simulations (OPLS-AA)

In contrast to GAFF, which aims to reproduce quantum mechanical properties, OPLS potential functions have been parametrised directly to reproduce experimental thermodynamic and structural data on fluids[67,68]; hence their description of the condensed phase should be superior to many alternatives that have been developed with limited condensed-phase data, such as AMBER[69], while their functional form is identical to the one used in AMBER or GAFF. However, a major limitation of OPLS is the necessity of having experimental data to generate parameters, and hence parametrisation could potentially even involve the generation of experimental liquid properties prior to actually generating parameters for the OPLS force field. Moreover, several versions of the OPLS force field exist, and it is often not clear which set of parameters have been used for published results. Moreover, the programs MCPRO as well as BOSS[196], both MC codes that allow the use of the OPLS force field, internally choose the most similar parameter if an exact match for the atom types provided for a molecule cannot be found.

The OPLS-AA force field comes, in principle, with a complete set of atomic charges for proteins, small molecules and others. Much effort has gone into the development and testing of alternative procedures for obtaining optimal fixed charges for nonpolarizable force fields for any organic molecule. To achieve this, a dominant route has been to perform a quantum mechanical calculations and to fit the charges to reproduce electronic properties, especially dipole moments and electrostatic potentials, which has led to popular alternatives such as RESP charges, and faster approaches, i.e. AM1 with bond corrections.

In 1995, Storer et al. reported the particularly attractive CM1A and CM1P approaches that provide high-quality partial charges from rapid, semi-empirical AM1 and PM3 quantum calculations[197]. For these methods optimised mapping procedures resulted in highly accurate dipole moments for organic molecules and further refinements have led to the introduction of Charge Model 3 (CM3), which encompasses a larger training set in conjunction with both semi-empirical (AM1 and PM3) and density functional (BLYP and B3LYP) methods[198]. In previous testing of charge models based on AM1 and PM3 for the OPLS-AA force field the CM1A charge model has ranked amongst the best, and optimal scaling factor have been determined

through charge perturbations[163]. Thus AM1CM1A charges scaled by 1.14 for neutral compounds have been proposed to be a good choice for the free energy simulation using OPLS-AA parameters.

4.4.3 Classical free energy studies of hydration

Starting geometries for the solutes studied here were created using molden[199]. The solutes were setup with GAFF[136], and the atomic partial charges were derived using the AM1BCC method[49] as implemented in the program ANTECHAMBER[164]. An initial minimisation was performed using the Sander module of AMBER8[200] and a generalized Born force field. The final geometries were then subject to a charge recalculation, providing charges that are according to GAFF geometries, and yielding the final set of force field parameters for the combination of GAFF with AM1BCC charges. The solutes and their charges obtained are shown in figure 4.5.

The generation of RESP charges followed the protocol of R.E.D. proposed by[201]. According to this procedure, RESP and ESP charge derivation is performed in three steps and the resulting charges are shown in figure 4.6:

1. Geometry optimization: According to the GAFF protocol this step has been performed on the HF level using the 6-31G* basis set

2. Molecular Electrostatic Potential (MEP) calculation (using Gaussian's reorientation algorithm)

3. Fitting the charges centered on the atoms to the MEP calculated in step 2

AM1CM1A charges for the simulations using the OPLS-AA parameters were obtained via single point energy calculations on the semi-empirical AM1 level, as implemented in the program AMSOL[202]. Initial geometries for these calculations were supplied as OPLS-AA minimised structures, and the resulting charges were subsequently scaled by 1.14, as suggested by Jorgensen and coworkers for neutral molecules[51].

All charges obtained, i.e. AM1BCC, RESP and AM1CM1A, were finally averaged for identical atoms. In fact, to makes charges for many equivalent atoms identical by averaging is common practice in classical simulations and avoids potential nonphysical charge asymmetries. For example, if charges on hydrogens in a methyl group are not the same, the energy of the system can be different for identical rotamers,

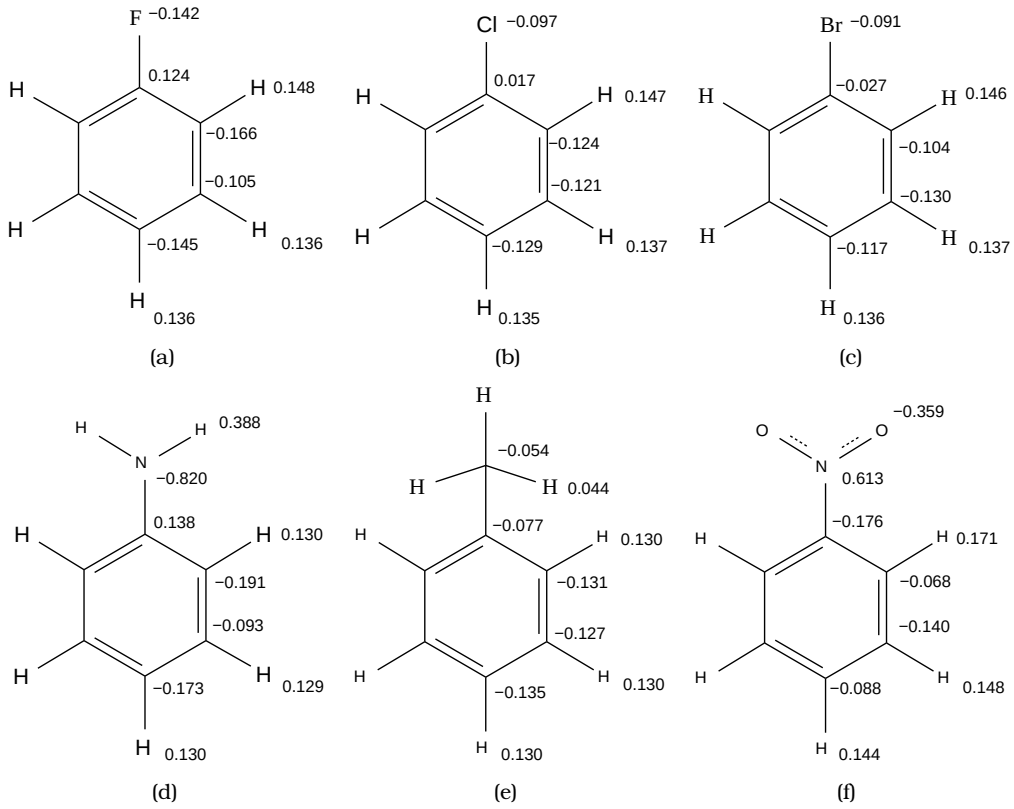


Figure 4.5: AM1BCC charges for fluorobenzene (a), chlorobenzene (b), bromobenzene (c), aniline (d), toluene (e) and nitrobenzene (f) generated using ANTECHAMBER and subsequent averaged to account for symmetrical charge distributions.

although it was shown previously that the averaging has little effect on computed free energies of hydration for neutral molecules.

The solutes were hydrated in a box of TIP4P water molecules [195] of the dimensions 25x25x25 Å generated using the program packmol [203] and subsequent equilibration was conducted using the in-house software ProtoMS 2.1 [47] allowing 200M moves for the 531 TIP4P water molecules contained in this box. To investigate the influence of the off-centre charge in the TIP4P water model, a similar box was constructed using the TIP3P water model [195], resulting in 530 TIP3P water molecules. The final snapshot was then used for subsequent RETI MC simulations. To calculate the relative free energies of hydration benzene was selected as a reference state; hence the solutes formed the perturbed states. To ensure reliable convergence, each perturbation was distributed over 16 evenly spaced values of the coupling parameter λ. Equilibrium

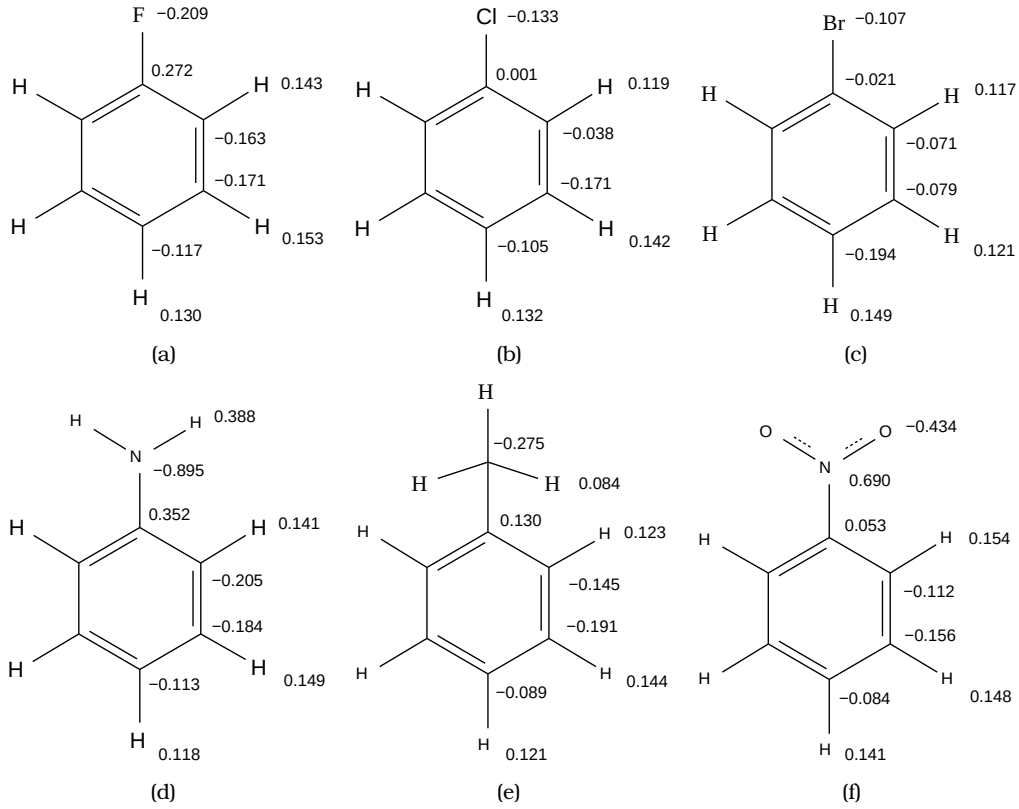


Figure 4.6: RESP charges obtained by R.E.D. for fluorobenzene (a), chlorobenzene (b), bromobenzene (c), aniline (d), toluene (e) and nitrobenzene (f).

configurations were generated in the isothermal-isobaric ensemble at 20°C and 1 atm using the Metropolis MC algorithm and preferential sampling was performed[72]. Periodic boundary conditions and a non-bonded cutoff of 9 Å which was quadratically feathered to zero over the last 0.5 Å were employed. At each of the λ windows, the system was equilibrated for 10M moves followed by 40M moves where data was collected. In the gas phase, data was collected over 5M moves. In the water phase, solute moves were attempted every 100 configurations and volume changes every 2500 configurations. The range of translations and rotations for the solutes was set between 0.2 and 0.5 Å and 5 and 10° respectively, to allow for an acceptance probability for the solute moves of approximately 30%.

It is common practice to compute hydration free energies for rigid solutes, i.e. no sampling of internal degrees of freedom is performed. These '*single conformation hydration free energies*' shall, according to findings of Mobley and coworkers[29], not be

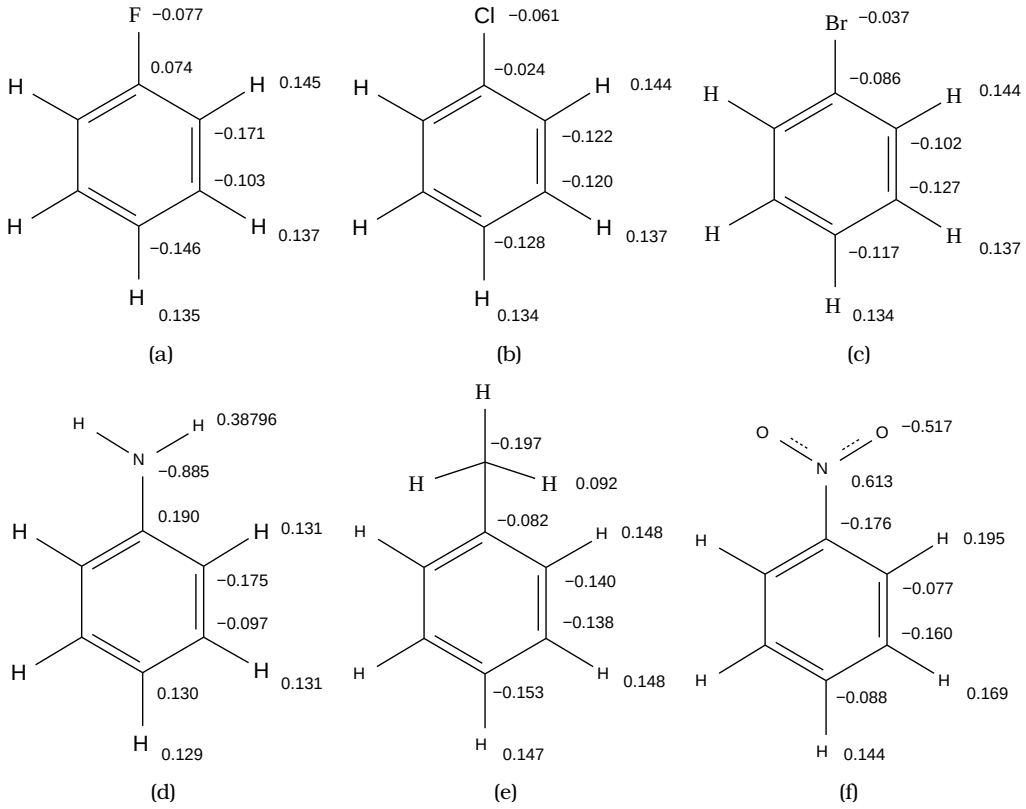


Figure 4.7: AM1CM1A charges scaled by 1.14 for fluorobenzene (a), chlorobenzene (b), bromobenzene (c), aniline (d), toluene (e) and nitrobenzene (f) generated using ANTECHAMBER and subsequent averaging to account for symmetrical charge distributions.

confused with *hydration free energies*, that allow full flexibility of the solutes's internal degrees of freedom. It has been shown that indeed flexibility can have effects on the computed hydration free energies, although for the solutes that are the subject of this study there is little concern regarding the influence of flexibility on the computed energies. Still, for the combination of AM1BCC/GAFF both options have been considered. For all other parameter sets only fully flexible solutes were considered, where bond angles and torsions were sampled with the exception of rings.

The relative hydration free energies were calculated using RETI. Here, relative hydration free energies were calculated for all solutes with the reference being benzene. A finite difference scheme was used to calculate the gradients with $\Delta\lambda = 0.001$ and the Zwanzig equation was used to calculate the free energies $\Delta G(\lambda + \Delta\lambda)$ and $\Delta G(\lambda - \Delta\lambda)$. For the 16 values of λ, the integral was then numerically estimated using trapezoidal

numerical integration.

4.5 QMMM studies of free energies of hydration

The novel QMMM free energy method described in section 4.3.2.2 was validated by calculating the relative hydration free energies of the solutes that were also subject in our classical studies. The setup conditions were kept identical and final snapshots of the classical simulations were used as a starting point for subsequent QMMM studies: solvent box with dimensions 25x25x25 Å, periodic boundary conditions, a 9Å molecule-based electrostatic and non-bonded cutoff feathered quadratically to zero over the last 0.5Å.

Two Hamiltonians were required to represent this system; an approximate MM Hamiltonian and the target QMMM Hamiltonian. In the QMMM Hamiltonian, the solute was represented using QM, while the surrounding solvent was represented using MM. The QM solute was modelled using BLYP/6-31G*, so including the effects of electron correlation. The electrostatic interaction between the QM and MM regions was modelled using the established method of including, within the QM Hamiltonian, the MM point charges, i.e. electrostatic embedding. The electrostatic cutoff was applied by including only point charges from atoms that were in solvent molecules within 9 Å of any QM atom (including periodic boundaries). Quadratic smoothing of the electrostatic interactions was applied by scaling down the point charges of affected atoms using the smoothing function applied over the last 0.5 Å. The van der Waals interaction between the QM and MM regions was modelled using the Lennard-Jones potential, which was also subject to the same 9 Å molecule-based cutoff with 0.5 Å smoothing function. In the MM Hamiltonian, the QM solute was approximated by a MM solute.

All Monte Carlo simulations were run in the isobaric-isothermal ensemble, at 293.15 K, and 1 atm pressure. Preferential sampling was used to enhance sampling around the solute, using a preferential sampling constant of 200 Å. Monte Carlo moves were chosen at random and solvent moves, solute moves and volume moves were attempted 98 %, 1.9 % and 0.1 % of the time. Solute and solvent moves consisted of random, rigid-body translations and rotations, with a 0.15 Å maximum

translation and 15.0° maximum rotation for the solvent molecules, and a 0.2 Å maximum translation and 2.5° maximum rotation for the solute. The volume moves changed the volume of the solvent box by a maximum of 53.1 Å^3. The QM solutes were kept flexible, and the gas-phase QM energy of the solute was subtracted from the QMMM energy, thereby removing the QM intramolecular energy of the solute from the calculation.

The Metropolis-Hastings Monte Carlo algorithm, described in the previous section, was employed to generate long QMMM trajectories using the SIRE Monte Carlo program[204], which used MOLPRO[205] to perform the QM energy calculations. 49.9M moves of the entire system using the approximate MM potential and 100K QMMM energy evaluations were attempted for four evenly spaced values of λ from 0 to 1. The first 1M moves from each simulation were discarded as equilibration, while the collected averages reported here correspond to the remaining 49M moves. The approximate MM Hamiltonian used the same MM water model as the solvent to approximate QM water, while GAFF was used to approximate the QM solutes.

The QMMM correction free energies required by the Warshel cycle in figure 4.3 are calculated using QMMM Monte Carlo. For correction free energies, RETI simulations were performed and the gradients were calculated as described previously, and these were used to calculate the free energies via TI. RETI moves were attempted every 50000 MC moves and the choice to swap even or odd pairs of replicas was made randomly at each RETI move.

The Warshel cycle also requires the calculation of the relative hydration free energy of the solutes, as estimated by the MM Hamiltonian. The MM relative hydration free energies were calculated using RETI, over 16 λ windows spaced evenly across a λ coordinate. All of the simulations were performed using PROTOMS 2.1[47], using the same protocol described in section 4.4.3 was used with the GAFF force field and AM1BCC atomic partial charges. Relative MM hydration free energies were calculated for all solutes with the reference being benzene. A finite difference scheme was used to calculate the free energy gradients, with $\Delta\lambda = 0.001$, and the Zwanzig equation was used to calculate free energies $\Delta G(\lambda + \Delta\lambda)$ and $\Delta G(\lambda - \Delta\lambda)$. After free energy gradients were calculated at 16 values of λ, the integral was numerically estimated using trapezoidal numerical integration.

4.6 Hydration free energy results

The results of this study are a set of calculated relative free energies of hydration for substituted benzenes using different parameter sets as well as by applying a novel QMMM method. When evaluating force field models, one is usually torn between using recent simulation methodology with the best proven scientific validity, the methods used for the actual parametrization of the fore field model, and the most commonly used methods in the literature, which may all be somewhat different. Here, we combined AM1BCC[49] and RESP[50] charges with GAFF[136] and AM1CM1A charges[51], scaled by 1.14, with OPLS-AA[67,68]. At the same time we try to agree with original simulation protocols for parametrisation; in particular we apply periodic boundary conditions, account for different combining rules[63] in OPLS-AA versus GAFF, and use a finite-ranged cutoff scheme. The relative free energies of hydration obtained for using AM1BCC charges in combination with GAFF are presented in table 4.1, where each simulation was repeated three times using a different random number seed and the errors given are the standard errors of the mean over these runs.

Hydration free energies in the literature are being discussed in terms of sampling of internal degrees of freedom of the solute, i.e. relative hydration free energies versus single conformation hydration free energies[29]. In our studies, and in table 4.1, protocol A refers to solutes that are entirely flexible, i.e. all internal degrees of freedom are being sampled with the exception of rings, hence the reported free energies correspond to relative free energies of hydration, while protocol B only allowed rigid body translations and rotations, resulting in single conformation hydration free energies. Comparing both estimates within a given water model clearly highlights the fact that sampling in these systems does not have a drastic impact on the computed free energies. This was to be expected for the benzenes substituted with halides, as little conformational motion is expected. However, for aniline, toluene and nitrobenzene, the effect could have been more pronounced, if geometries that were not consistent with the force field had been supplied. This is not the case for our study as geometries were supplied, and in fact minimized, using the underlying force field.

Moreover, the perturbations considered here all appeared well behaved. As an example we show the results of an energy decomposition for the perturbation of benzene to aniline, yielding a hydration free energy of -4.89 kcal mol^{-1} compared to the

solute	Exp.	TIP3P		TIP4P	
		Protocol A	Protocol B	Protocol A	Protocol B
aniline	-4.63	-4.92 +/- 0.01	-5.03 +/- 0.05	-4.89 +/- 0.08	-5.04 +/- 0.10
brbenz	-0.60	0.59 +/- 0.05	0.68 +/- 0.03	0.71 +/- 0.05	0.73 +/- 0.01
clbenz	-0.26	0.45 +/- 0.01	0.47 +/- 0.04	0.64 +/- 0.01	0.69 +/- 0.01
fbenz	0.06	0.80 +/- 0.03	0.79 +/- 0.01	0.98 +/- 0.04	1.00 +/- 0.03
nitrobenz	-3.26	-3.27 +/- 0.03	-3.16 +/- 0.13	-3.60 +/- 0.09	-3.27 +/- 0.10
toluene	-0.03	0.26 +/- 0.03	0.18 +/- 0.04	0.23 +/- 0.03	0.18 +/- 0.03

Table 4.1: Relative hydration free energies obtained for GAFF using AM1BCC charges. All figures are given in kcal mol^{-1}, and the error estimate corresponds to the mean unsigned error, caluclated over 3 simulations. Two water models, TIP3P and TIP4P have been used in protocol A and B, using flexible or rigid solutes respecyively.

experiment standing at -4.63 kcal mol^{-1}. The equilibration prior to data collection yields a well equilibrated box of TIP4P water molecules and the recorded total energy of the system behaves stably from the start of the simulation. This is show in figure 4.8(a). Similar effects can be observed for intermolecular Lennard-Jones and Coulombic interactions, all characterised by very stable estimates for the averaged energies, shown in figures 4.8(b) and 4.8(c). The use of RETI allows one replica to travel along the reaction coordinate λ, which is shown in figure 4.8(d), and where only 2 replicas are being shown due to clarity. The acceptance rates for the RETI moves were in the region between 40 and 50 %.

Most force fields have been parametrised using simulations that employed finite-ranged non-bonded interactions. This would suggest one should use the corresponding truncation protocol for non-bonded interactions in order to be faithful to the parametrisation of the model. At the same time truncation schemes have been shown to yield qualitatively incorrect results, as compared with more sophisticated and increasingly common methods such a Ewald summation or the use of a reaction field[206]. A finite-ranged treatment for non-bonded interactions is ultimately inappropriate for the inherently long-ranged Coulombic interactions of charged species. However, OPLS-AA[67,68] was developed using quadratically tapered cutoffs and a long-range Lennard-Jones correction, and AMBER[69,161,162] was developed with residue-based abrupt cutoffs, which result in discontinuous interaction energies and forces at the boundary. Studies supportive of this observations have confirmed a drastic difference on the computed free energies, i.e. up to 1.56 kcal mol^{-1} in the case of a

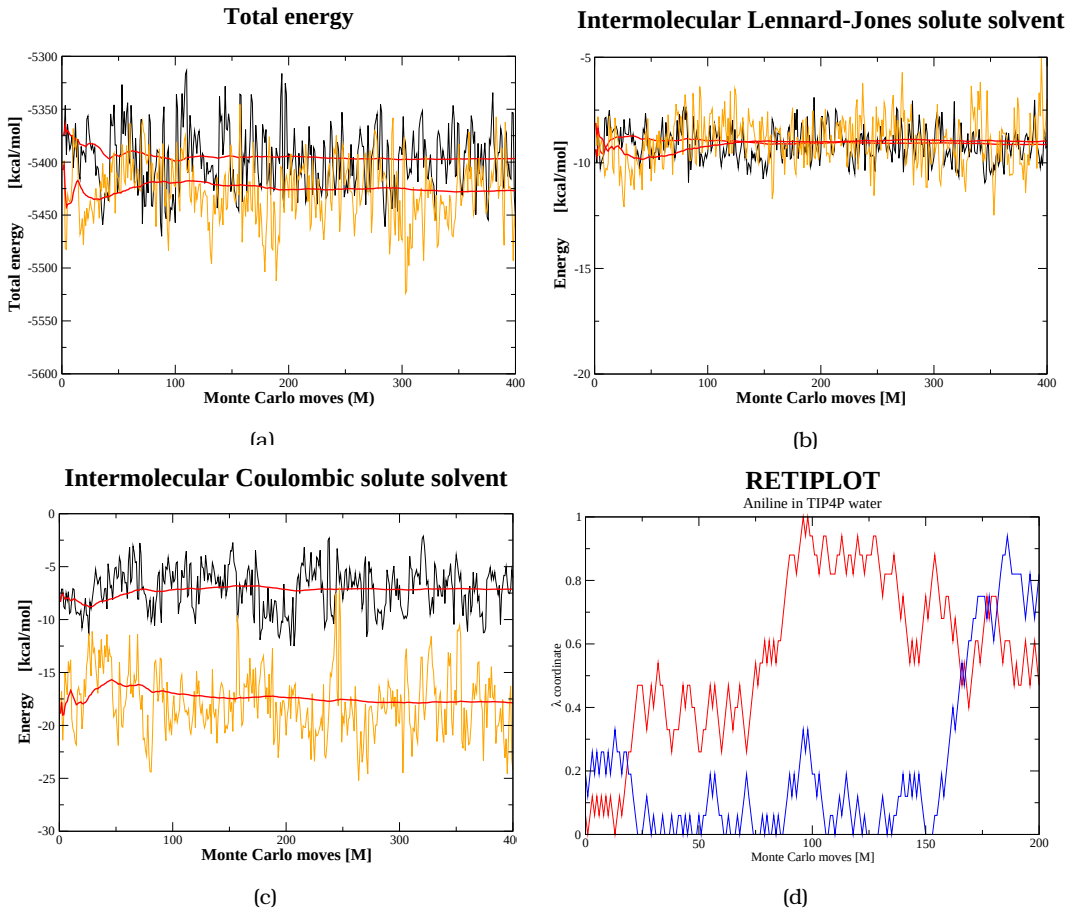


Figure 4.8: Energy decomposition for the perturbation benzene to aniline in TIP4P water using GAFF and AM1BCC charges: Total energy of the systems (a), intermolecular Lennard-Jones interactions (b), intermolecular Coulombic interactions (c) and a plot for two replicas showing good exchange thus accelerate sampling (d). The orange line in the graphs corresponds to aniline, while the black line corresponds to benzene. The red lines accompanying the black and orange lines represent the statistical averages.

TRP analogue[160]. This difference was entirely attributed to the attractive long-range Lennard-Jones terms, and was being covered using a correction term. Generally although Lennard-Jones interactions are rather small outside the cutoff ranges, they are still attractive everywhere, and hence can contribute significantly to the solvation free energy.

Issues relating to the finite cutoff employed in this study are not expected to have a drastic effect of the free energies of hydration as no charged solutes are present in the series. To validate this assertion we have extended the cubic boxes of TIP3P and TIP4P water molecules from 25x25x25 and using a 9 Å cutoff quadratically smoothed over the last 0.5 Å, to 50x50x50 and increasing the non-bonded cutoff to 20 Å. The results for both cutoff ranges and both water models did, as expected, not result in significant differences in relative hydration free energies and would not have changed any of the predictions by more than 0.1 kcal mol^{-1} for GAFF in combination with AM1BCC charges.

The degrees of freedom contributing most to the free energy of hydration are those associated with the rearrangement of the solvent. The original TIP3P and TIP4P water models are rigid and include no Lennard-Jones terms on the hydrogens, but an off-centre charge is introduced in TIP4P to mimic the lone-pair electrons on the oxygen, while in TIP3P the charge is centred on the oxygen atom[195]. However, for the solutes studied here and using both water models, results according to table 4.1 indicate a high degree of agreement, i.e. correlation coefficients of 0.98 and PIs of 0.93 for all simulations using GAFF/AM1BCC and either TIP3P or TIP4P water.

Results for GAFF with RESP charges are presented in table 4.2. AM1BCC, a charge model to reproduce RESP charges[49], in fact outperforms RESP charges with respect to the accuracy of the results as well as the rank-ordering of the solutes, i.e. 0.98 versus 0.96 and 0.93 versus 0.82 for the correlation coefficients and PIs obtained for AM1BCC versus RESP charge models with GAFF respectively. With respect to the fact that the generation of the RESP charges is substantially more elaborate and for this series still performed worse than AM1BCC, it could not be shown that high-level ab initio calculations justify their use in generating atomic charges.

However, the dominant route to generating atomic charges has been the use of quantum mechanics and semi-empirical methods, and to fit the charges to reproduce

solute	RESP	AM1CM1A	QMMM
aniline	-4.81 +/- 0.03	-4.90 +/- 0.05	-5.5 +/- 0.20
brbenz	0.77 +/- 0.01	0.41 +/- 0.01	0.35 +/- 0.10
clbenz	0.55 +/- 0.01	0.68 +/- 0.02	-0.1 +/- 0.06
fbenz	0.69 +/- 0.01	0.99 +/- 0.01	0.2 +/- 0.05
nitrobenz	-4.99 +/- 0.13	-2.25 +/- 0.09	-5.3 +/- 0.10
toluene	0.48 +/- 0.05	0.03 +/- 0.02	0.2 +/- 0.01

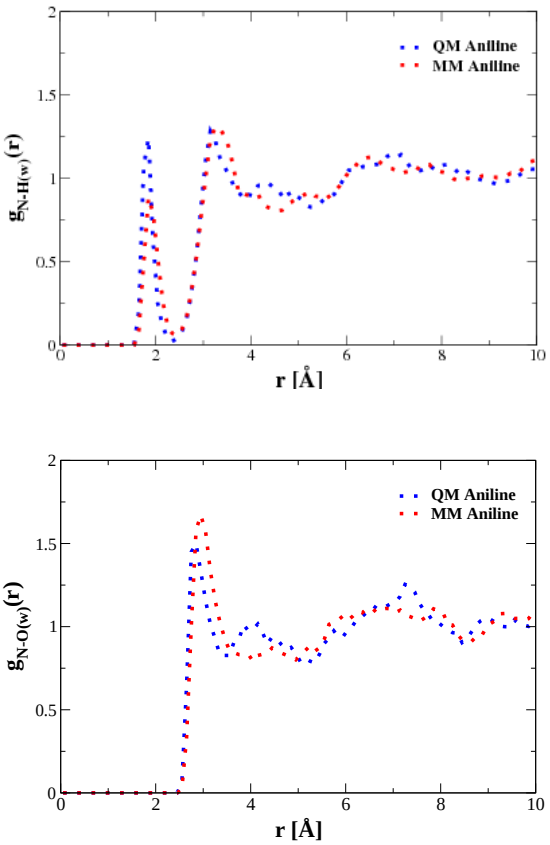
Table 4.2: Relative hydration free energies obtained for GAFF using RESP charges (labelled RESP), GAFF/AM1BCC in combination with BLYP/6-31g* labelled (QMMM), and OPLS-AA using 1.14*AM1CM1A (labelled AM1CM1A). All figures are given in kcal mol^{-1} and the error estimate is the mean unsigned error calculated over 3 simulations.

electronic properties, especially dipole moments and the electrostatic potential. This has led to the development of RESP charges[50]. In a similar way, AM1CM1A charges have been developed to use with the OPLS-AA force field[51]. The resulting free energies of hydration using AM1CM1A charges and OPLS-AA are presented in table 4.2 and indicate a considerable degree of agreement with the results obtained using AM1BCC charges with GAFF. In fact, correlation coefficients and PIs are essentially identical and stand at 0.98 and 0.95 versus 0.98 and 0.93 using AM1MCC/GAFF and AM1CM1A/OPLS-AA parameters, respectively. Given that OPLS-AA has been parametrised to reproduce not quantum mechanical properties but pure liquid properties, one might have expected a superior performance of OPLS-AA compared to GAFF. Looking at the predictions of using AM1BCC/GAFF and AM1CM1A/OPLS-AA highlights that both force fields perform equally well. The OPLS-AA force field is a diverse collection of force field parameters, that, depending on their source, will perform differently. For example, the OPLS-AA force field has been re-parametrised for ammonia, secondary and tertiary amines, as the initial parametrisation used parameters derived for primary amines for essentially all amines[207]. However, these parameters are not present in all distributions of OPLS-AA, as the force field is constantly being improved to cover a wider range of potential ligand types. Here, we made use of the OPLS-AA parameters that come with MCPRO version 2.1.

The combined Warshel-RETI-MMBIF method[52] was used to calculate free energies of hydration for the series of substituted benzene molecules, using BLYP/6-31* as the QMMM Hamiltonian and GAFF/AM1BCC as the approximate Hamiltonian. Although minor changes are being introduced to perturb the reference solute benzene

to all of the more polar solutes, i.e. fluorobenzene, bromobenzene, chlorobenzene, aniline, toluene and nitrobenzene, these calculations could be very informative. For example, the benzene to fluorobenzene perturbation will change the lowest non-zero multipole moment of the solute from a quadrupole to a dipole. The resulting changes in the electric field around the molecule may in turn cause a significant solvent reorganisation. The extent to which the MM and QMMM representations are able to capture this effect will be of interest. Using the QMMM approach, the main corrections are predicted for the halobenzenes with -0.4, -0.8 and -0.8 kcal mol^{-1} for bromo-, chloro- and fluorobenzene respectively. Therefore it provides a more accurate estimate for halobenzenes, compared to AM1BCC and RESP. However, looking at aniline and nitrobenzene, this trend is not observed. In fact, aniline, corrected by -0.6 kcal mol^{-1} using the QMMM approach, is predicted too negative, by, say, roughly 1 kcal mol^{-1}, while nitrobenzene, corrected by +2.0 kcal mol^{-1}, becomes too positive, and is predicted roughly 1 kcal mol^{-1} below its' experimental value. As for all other results presented in this section, the standard error of the mean of three independent simulations, using a different random number seed, is consistently below 0.1 kcal mol^{-1} (see table 4.2 for details), and hence allows the calculation of the free energies of hydration with high precision. Apart from calculating free energies of hydration as one measure of the fitness of a force field, it should be emphasised that there are other measures of fitness that are important, such as other experimental observables including the differences in geometry.

Systematic studies of hydration structures are clearly warranted to provide a detailed understanding of the hydration process. Therefore we have also evaluated the solvation structures of our solutes in infinitely dilute solutions of TIP4P water via radial distribution functions (RDFs)[159]. This has been accomplished by implementing a monitor function in SIRE[204], which records the RDFs histograms up to 10 Å using a bin spacing of 0.1 Å. We note that more specific distribution functions may be computed, and combining these axial, spatial and cylindrical RDFs results in the overall RDF; in principle these give a more detailed picture of the hydration processes. However, since we are mainly interested in defining an optimal set of parameters to be used for the simulation of DHODH, we have not aimed at implementing these more distilled versions of the RDFs. The RDFs recorded for aniline are shown in figure 4.9.



(b)

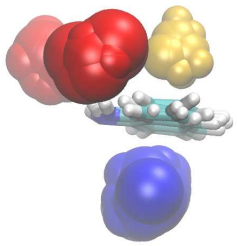
Figure 4.9: RDFs recorded for aniline using a QMMM representation. Figure (a) plots the $g_{N-H(w)}(r)$ distribution function and figure (b) the $g_{N-O(w)}(r)$ distribution function against the distance r given in Å.

The primary structure of water around aniline is shown in figure 4.10. This figure has been created from snapshots from QM aniline recorded during the simulation. Inspection of the trajectories revealed that the hydrating water molecules, form as expected, both hydrogen-bond donating and accepting pairs with the amino group. Two cups along the direction of the N-H bonds shown in red in figure 4.10 are due to hydrogen-bond accepting water molecules, whereas elongated features below the nitrogen shown in blue in figure 4.10 are due to hydrogen-bond donating water molecules. In addition, water tends to form a π-type complex with the aromatic region of aniline, which is shown in yellow in figure 4.10.

The recorded RDF between the nitrogen atom of aniline and the hydrogen atoms of the surrounding waters are shown in figure 4.9(a) and between the nitrogen atoms of aniline and the oxygen atom of water in figure 4.9(b). Previous discussions in the literature are based on the analysis of the nitrogen-oxygen RDF, and the application of spatial distribution functions has shown that in fact the recorded nitrogen-oxygen RDF, i.e. $g_{N-O(w)}(r)$, as shown in figure 4.9(b), is a combination of two separate RDFs: the below-the plane SDF, due to the hydrogen-bond donating water, with a minimum at 2.8 Å and corresponding minumum at 3.2 Å, thus indicating the presence of a strong hydrogen-bond; and the above-the-plane SDF, due to hydrogen-bond accepting neighbours, shows a broad first peak with maximum and minimum shifted to 2.98 and 4.4 Å respectively, therefore suggesting a weaker hydrogen-bond. The average of both SDFs results in maximum and minumum of 2.8 and 3.4 Å respectively. Inspection of figure 4.9(b) shows that the RDF recorded is in exact agreement with these findings. The complementary set of RDF for the nitrogen atom of aniline and the hydrogen atoms of surrounding water molecules is shown in figure 4.9(a). As for the $g_{N-O(w)}(r)$, the position and shape of the first peak in $g_{N-H(w)}(r)$ with maximum and minimum located at 1.8 and 2.4 Å respectively, confirm the presence of a strong hydrogen-bond between the hydrogen of water and the nitrogen. For both, $g_{N-H(w)}(r)$ and $g_{N-O(w)}(r)$, the relative positions of the first peaks sugest a linear hydrogen bond and they indicate the presence of a strong donating hydrogen bond between the nitrogen atom and the below-the-plane water, as well as a weaker hydrogen-bond for the hydrogen-bond accepting waters above-the-plane. Again, these findings correspond with findings of other groups, and therefore we may be confident that the current QMMM approach reproduces the important structural features upon solvating aniline in water well. Additionally, figure 4.9 shows that the recorded RDFs do not differ between the MM and the QM representation, suggesting that both methodologies may capture the solvent structure appropriately.

4.7 Conclusions

The results presented here, were intended to identify an optimal set of charges in combination with other force field parameters, to simulate protein-ligand binding in



(a)

Figure 4.10: Hydration network for the simulation of aniline using a MM representation (a). The water molecules, shown in sphere, have been extracted from simulation snapshots and are colour-coded blue, red and yellow, for the hydrogen-donating, hydrogen-accepting, and stacking interaction with the aromatic ring respectively.

DHODH. A summary of the calculated relative free energies of hydration are given in table 4.3.

solute	Exp.	AM1BCC	RESP	AM1CM1A	QMMM	Mobley
aniline	-4.63	-4.89	-4.81	-4.90	-5.5	-5.22
brbenz	-0.60	0.71	0.77	0.41	0.35	0.33
clbenz	-0.26	0.64	0.55	0.68	-0.1	0.1
fbenz	0.06	0.98	0.69	0.99	0.2	-0.5
nitrobenz	-3.26	-3.60	-4.99	-2.25	-5.3	-2.7
toluene	-0.03	0.23	0.48	0.03	0.2	-0.01

Table 4.3: Summary of the calculated relative hydration free energies in this study, together with the experimental values (column Exp.) and the results of a study of Mobley and coworkers [132] (column Mobley). All values are given in kcal mol^{-1} and the mean unsigned error is consistently below 0.1 for AM1BCC, RESP, AM1CM1A and QMMM. The standard error for the study of Mobley is below 0.5 kcal mol^{-1}.

The inspection of this table underlines the fact that all force fields and charge methods yield qualitatively similar results, and methods that apply a higher level of theory, i.e. QMMM, or use higher levels of theory to generate appropriate charge sets, did not generally provide more accurate results. This is in agreement with reports of other groups [131,132,160]. Although these published studies did attempt the calculation of absolute free energies of hydration, instead of the relative free energies of hydration presented here, similar relative hydration free energy results may be derived. This

is shown for a study of Mobley and co-workers[132] in table 4.3 in column labelled Mobley.

The QMMM protocol has provided more accurate results for halide-substituted benzenes compared to RESP and AM1BCC. Since our main focus for the perturbations in DHODH are fluorine atoms, one could argue that the novel QMMM method should be used for calculating the free energies of binding for DHODH. Moreover, the QMMM method has shown to reproduce liquid properties, other than hydration free energies, well, i.e. the RDFs and the resulting solvent configurations were in agreement with findings of other groups[28]. However, when this project was started, SIRE[204] was still in a development phase and this study was not only used to calculate hydration free energies, but to validate and help debugging the code. For this reason the BLYP method was the only method available at that time. Also, we do not give any detailed information regarding the computer time required to apply any of the protocols described. However, all of the MM calculations reported here can be run within 24 hours using 8 standard dual-core processors. For the QMMM method reported, this estimate is significantly increased. For example, a single simulation of fluorobenzene using a single standard dual-core processor requires 35 days, while a single run for nitrobenzene requires 73 days. This makes not only debugging the code tedious, but also prevented us from applying QMMM on DHODH, as in that case debugging SIRE[204] for the use in binding free energy simulations would have been required.

Given the ease of generating charges using AM1BCC as implemented in ANTECHAMBER, and of AM1CM1A charges using AMSOL, either of these methods would be suitable. However, GAFF is more widely applied compared to OPLS-AA, and does not require experiments to derive new parameters. Therefore, to remedy the problems of the QMMM approach and to allow a wider validation of the performance of QMMM, work in our group is currently focusing on the development of novel QMMM methods[30], while for the purpose of studying DHODH we select GAFF and AM1BCC charges.

5

Assessment of crystallographic water molecules

Having defined an optimal set of force field parameters to be used for the simulation of the Baumgartner series of inhibitors of DHODH[48], we shall now elaborate on the thermodynamic end states, and indeed, problems associated with them. Figure 5.1 shows the structure 2FQI deposited in the PDB data bank.

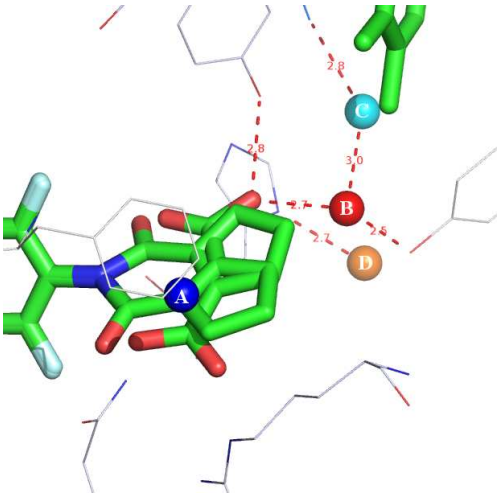


Figure 5.1: The structure of compound 5 and part of the FMNH2 molecule protruding towards the inhibitor, both shown in stick representation and using green, red, blue and cyan for carbon, oxygen, nitrogen and fluorine respectively. Key residues are shown in line representation using light-grey, red and blue for carbon, oxygen and nitrogen atoms respectively. The crystallographic water molecules, i.e. labelled A, B, C and D, that establish two different solvation pockets, are shown in sphere representation and are colour-coded according to their B-factors, 25, 46, 29 and 42 Å^2 for waters A, B, C and D respectively. The figure has been created using PyMOL[155].

The dual binding modes, resolved by the crystallographers, may or may not exist. The answer to this question is essential if we were to design and further profile a new tightly binding small molecule for DHODH, and also demonstrate the limits of diffraction experiments. However, it also highlights the need for free energy simulations applying the previously developed dual topology paradigm, as described in chapter 2, to confirm and complement the crystallographic observation, and hence point towards either of the binding modes, or indeed, both binding modes. To be able to rigorously define two separate thermodynamic end states from the single structure deposited in the PDB and shown in figure 5.1, we need to rigorously define the hydration pattern of each binding mode. A thorough definition of these may result in perturbations that reflect the actual or theoretically plausible modes of binding, and thus thermodynamic cycles constructed for the perturbation of one binding mode into another should, in principle, and under the assumption of complete sampling, close to within an error typical for perturbations applying the dual topology scheme[32,46]. Not considering an active site water molecule, may not only result in a free energy difference between two binding modes that lacks the energetic contribution of removing or creating a particular water, but subtle conformational changes resulting from the presence or absence of a water molecule may diverge the sampled free energy path compared to the experiment[123].

Two major hydration sites in the ligand binding domain of DHODH can be identified crystallographically[48]. The first hydration site consists of a single water molecule labelled A in figure 5.1. This site can be understood as an isolated site with one water molecule stacking with a phenyl ring of the neighbouring PHE and H-bonding with nearby ARG and GLN residues and, depending on the assumed binding mode, may interact with the ligand's carboxy or phenyl ring moiety. Water A can be found in all crystal structures of the Baumgartner series[48].

A second hydration cluster, consisting of water molecules labelled B, C, and D in figure 5.1, lies at an interfacial region of the protein. A close-up view of this water pocket is available in figure 5.2. This interface is in fact twofold as crystallographic water molecules shield the inhibitor from the prosthetic group FMN to some extent, as well as providing a tunnel to access bulk solvent. Waters B, C, and D in figure 5.1, all part of this second hydration cluster in DHODH, build up a triad of water molecules. This is visualised by the blue dashed lines connecting waters B, C and D

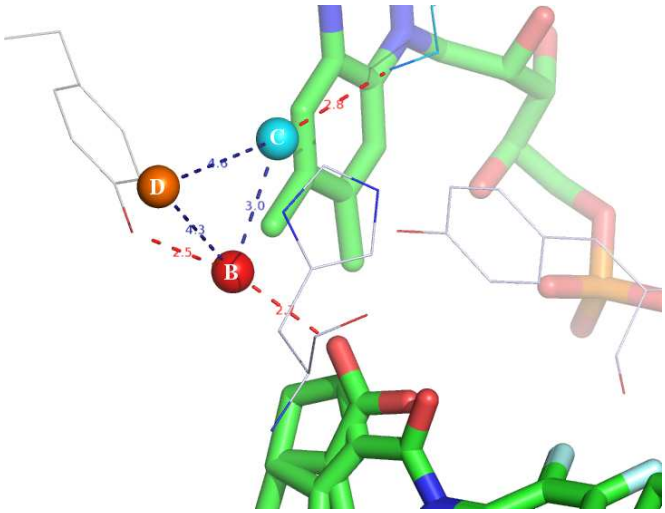


Figure 5.2: Close-up view of compound 5 and part of the FMNH2 molecule protruding towards the inhibitor, both shown in stick representation and using green, red, blue and cyan for carbon, oxygen, nitrogen and fluorine respectively. Key residues are shown in line representation and the crystallographic water molecules, i.e. labelled B, C and D, are shown in sphere representation using a colour-coding representative for their B-factors (46, 29 and 42 Å^2 for waters B, C and D respectively). The figure has been created using PyMOL[155].

in figure 5.2, while interactions involving protein residues are shown in red dashed lines. Here, water B is bridging the interactions between a nearby TYR residue and the inhibitor, and could in principle also H-bond with the FMN molecule. Water C is H-bonded to water B and shows an additional H-bond with a nearby carbonyl of a GLY residue. Water D is very close to the bulk region, hence it lies at the exit of the hydration tunnel, is H-bonded to a nearby HIS residue and seems to direct the protonation state of this residue. All 3 waters together build a H-bonding network with the inhibitor, the protein and possibly the prosthetic group, while H-bonds between individual water molecules may exist. It is this region, that, at least in crystallographic experiments, shows a subtle rearrangement of water molecules[48] depending on the inhibitor, yielding a hydration cluster with up to three crystallographically observed waters, that may mediate intermolecular interactions while lying within the tunnel and not in the nearby bulk solvent. The B-factors, a measure for the effective diameter of an atom's electron density[65], are colour-coded in figure 5.2 and stand at 46, 29 and 42 Å^2 for waters B, C, and D, respectively.

In fact, the displacement of water molecules in binding sites by judicious lig-

and modification has emerged as a strategy to optimize lead compounds[119,208] while implicit solvent theories can be used to model successfully the effect of bulk desolvation in protein-ligand binding[18]. However, implicit solvents are unlikely to account for specific water-ligand interactions, while the explicit consideration of a few water molecules in the vicinity of a ligand has been shown to improve the quality of predictions from docking algorithms[209,210]. However, these hydration sites are often neglected in docking studies, or are uncertain due to difficulties in resolving water molecules by crystallography[87,211]. Looking at the B-factors of water molecules in the Baumgartner series[48], it appears problematic to assign a reliable hydration state. Therefore some alternative computational methodologies have been proposed to aid this task.

Empirical techniques based on interaction energies or QSAR descriptors to assess the hydration of protein binding sites have been reported in the literature, but their transferability across targets and limited accuracy due to their empirical nature may raise concerns[212,213]. For instance, the method CMIP has been reported to overestimate the number of hydration sites, possibly because it neglects entropic contributions to the affinity of a water molecule for a given site[213,214]. More rigorous MC or MD simulation can be conducted to equilibrate the water distribution in a binding site. In particular, Lazaridis and Li have used MD simulations and inhomogeneous fluid solvation theory to evaluate the binding enthalpies and entropies for interfacial water molecules in protein-ligand complexes[215], and was subsequently extended to both locate all water molecules in a protein binding site and evaluate the favourability of their displacement[216,217]. However, a major drawback with MC/MD approaches is that diffusion of water molecules in and out of cavities at the protein interface can be excessively slow or even kinetically trapped[123]. Hence, depending on the particular structure of the binding site, it may be impossible for standard free energy simulation methods to obtain results that do not depend markedly on the initial setup, or that do not show marked differences upon inclusion of a water molecule. In the following two sections we introduce alternative methods to estimate active site hydration by applying simulation methodology and statistical thermodynamics. These approaches have subsequently been utilized to predict the hydration of the ligand binding domain of DHODH for the Baumgartner series.

5.1 (J)ust (A)dd (W)ater (M)olecule(S): JAWS

The equilibrium hydration of a binding site depends on the difference between the free energy of a water molecule in bulk solvent and in the binding site:

$$\Delta G_{bind}(water) = G^{site}_{water} - G^{bulk}_{water} = -\ k_B T\ ln\ \frac{Q^{site}_{water}}{Q^{bulk}_{water}} \tag{5.1}$$

where $\Delta G_{bind}(water)$ is the free energy of binding of a water molecule, G^{site}_{water} and G^{bulk}_{water} are the free energies of the water molecule in the binding site and in bulk respectively; T is the temperature and k_B the Boltzmann constant, and Q^{site}_{water} and Q^{bulk}_{water} are the partition functions for the water molecule in the protein binding site and bulk solvent respectively.

A direct calculation of $\Delta G_{bind}(water)$ is, however, not practical when using for instance the Zwanzig equation[11], as the exchange of water molecules between bulk and the ligand binding domain can be very slow. Gilson and coworkers proposed a method where this quantity can be estimated using a series of unphysical intermediates that allow the transfer of a water molecule from bulk to the protein binding site. This method is called *double decoupling method* and is illustrated in figure 5.3[218].

In the double decoupling approach, the removal of a water molecule from bulk and from the binding site are being compared, yielding the absolute free energy of binding of the water molecule. Thus, equation 5.1 can be written as

$$\begin{aligned}\Delta G_{bind}(water) = &-\ kT\ ln\left(\frac{Q^{bulk}_{ideal}}{Q^{bulk}_{water}}\right) - kT\ ln\left(\frac{Q^{site,constr}_{ideal}}{Q^{bulk}_{ideal}}\right)\\ &- kT\ ln\left(\frac{Q^{site,constr}_{water}}{Q^{site,constr}_{ideal}}\right) - kT\ ln\left(\frac{Q^{site}_{water}}{Q^{site,constr}_{water}}\right)\end{aligned} \tag{5.2}$$

which can be rewritten as

$$\begin{aligned}\Delta G_{bind}(water) = &-\ \Delta G_{hyd}(water) + \Delta G_{constr}(ideal, site)\\ &+ \Delta G_{trans}(water, site) - \Delta G_{constr}(water, site)\end{aligned} \tag{5.3}$$

where the first term $-\Delta G_{hyd}(water)$ corresponds to the free energy change for removing the intermolecular interactions of a water molecule in bulk, i.e. step B in figure 5.3. It is the negative of the excess hydration free energy of water[159] and according to previous results for computing the free enery of hydration of a TIP4P

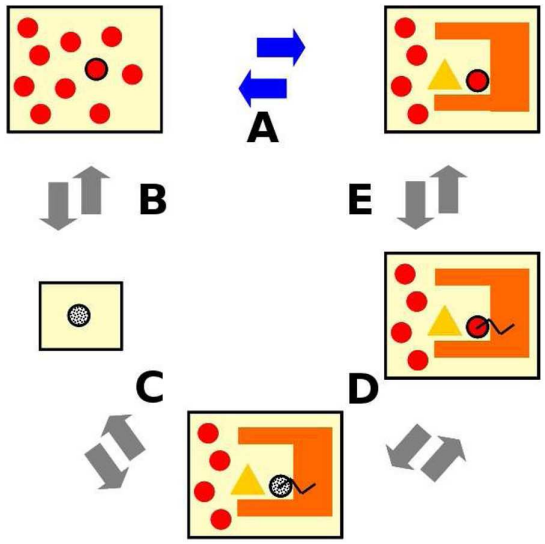


Figure 5.3: Reversible transfer of a water molecule from bulk solvent to a well-defined position in a protein binding site: (A) direct process; (B) decoupling of one water molecule from bulk; (C) localization of an ideal particle into a binding site; (D) conversion of a localized ideal particle into a water molecule; (E) removal of the constraint. Water molecules are depicted by red spheres, ideal particles by gray spheres. The orange shape represents a binding site. The yellow triangle represents a ligand. The black zigzag depicts a volume constraint. The figure has been taken from [123].

water model [122] and experimental data [219] this value was set to 6.4 kcal mol^{-1} in this study.

The second term in equation 5.3, $\Delta G_{constr}(ideal, site)$, corresponds to the free energy of constraining the now ideal particle, i.e. as the particle is not interacting, to occupy a volume V^{constr} in a binding site instead of the volume V° available to a water molecule in bulk solvent, shown in step C in figure 5.3. Both end states now only differ in translational entropy, and hence this term is equal to the ratio of available volumes as shown in equation 5.4 [6], where an appropriate value of V° is the inverse of the concentration C° of bulk water, i.e. 55.55 mol/L [159], and V^{constr} depends on the nature of the constraint.

$$\Delta G_{constr}(ideal, site) = -kT\ ln\left(\frac{V^{constr}}{V^{\circ}}\right) \tag{5.4}$$

The third term in equation 5.3, $\Delta G_{trans}(water, site)$, corresponds to the conversion of the localised ideal particle into a water molecule, see step D in figure 5.3. Several free energy techniques may be employed to compute this term. For example, the

Lennard-Jones terms may be first turned off, followed by the atomic partial charges on the water molecule. It is important that the particle is constrained. Hypothetically, when the intermolecular interactions are removed and no constraint is being applied, the particle could, in principle, sample the entire simulation volume and cause large numerical errors[220]. In our studies a hard-wall potential was used whereby the particle is constrained to occupy a sphere of radius R. The choice of R is not trivial and has great impact on the computed free energies. If R is too large, the process will not be reversible and numerical errors will arise, while if R is too small, important configurations of the fully interacting water will be missed[32,218]. For our calculations a radius of 1.4 Å was found to be adequate, as a small variation on this number, for example between 1.4 and 2.0 Å, did not results in different computed free energies. while radii under 1.0 Ådid result in very high energies and radii over 2.8 Å in large error statistics.

Finally, the last term in equation 5.3, $\Delta G_{constr}(water, site)$, corresponds to the free energy change for removing the constraint. This is illustrated in step E in figure 5.3. This term must be zero, if one can demonstrate that the confinement does not affect the calculated $\Delta G_{bind}(water)$.

Moreover, the analysis above can be extended to simulate the insertion of water molecules into N hydration sites (with indices i = 1, ... , N) in a binding site via equation:

$$\begin{aligned}\Delta G_{bind}(Nwater) = &- kT\ ln\left(\frac{Q_{Nwater}^{sites,constr}}{\Pi_{i=1}^{N} Q_{ideal}^{site1,constr}}\right) \\ &+ \left(\sum_{i=1}^{N} \Delta G_{constr}(ideal, site\ i) - \Delta G_{hyd}(water)\right)\end{aligned} \tag{5.5}$$

Evaluation of equation 5.5 is complicated due to the existence of coupled interactions between water molecules located at each hydration site. In principle, N water molecules could be iteratively transferred to an increasing number of hydration sites (i = 1, 2, ..., N) until a global minimum in ΔG_{bind} is found at which stage the optimum integer number of water molecules has been determined. However, this would not be feasible as the number of required free-energy calculations is very large. Furthermore, prior knowledge of the positions of each of the N putative hydration sites in the binding site would be needed, which cannot always be guaranteed experimen-

tally, and alternative distributions of the same number of hydration sites have to be considered for reliable estimates, i.e. sampled.

Thus Michel and coworkers have elaborated approximations from equation 5.5 to yield a more practical methodology. These approximations on the double decoupling approach explained so far, led to JAWS[123]. The usual expression for the potential energy function $U(r)$ that describes N water molecules in a protein binding site can be modulated using N scaling parameters θ_i, as shown in equation 5.6, where U_{inter} is the intermolecular energy of water molecule i and U_0 gives the remaining energy terms.

$$U(r, \sum_{i=1}^{N} \theta_i) = U_0(r) + \sum_{i=1}^{N} \theta_i U_{inter}(r, water\ i) \tag{5.6}$$

In JAWS, the θ_i are treated as degrees of freedom, which can be sampled during a MC simulation in the same spirit as in the λ-dynamics method developed by Kong and Brooks[108]. To avoid confusion through the introduction of a new coupling parameter, the symbol θ_i is used to distinguish this set of parameters from the coupling parameter λ used in our RETI studies. According to the θ coupling, the θ-water i behaves as an ideal particle when $\theta_i = 0$ and as a regular water molecule when $\theta_i = 1$. By collecting statistics during an MC simulation, the probability that a θ-water is water-like or ideal-like is readily determined and the ratio of these two probabilities is formally related to the free energy change $\Delta G_{trans}(water, site\ i)$ for transferring a water molecule from the gas phase into site i using equation 5.7.

$$\Delta D_{trans}(water, site\ i) = -kT\ ln\ (\frac{P(\theta_i \rightarrow 1)}{P(\theta_i \rightarrow 0)} \tag{5.7}$$

It is important to highlight that a reliable free-energy estimate can only be achieved with equation 5.7 if the θ_i-water molecule samples readily both high and low θ values over the course of the simulation. If high energy barriers must be crossed, or if the free energy difference is large, excessive computational resources might be required before a statistically significant number of transitions can be observed. This difficulty can be overcome by adding a biasing term, $V(\theta_i)$ in equation 5.8[221], for each of the N θ-water molecules to the potential energy in equation 5.6. The excess hydration free energy of water has been proposed by Michel et al. and hence the biasing term was set to 6.4 kcal mol^{-1} in the current study. As a consequence, water molecules

that appear water-like in the simulation prefer the protein binding site environment instead of bulk water.

$$V(\theta_i) = \Big(- \Delta G_{hyd} + \Delta G_{constr}(ideal, site_i) \Big) \theta_i \tag{5.8}$$

The V terms correspond exactly to what is necessary to correct and estimate a binding free energy for each of the θ-water molecules. In other words, addition of the V terms and collecting statistics about the probabilities that a θ_i parameter is unity or zero, permits the direct partitioning of N θ-water molecules between bulk and the binding site. Each $V(\theta_i)$ term penalises high water-like θ_i values by an amount that accounts for the free energy change for desolvation and localization of a water molecule at hydration site i. Therefore, evaluation of the ratio of probabilities of high and low θ_i values during a simulation with the $V(\theta_i)$ terms activated, allows direct estimation of a free energy of binding ΔG_{bind} for each θ-water, in the presence of N - 1 other θ-water molecules.

Technically, JAWS is carried out in two phases, i.e. JAWS phase one and two. In JAWS phase one, potential hydration sites are detected by placing N θ-water molecules randomly onto a grid and allowing these waters to freely sample the entire region in a MC simulation, where the θ-water molecules are translated and rotated, and for 50% of the moves a random variation in its θ value is attempted. θ-waters with θ values greater than 0.995 are considered water-like and the nearest grid point is incremented by 1, yielding the probability distribution of water occupancies. At the end of JAWS phase two, these occupancies are converted into an integer number of potential hydration sites (using a clustering algorithm). In JAWS phase two, a θ-water molecule is placed at each hydration sites identified in JAWS stage one, and is constrained to occupy a volume of 27 Å^3. Now the biasing potentials $V(\theta_i)$ are turned on and statistics about their θ value is collected in a new MC simulation, estimating the free energy of binding of the water molecule from the ratio of probabilities of observing a θ-water at high, i.e. $\theta > 0.995$, or low, i.e. $\theta < 0.05$, θ values.

The incorporation of the V biasing terms into the potential energy function before monitoring the θ_i values to obtain a free energy of binding directly is advantageous, compared to correcting the transfer free energy from equation 5.7 after the simulation. Even if a water molecule at a possible hydration site is never turned off or on completely during the simulation and hence its binding free energy is undetermined,

the θ_i provides a good indication whether this site would be occupied or unoccupied by a water molecule, while the correction terms could not be added after the simulation, if insufficient statistics were collected to compute a transfer free energy using equation 5.7. Additionally, equation 5.8 provides the benefits of a biasing potential by penalising high values of θ which correspond to water-like states. Hence, equation 5.8 facilitates transitions between high and low θ values and enhances convergence of the binding affinities.

The introduction of θ moves to convert particles from water-like to ideal-like states has important advantages over more typical FEP approaches. Most importantly, a single simulation is sufficient to determine multiple hydration sites, as several water molecules can be assigned a θ coupling parameter. Moreover, cooperative hydration of the binding site by clusters of water molecules may naturally arise over the course of the simulation and estimates of ΔG_{bind} for each water molecule can be obtained from a single simulation. Using a double decoupling approach would using FEP for example, would require substantially more computer time.

5.2 Grand Canonical Monte Carlo (GCMC) methods

GCMC methods aim to change the number of water molecules during a simulation and hence make use of the Grand Canonical ensemble[222]. GCMC methods, first proposed by Adams in 1974[223,224], have been extensively studied in the literature to model systems where the number of particles in the system can change as a function of the external conditions. In a similar fashion to the ensembles outlined in chapter 2, a GCMC ensemble can be thought of as an extension to the standard canonical ensemble, where on top of the exchange of energy between individual systems, the number of particles can also be exchanged. The number of particles in the system is commonly controlled using the chemical potential. Hence in a Grand Canonical Monte Carlo (GCMC) simulation, the chemical potential, volume and temperature are kept constant whilst the number of particles in the system is allowed to change, i.e. the μVT ensemble. The grand canonical partition function may be expressed as[6]

$$Q(\mu VT) = \sum_{N=0}^{\infty} \frac{exp(\beta\mu N)V^N}{\Lambda^3 N!} \int ds^N exp\Big[-\beta U(r^N)\Big] \tag{5.9}$$

where N is the number of particles, β is $(k_BT)^{-1}$ with k_B as the Boltzmann constant and T the temperature; $U(r^N)$ is the potential energy of the system, μ is the chemical potential and Λ is the De Broglie wavelength $(\Lambda = h/(2\pi m k_B T)^{1/2})$.

As the number of particles in a GCMC simulation may change, three different types of MC moves can be imagined: an insertion, corresponding to an increase in the number of particles in the system from N to $N+1$; a deletion, corresponding to a decrease in the number of particles in the system from N to $N-1$ and finally a move that randomly displaces a particle, hence does not change the number of particles, which then follows the procedues of the standard canonical ensemble. All three moves may be accepted based on the following acceptance tests[6]:

$$P_{in} = min\Big[1,\ \frac{V'}{N+1} exp(\frac{B-\Delta E}{k_bT})\Big] \tag{5.10}$$

$$P_{del} = min\Big[1,\ \frac{N}{V'} exp(\frac{-B+\Delta E}{k_bT})\Big] \tag{5.11}$$

$$P_{dis} = min\Big[1,\ exp(\frac{-\Delta E}{k_BT})\Big] \tag{5.12}$$

Here, N is the number of particles in the simulation, B is the Adams parameter[224] that relates to the excess chemical potential μ' via $(B = \mu'/k_BT + lnN)$ and β is $(k_BT)^{-1}$. $V' = V/\Lambda^3$, where Λ is the de Broglie wavelength.

The computation of a free energy change in the μVT ensemble does not correspond to the usual experimental NPT conditions. Therefore, the Adams factor B is used in the simulation instead of μ. Since B and μ differ by a constant, performing a simulation at constant B is equivalent to performing a simulation at constant chemical potential μ'. Hence, to select a constant chemical potential μ' that yields an ensemble that closely matches experimental conditions, many simulations have to be carried out with varying magnitudes of B.

In GCMC simulations, move types refering to insertions and deletions of molecules are problematic particularly for dense systems, and their acceptance rates are usually limited to less than 1 %[225]. This is because many attempts to insert a particle at random are often rejected, since there is a high probability that the new particle will overlap with existing particles in the system. Developments that have attempted to overcome this problem were first proposed by Mezei and coworkers, i.e. the cavity-biased GCMC method[226,227] as well as the dynamic grid method proposed by Roux[225]. Using a cavity bias, insertions are attempted at predefined points in

the simulation, which have a radius equal to, or greater than, that of the radius of the test particle. A similar approach is the contact cavity-bias[228] method, that involves the computation of the radial distribution function of a test particle based on existing molecular centres, which has shown to increase the probability of finding a cavity by 5. The third approach to increase the acceptance rates of insertions has been realized with the dynamic grid approach[225], that involves the generation and maintenance of a dynamic grid. However, none of these methods to increase acceptance rates for insertions have subsequently been employed in this study, and thus no further details are given here, but the interested reader is referred to the original publications[225,226,227].

Initial validations of the GCMC method to assess hydration states, performed by M. Bodnarchuk as part of his PhD project, have shown no substantial benefit using a cavity-bias algorithm and thus the standard random insertion GCMC method has been applied here.

5.2.1 Free energy calculations of water molecules using GCMC

In 1996 Mezei and Guarnieri proposed a method relying on GCMC simulations for the study of DNA hydration, i.e. the *simulated annealing of the chemical potential method*[229]. As suggested by the name, the method involves simulations at high chemical potentials, with the effect being a flooding with water molecules, and corresponding to an increase in the probability of an insertion move being accepted, while after equilibrating at high values of the chemical potential, the resulting structures are then annealed at a lower value of the chemical potential. This approach is similar to the general idea, outlined in the last section, that several simulations need to be carried out in order to obtain an ensemble that is comparable to experimental NPT conditions. As a result, the weaker binding water molecules leave the simulation upon reducing the chemical potential and only the tightest binding water molecules remain in the simulation, once sufficiently low values of the chemical potential are simulated. Moreover, this method allows not only for the exact calculation of the free energy of binding of a water molecule, but also gives insight into differences in binding free energies of water molecules in a simulation, as at different values of the chemical potential, cooperativity effects of weak binders can be explored until this picture has been reduced to only the tightest binding waters.

The JAWS methodology proposed by Michel[123] as well as the GCMC method have been implemented in the in-house software ProtoMS by Michael Bodnarchuk as his PhD project. Here we make use of the new algorithmic features present in ProtoMS, as DHODH shows ideal features that would allow for the validation of these methods. Hence, for the remainder of this chapter we will start with the setup of the DHODH system for the simulations, followed by a more detailed and technical protocol applied to run JAWS and GCMC. We then present the results obtained using both approaches, investigate performance issues related to both protocols, and conclude with a thorough picture of the specific hydration of each ligand and each possible binding mode for the Baumgartner series. This should allow us to move towards our final aim: the prediction of binding modes and the relative free energies of binding in DHODH.

5.3 System Setup for JAWS and GCMC

All PDB structures of the Baumgartner series of human DHODH, i.e. structures 2BXV, 2FPT, 2FQI, 2FPV, 2FPY[48] deposited in the PDB[85] for compounds 3, 4, 5, 6 and 7 respectively, were selected as starting points for this study. Hydrogen atoms have not been resolved by the crystallographer and were added to each protein using WHAT-IF[91], Reduce[230] and Molprobity[231]. An inspection of all the resulting structures, revealed marginal differences in the protonation states of the protein residues obtained. Together with a visual inspection of all critical residues and taking on board crystallographic observations, e.g. the protonation state of HIS56 due to the presence of a stabilizing water molecule[48], a consensus was achieved. The proteins were setup with the AMBER99 force field[69]; inhibitors, the cofactor FMN and the natural substrate ORO were setup with the GAFF force field[136], and the atomic partial charges were derived using the AM1/BCC method[49], as implemented in the software antechamber[164].

The charges for symmetric sites on the inhibitor atoms were subsequently averaged to prevent artificially favourable rotameric states. Owing to the intramolecular hydrogen bond that may be present in all the inhibitors, and indeed was observed to be strong for all subsequent simulations, a formal charge of -1 was assigned to all inhibitor molecules. A formal charge of -2 was assigned for the FMNH2 molecule and

a formal charge of -1 was assigned to ORO, all of which were protonated using the PRODRG server[232].

To avoid steric clashes, each protein was energy minimized using the Sander module of AMBER8[200] and a generalized Born force field (the igb keyword was set to 1) for each ligand and each binding pose. Once the structures were minimised, all crystallograhic water molecules were removed and scoops have been created from the minimised proteins to reduce the computational cost. Hence protein residues that have one heavy atom within 15 Å of any heavy atom of the brequinar binding mode of each of the ligands were retained resulting in one scoop for each protein structure. These scoops contain 203, 206, 205, 204 and 209 residues for structures 2BXV, 2FPT, 2FQI, 2FPV and 2FPY[48] respectively.

In subsequent JAWS and GCMC simulations, the backbone of the energy minimized protein was kept rigid, and all non-bonded interactions were evaluated up to a distance of 10 Å, feathered over the last 0.5 Å and applying a residue-based cutoff. Additionally, and for validation purposes, several JAWS and GCMC simulations were compared to runs where the protein was allowed to move freely, i.e. allowing sidechain as well as backbone movements. Unless indicated in the results section of this chapter, the reported hydration sites are a result of the simulations where the sidechains were allowed to move but the backbone was kept rigid. The degrees of freedom of the side chains of protein residues with any atom within 10 Å of a ligand atom were sampled during the MC simulations. The inhibitors, the cofactor and the natural substrate were treated as fully flexible, and all bond-angles and dihedrals were sampled during the simulation, with the exception of rings and bond lengths.

The resulting complexes did not contain any crystallographic water molecule and were hydrated by a sphere of TIP4P water molecules[195] of 22 Å radius and centred on the geometric centre of the inhibitors. It is typical for MC simulations to usually reinsert crystallographic water molecules in the active site. However, since we want to determine the hydration pattern using JAWS and GCMC, these waters have not been reintroduced. To prevent evaporation in subsequent simulations, a half-harmonic potential with a force costant of 1.5 kcal $mol^{-1}Å^{-2}$ was applied to water molecules whose oxygen atom distance was greater than 22 Å from the center of geometry of the ligand. A similar sphere of TIP4P waters was applied to solvate the inhibitors for the free state.

For the purpose of a JAWS or GCMC study, additional water boxes need to be constructed. In JAWS these boxes are being used to flood the systems with θ-water molecules, and in a similar fashion in GCMC these waters may be used to attempt insertions. These boxes will be explained in the protocol sections for JAWS or GCMC.

Biological affinities of the inhibitors, measured by their IC_{50}s, was used from a single assay, and the experimental values were converted to binding free energies via the Cheng-Prusoff equation[233] and by assuming that the ratios of dissociation constants behave similar to the ratios of IC_{50}s. The statistical error presented for the JAWS phase two runs is the standard error of the mean, or mean unsigned error, obtained by running two independent simulations using a different random number seed[6].

5.4 JAWS simulation protocol

At the beginning of a JAWS simulation the binding site must be defined. Here, a three-dimensional grid is formed by overlapping spheres centred on user-selected sets of atomic coordinates that cover the area where hydration should be determined. The origin of the grid used lies at 38.0, 34.0 and 37.0 for x-, y-, and z- cartesian coordinates and extends to 10, 12 and 11 Å in x-, y-, and z-direction for all complexes generated. A default radius of the spheres of 2.5 Å was applied. The resulting volume was filled with grid points, evenly spaced at 1 Å in three dimensions.

The algorithm then consists of two distinct phases: finding the potential hydration sites, i.e. JAWS phase one, and then determining their occupancies, i.e. JAWS phase two. In the implementation of M. Bodnarchuk this is a two step approach in the sense that at least 2 separate simulations need to be run for the determination of the binding free energy of a single water molecule, while Michel proposed an implementation that allows the determination of the site as well as its occupancy in a single simulation[123]. The θ-water molecules were constructed based on the TIP4P water model using packmol and resulting water boxes had a volume of 1320 Å^3 and contained 100 water molecules, hence they represent a densely packed water box that exceeds the density of normal TIP4P water at 298 K and 1 atm.

The entire simulation system consists of the protein scoop, the inhibitor, the cofactor and the substrate as well as θ-water molecules, i.e. the box of θ-water molecules,

and regular water molecules, i.e. water molecules from the solvating TIP4P sphere, i.e. the standard solvent. In the first phase, putative hydration sites are detected by placing N θ-water molecules randomly on the grid. The number N can be manually specified, and a default of 1 water molecule per 30 Å^3 of grid volume is added randomly onto the grid. This corresponds to the density of liquid water at 298 K and 1 atm, a reasonable upper estimate of the water density in the binding site[123]. An MC simulation is performed that includes allowing the N θ-water molecules to sample freely the entire grid without any biasing potentials $V(\theta_i)$. An attempted MC move of a θ-water includes rigid body translations and rotations and, for 50% of the attempted moves, a random variation of its θ_i value, while the insertion of a new θ-water is attempted 23% of the time. The remainder are solvent, protein resiude and solute Metropolis MC moves attempted 23 %, 3.5 %, and 0.5 % of the time.

Initial equilibration was performed for 5M configurations in which only water molecules were allowed to move, followed by 10M moves of general equilibration, and 30M moves of data collection where all parts of the system were moved, i.e. θ_i, bond angles and dihedral angles for protein side chains, as well as all degrees of freedom for the inhibitors, FMN and ORO, with the exception of rings and bond lengths. To determine the most likely hydration sites, the θ_i values are inspected every 100 MC steps; if a θ_i is greater than a threshold value, the θ_i-water molecule is considered water-like and the coordinates of its neareast grid point are being saved. Here, we used a threshold of 0.995. As the simulation proceeds, the statistics collected in this fashion on the grid are interpreted as a probability distribution of water occupancies. At the end of the simulation, the fractional water occupancies are converted into an integer number of potential hydration sites, and a grid map is constructed using a procedure developed by Astex pharmaceutics. This grid file can be used in AstexViewer[234] to visualise all potential hydration sites. The procedure to generate the grid map can be summarised in the following steps:

1. A pdb is constructed containing all unique water grid points for water-like θ_i water molecules and their number of appearance over the entire simulation

2. Looping over the lines in the pdb file and adding the coordinates and number of appearance to a dictionary

3. Normalise all values based on the maximum appearance of a certain grid point

4. Generate a box that is large enough to contain the coordinates in the dictionary
5. Generate a grid based on the box dimensions, and with a grid spacing of 0.2 Å
6. Loop over coordinates in the dictionary
7. Mark a sphere of radius 1 Å at the coordinates using the normalised values
8. Save the resulting grid map

Once all potential hydration sites have been determined, a θ-water molecule is then positioned at its potential hydration site and a new simulation is attempted. The exact positioning of the water molecule is based on the coordinates of the oxygen and it is constrained to occupy a volume corresponding to a 27 Å^3 cube.

For this purpose the solvating sphere of TIP4P water molecules is being changed, such that the sphere now contains all waters of the former TIP4P sphere, all potential θ-waters suggested in the previous simulation and converted to standard TIP4P water molecules, but not the particular θ-water molecule that has been selected for phase two of the algorithm. This is because the final hydration pattern suggested by JAWS phase one was allowed to build up while the simulation was progressing, and hence omitting any of those waters in JAWS phase two, could potentially harm the sensitive cooperativity that is a featured of clusters of water molecules.

For the single θ-water molecule, the biasing potential $V(\theta_i)$ is turned on and statistics about the θ values are collected in a new MC simulation for this now constrained water molecule, i.e. using the hardwall constraint described previously. As for JAWS phase one, identical move ratios are applied, and data was collected for 40M MC moves, after a phase of 10M moves of general equilibration. A binding free energy is then estimated from the ratio of probabilities of observing this θ-water at high ($\theta > 0.995$) or low ($\theta < 0.05$) θ values.

This process is then repeated for every water molecule identified in JAWS phase one, revealing the complete hydration pattern detected by JAWS with their corresponding free energies of binding. This free energy estimate has so far, given evidence in the literature, reproduced free energies of binding compared to the more rigorous double decoupling method[123].

5.5 GCMC simulation protocol

Similar to JAWS simulations, the GCMC region must be defined at the beginning of the simulation. To create similar simulation conditions, the origin of this site was defined identically to our JAWS protocol, i.e. 38.0, 34.0 and 37.0 for x-, y-, and z-cartesian coordinates extending to 10, 12 and 11 Å in x-, y-, and z-direction. A set of grid points is then generated along these coordinates, using a grid spacing of 1 Å.

In subsequent insertion or deletion attempts, random grid points are tested against a 2.5 Å cutoff for any other atoms and, when successful based on equations 5.10, a water molecule is inserted or deleted, while inserted water molecules may also be dislocated according to the acceptance test for the standard canonical procedure in equation 5.10.

In all simulations, all complexes were equilibrated for 10M moves in the NVT ensemble to remove bad contacts, and only allowing water molecules to move, followed by a further 20M moves of equilibration where the sidechains of protein residues, all degrees of freedom for the solutes, cofactor, and substrate with the exception of rings and bond lengths, and rigid body translations and rotations for the solvent were sampled.

In all simulations solvent moves were attempted 44% of the time, protein moves were attempted 6.5% of the time, insertion/deletion were attempted 2.4% of the time, and translational solvent moves of inserted water molecules were attempted 44% of the time.

To allow an exchange of water molecules, one has to define the necessary reservoir of test particles. For the simulations performed, the same reservoir of water molecules was used as in our JAWS simulations using 100 water molecules in a box of 1320 $Å^3$.

Subsequent GCMC was carried out analogously to the simulated annealing of the chemical potential approach using the custom version of ProtoMS supplied by M. Bodnarchuk, and thus, separate simulations were run for varying values of Adam's B factor. As mentioned earlier, all simulations were conducted using a protein with flexible sidechains, while the backbone was kept rigid. However, additional runs were performed where full protein flexibility was allowed. Protein flexibility did not change the resulting hydration networks presented here, and hence results reported all refer

to simulation conditions with a flexible side-chain but a rigid backbone of protein residues.

GCMC simulations at different values of B will provide us with a rough estimate of the free energy associated with the inserted water molecules. As outlined above, at the beginning of the simulation, the defined region is flooded and weakly binding waters may be accommodated. As the value of B is lowered, weaker binding waters will be deleted until the strongest binding waters will remain, and finally all waters leave the region, as the value of B is decreased further. To compare the results obtained by this approach with the results provided by JAWS, a post-processing method of the simulation snapshots was used to generate grid maps. One problem associated with the conversion of snapshots into grid maps is that actual water coordinates are recorded in GCMC, instead of grid points that account for the appearance of water molecules in JAWS. Hence, resulting coordinates for water molecules from our GCMC runs were rounded to the nearest integer. This could potentially introduce a deviation of roughly 0.5 Å. However, this approach was adopted here as it has been shown to greatly assist and complement the findings using JAWS.

Thus, instead of iteratively finding the correct B value for a particular water molecule, we follow here a different approach. The grid maps created are being used more intuitively to assess the change of hydration with respect to B and compare these to the grid maps provided by JAWS. Water molecules found to be essential are subsequently taken from our GCMC runs and are fed into a phase two run of the JAWS algorithm, providing us with a reliable estimate for its free energy of binding. This is fortunate, as the fine-tuned annealing of B is not necessary, and thus alleviating computer resources. As outlined below, we also attempted to validate our findings for critical water molecules using the rigorous double decoupling approach. This was accomplished by turning off the partial charges gradually, followed by turning-off gradually the Lennard-Jones terms on the oxygen atom of the water molecule.

In all simulations using the double decoupling method, evaluation of the potential energy employed a 10 Å residue-based cutoff. To guarantee convergence, a well-defined volume for the water being annihilated was sought. This was realised with a spherical hard-wall constraint of 1.4 Å radius, positioned at the initial oxygen coordinates and prohibited to leave its volume, i.e. constraint. Furthermore, other molecules or atoms were not permitted to diffuse into this excluded region.

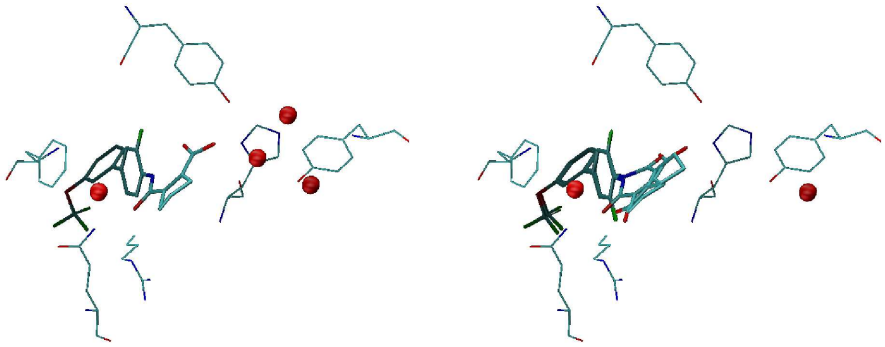
The double decoupling approach was carried out using 16 evenly spaced windows of RETI[40]. Initial coordinates came from the preceding JAWS run. For each window, further equilibration began with 10M MC configurations of water-only sampling, followed by 20M MC moves of general equilibration and 40M MC moves of data collection.

5.6 Results

This section describes the results obtained using the JAWS method and a GCMC approach to define hydration states in the protein DHODH complexed to the Baumgartner series of inhibitors. A maximum of four water molecules has been observed crystallographically[48]. In the following we use the same notation for these water molecules as in figure 5.1. Water molecule A lies in an enclosed cavity, is stacking with a PHE ring and may interact with the inhibitor. Water B, C, and D lie in a tunnel that exits towards the bulk solvent with water D lying almost at the exit of the tunnel while stabilising the conformation of a nearby HIS residue and H-bonding with other water molecules. Water C lies between waters B and D, is H-bonded with the carbonyl group of a GLY resiude and with waters B and D. Water B was described a a structural water molecule that bridges the interaction of the inhibitor with a TYR residue, while additional H-bonding may exist with water C.

Although the Baumgartner series shows all signs of a strictly congeneric series of compounds, crystallographically resolved water molecules have been observed to change upon complexing the different inhibitors[48]. The crystal structures of compounds 3 and 4, with PDB codes 2BXV and 2FPT respectively, are shown in figure 5.4 as an example.

Given the minute structural alterations on the inhibitors, it is surprising to find between two and four water molecules in the binding pockets. This is shown in figure 5.4. It is important to remember that these hydration patterns are in fact the sum of each individual binding mode of each inhibitor. For example, compound 3, shown in figure 5.4 (a) has been resolved in a single binding mode, while compound 4, shown in figure 5.4 (b), has been resolved as a dual binding mode inhibitor[48]. Hence, these pictures, or in other words the crystal structures associated with these dual binding modes, do in fact not allow one to unambiguously define the hydration



(a) (b)

Figure 5.4: Crystal structures for compound 3 (a) and compound 4 (b) showing the different hydration patterns observed crystallographically[48]. In these images the compounds are all shown in liquorice representation in a side orientation to allow for a clearer differentiation of both hydration sites observed, while crystallographic water molecules are represented using a red sphere representation for the oxygens only. For clarity, only key resiudes are shown in line representation. The pictures have been created using VMD[235].

pattern of each binding mode, and thus any discussion of the experimentally observed hydration pattern becomes arbitrary and underlines the need for a thorough assessment of them.

As outlined above, we employed a combined approach to achieve this. In summary, we assess the hydration states using the JAWS methodology[123] as well as a simulated annealing of the chemical potential GCMC approach using varying values for Adam's B factor. As will be explained in more detail later in this chapter, the grid maps provided by JAWS were difficult to interpret, leaving us with too many options for estimating the free energies of binding of water molecules using the JAWS phase two phase. However, we also find that the GCMC approach adopted here, results in more intuitive and cleaner grid maps, and subsequently allows the interpretation of our JAWS maps. Hence, we do not aim to calculate the exact free energy of binding for water molecules using the GCMC methodology, but instead we reinterpret our JAWS grid maps using GCMC, and make all waters identified by those subject to a JAWS phase two run, aiming to estimate the free energy of binding of each water molecule. This allows us to define the hydration pattern for each compound, and indeed each binding mode.

To further simplify the discussion, we start with an overall assessment of our

findings related to these hydration patterns in subsection 5.6.1. We then look at each compound, and binding mode, individually, and present our analysis in subsections 5.6.2, 5.6.3, and 5.6.4.

5.6.1 Overall assessment of hydration patterns in DHODH

The two hydration sites present in the current DHODH system have been the subject of GCMC studies. Water A has been identified as a strong binder by GCMC for all compounds, and indeed all binding modes. The estimation of the free energy of binding of water molecule A was performed using JAWS phase two and revealed free energies of between -2.4 and -6.1 kcal mol^{-1}.

The triad of water molecules, consisting of water B, C and D, does, however, not appear in all structures. Interestingly, the hydration pattern for the non-brequinar binding mode is identical for all compounds, while for the brequinar binding mode, changes of the hydration pattern have been observed. The protein-ligand complexes of compounds 3 and 4 show an identical hydration pattern with all 3 waters present in both binding modes; hence compounds 3 and 4 do not differ in their hydration pattern in any binding mode. For compounds 5, 6 and 7 in the brequinar binding mode, the triad is reduced to two water molecules, while for the non-brequinar binding mode of compounds 5, 6 and 7 the triad was found to be energetically prefered.

The free energies of binding according to JAWS phase two for water molecules A, B, C and D that were previously identified by GCMC are given in table 5.1. In this table we use BQ and NBQ to indicate the brequinar and non-brequinar binding mode respectively, and use the same notation for the water molecules as previously, i.e. waters A, B, C and D.

Given the process of generating a grid map for both, GCMC and JAWS, it raises the question of how well resolved the actual water placement coordinates are. Clearly, the rounding of coordinates in GCMC, and the detection of not the actual water coordinates but their representative grid point in JAWS, may introduce some noise leading to subtle differences in the placement of these waters. This is illustrated in figure 5.5 on the example of a GCMC simulation for water D, i.e. hydration cluster two. The grid map in this picture has been generated with rounded coordinates of the oxygen atoms of the water molecules. The actual coordinates of these oxygens that were used to generate the grid map are shown in red. It can be seen that the

	BQ				NBQ			
	A	B	C	D	A	B	C	D
compound 3	-4.1	-2.6	-2.3	-5.1	-5.1	-3.9	-5.0	-4.8
compound 4	-3.0	-2.4	-2.5	-3.8	-4.0	-4.0	-4.2	-4.0
compound 5	-2.4	-2.2	+1.5	-2.0	-2.0	-5.0	-3.5	-2.5
compound 6	-6.1	+6.5	-2.5	-6.7	-4.6	-3.6	-1.8	-4.0
compound 7	-3.1	+5.8	-4.5	-3.0	-3.9	-1.5	-5.5	-3.6

Table 5.1: Estimated free energies of binding using the JAWS phase two algorithm for waters A, B, C and D, for the brequinar (BQ) and non-brequinar (NBQ) binding mode. All figures are given in kcal mol^{-1} and are a result from at least 2 independent simulations, using a different random number seed. The mean unsigned error associated with these energies are below 0.3 kcal mol^{-1} for all cases.

grid map covers a significant amount of the actual coordinates of the oxygen atoms. Nevertheless parts of the recorded coordinates of the oxygen atoms are lying outside the grid map sphere.

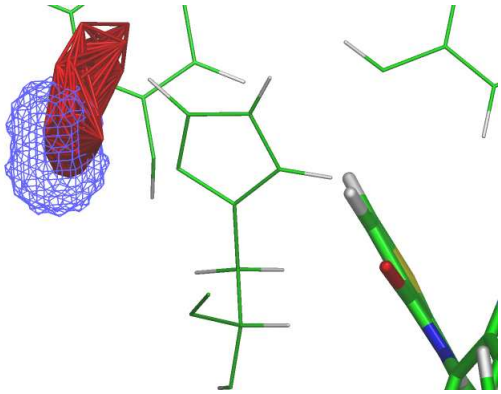


Figure 5.5: Differences for the coordinates of water molecules resulting from the rounding procedure. The blue grid maps show the location of the suggested site for water D. This is based on rounded coordinates, while the red blob partly lying within the blue grid map sphere are the actual oxygen coordinates observed for this simulation. The picture has been created using OpenAstexViewer[234].

To test whether small differences in water placements would result in significantly different computed free energies of binding of the respective waters, we have run additional JAWS phase two simulations using alternate water coordinates. For example, according to the GCMC results, water C is present in the brequinar binding mode of compound 4, while it is energetically unfavourable for compound 5 in the brequinar binding mode. To test whether this difference is merely an artefact of using specific, and rounded, coordinates, or indeed is likely to result from an actual

change in hydration pattern, we use different simulation setups: Two simulations where we use the actual water coordinates from GCMC snapshots, while making sure that these coordinates are distinctly different. Further simulations where then run that use coordinates for the water molecule such that it is lying inside the grid map for both, the water predicted to be favourable in compound 4 as well as the water predicted unfavourable in compound 5. Additionally, we also use the coordinates provided in the crystal structures in further GCMC simulations. In principle, this will result in slightly different energies, as in JAWS phase two we use a hardwall constraint, leading to very particular and well defined water locations. However, none of these simulations did affect the conclusions for the predicted water networks, and hence all free energies reported here refer to the actual predicted water locations via GCMC.

5.6.2 Detailed results for compounds 3 and 4

The crystallographic structures of compounds 3 and 4, PDB codes 2BXV and 2FPT respectively[48], in principle, suggest a dramatic change in hydration pattern, as can be seen in figure 5.4 on page 118. As with all structures of the Baumgartner series, water A is present in 2BXV and 2FPT, but the second hydration cluster is reduced from a full triad of waters in 2BXV to a single water molecule that stabilises the nearby HIS residue, and thus has been proposed to direct the protonation state of this residue. It is more intuitive that this water, i.e. water D, will be present, given its proximity to bulk solvent.

Figures 5.6, 5.7, 5.8 and 5.9, on pages 124 to 127, graphically illustrate the findings of JAWS as well as GCMC simulations for compounds 3 and 4 in the brequinar and non-brequinar binding mode respectively, using a grid representation for the water clusters detected. A visual inspection of these figures reveals the usefulness of a combined approach to assess hydration sites in DHODH.

JAWS, although picking up very similar sites compared to GCMC, appears less detailed and more diffuse. The representative grid maps are shown in figures 5.6 (a), 5.7 (a), 5.8 (a) and 5.9 (a) using green contours for compounds 3 and 4 in the brequina and non-brequinar binding mode respectively. For JAWS phase two, the hydration sites observed in the map need to be distilled down to individual water

molecule locations. Unless the sites are clearly identified in this way, the allocation of waters is very problematic leading to unoptimised hydration networks.

On the other hand, the figures from the GCMC method in a simulated annealing of the chemical potential approach employed here, appear much cleaner. This is shown in figures 5.6 to 5.9, subfigures (b) to (f). At high values of B, thus positive chemical potentials, the binding site is flooded and appears similar to the one observed for the JAWS analysis. As B is decreased, i.e. the chemical potential is decreased, weaker binding waters disappear until only tightly binding water molecules are present in the system. However, and most importantly, it becomes clear how grid maps created for the GCMC runs may be used to reinterpret the grid maps resulting from the JAWS analysis phase one.

Although the cooperativity of water molecules is in principle incorporated in the JAWS algorithm - in fact the dominating ensemble naturally evolves during the simulation as it does in the GCMC simulations- it is not trivial to see in the JAWS grid maps.

Therefore, in the present study, the JAWS grid maps were reinterpreted using the GCMC results at different values of B. Values of B that correspond to a favourable binding free energy of the water molecules, i.e. a value of $B = -10$ roughly corresponds to a chemical potential of -2.0 kcal mol^{-1}, have subsequently been selected and the coordinates of the inserted water molecules extracted. These water molecules were then subject to a JAWS phase two simulation, i.e. an individual water molecule constrained with a hardwall is sampled in terms of on and off states as well as rotational degrees of freedom, while the solvating sphere of TIP4P molecules now also contains all other water molecules that were identified using GCMC and appeared to be favourable binders. For estimates of the free energies of binding of individual water molecules, please refer to table 5.1 on page 120.

According to the simulations, no change in hydration pattern is observed for either compound 3 or 4 and for either binding mode. This result is somewhat surprising, as Baumgartner reports that *" the water molecule bridging the hydrogen bond between the carboxy group and TYR147 is present in both structures"*, where *'both structures'* is a referral to compounds 4 and 5 in the literature[48]. However, this water is not present in the PDB structure 2FPT, i.e. compound 5. Finally, we propose a consistent

hydration pattern for compound 3 and 4 and for both binding modes consisting of water A and the triad of waters at the second hydration cluster in DHODH.

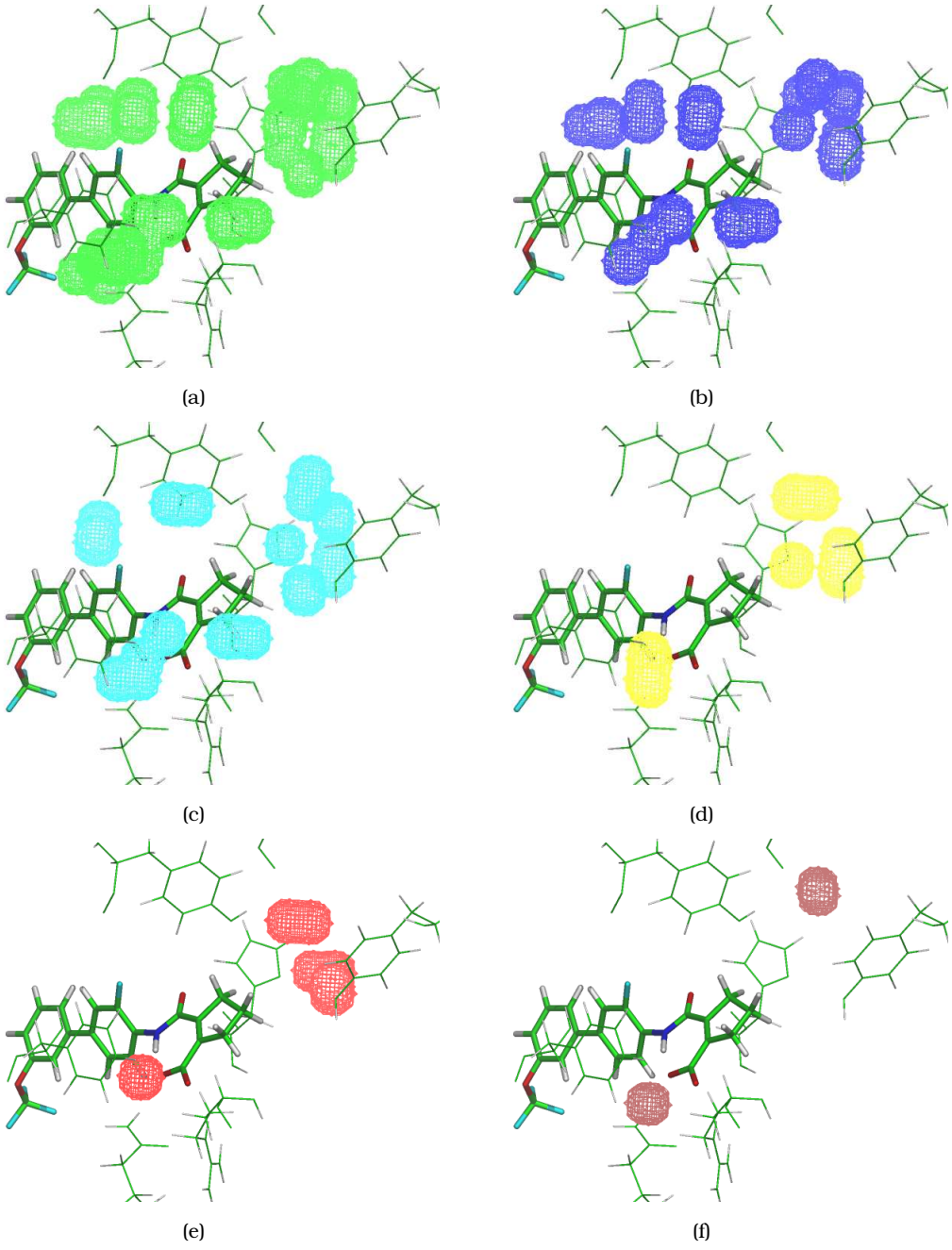


Figure 5.6: Hydration pattern for compound 3 in brequinar binding mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). Only the inhibitor molecules, shown in stick representation, and key residues, shown in line representation, are shown for clarity. All figures have been created using OpenAstexViewer[234].

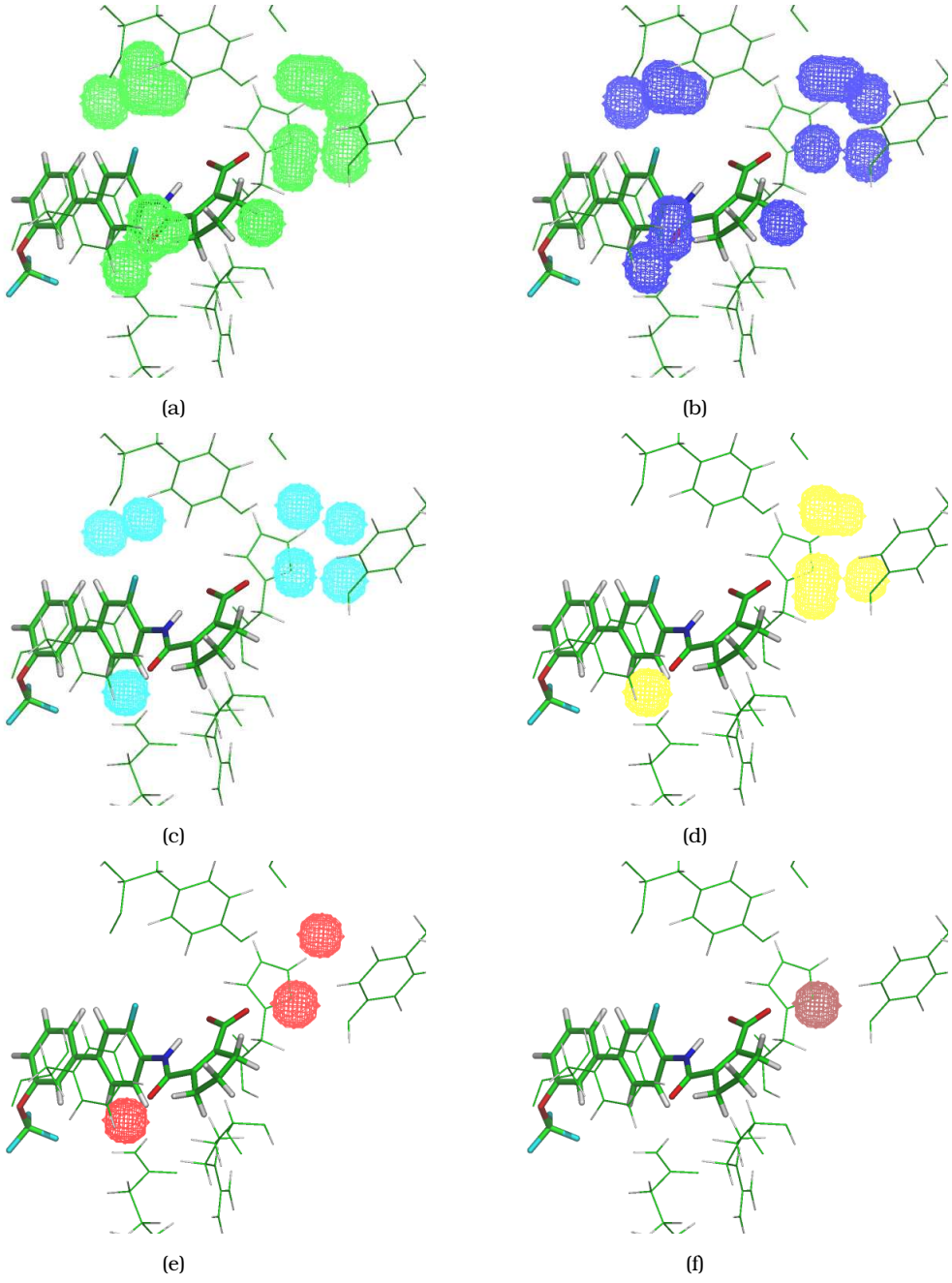


Figure 5.7: Hydration pattern for compound 3 in the non-brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). Only the inhibitor molecules, shown in stick representation, and key residues, shown in line representation, are shown for clarity. All figures have been created using OpenAstexViewer[234].

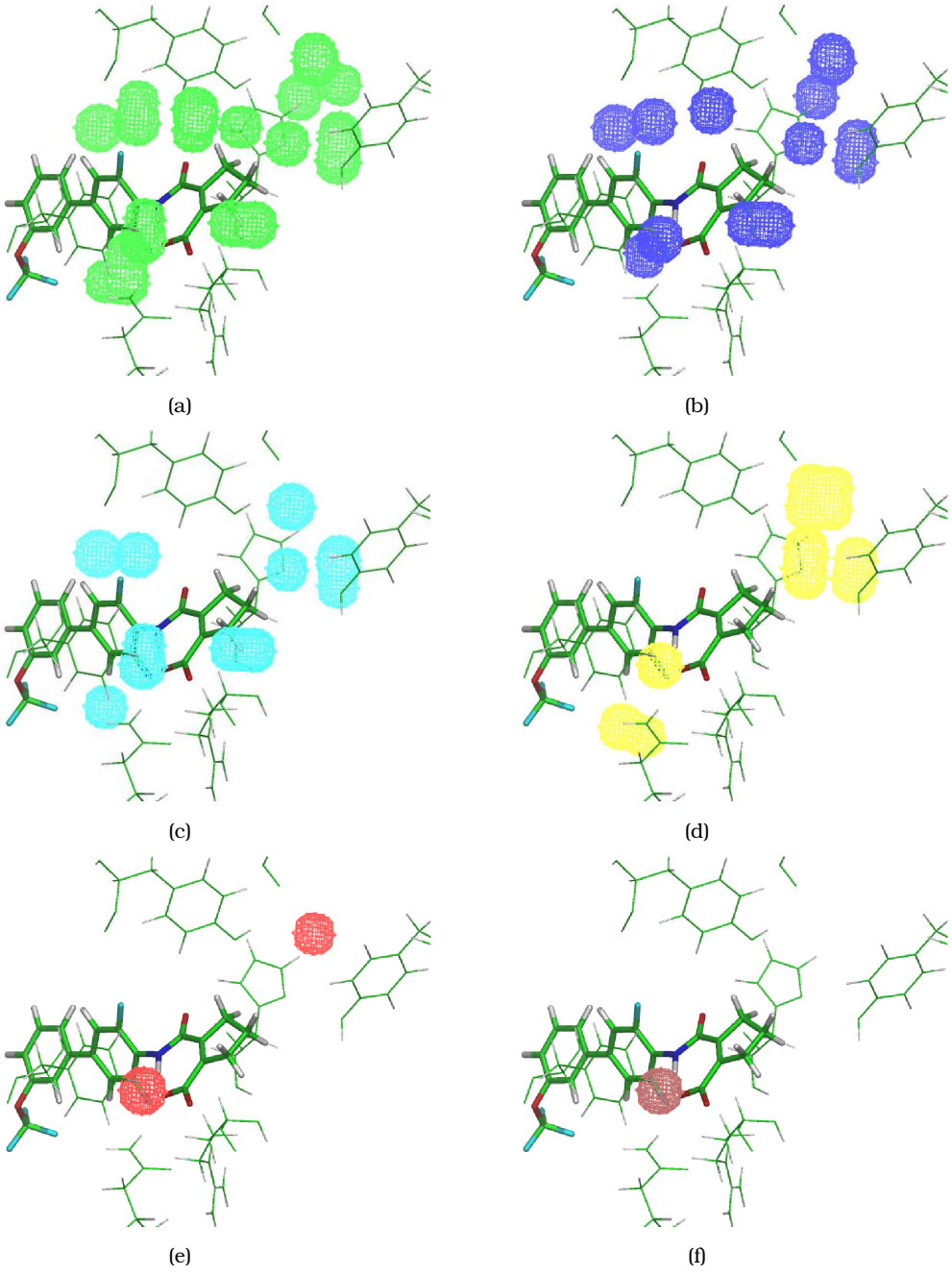


Figure 5.8: Hydration pattern for compound 4 in the brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). Only the inhibitor molecules, shown in stick representation, and key residues, shown in line representation, are shown for clarity. All figures have been created using OpenAstexViewer[234].

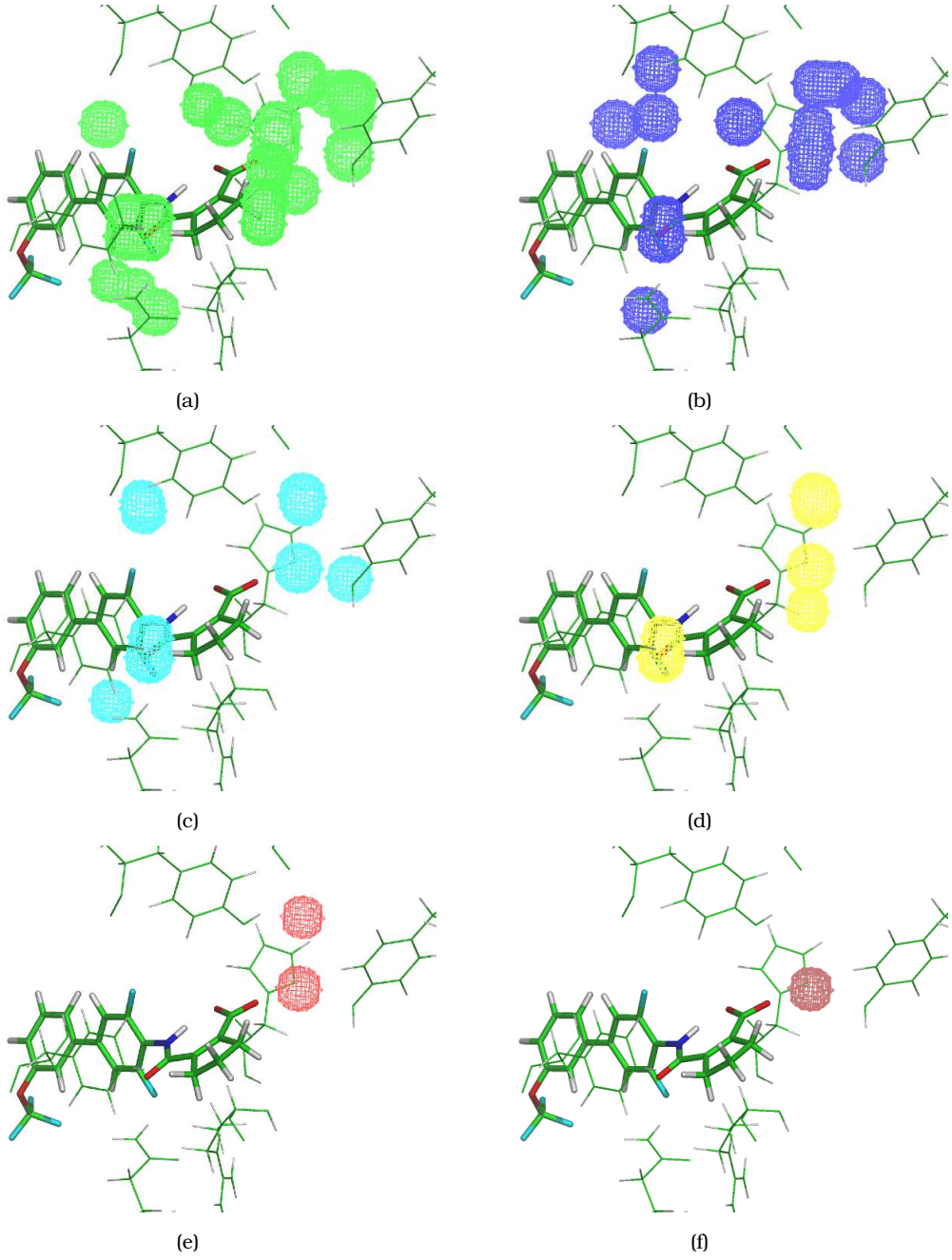
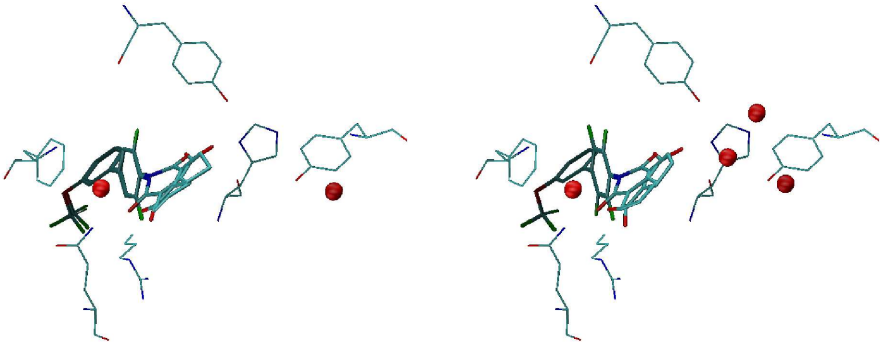


Figure 5.9: Hydration pattern for compound 4 in the non-brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). Only the inhibitor molecules, shown in stick representation, and key residues, shown in line representation, are shown for clarity. All figures have been created using OpenAstexViewer[234].

5.6.3 Detailed results for compound 5

The crystallographic structure of compounds 5, PDB code 2FQI, and as a comparison compound 4, PDB code 2FPT, are shown in figure 5.10. Again, even though the original publication reports an identical hydration pattern for compounds 4 and 5[48], inspection of figure 5.10, created from the original structures deposited in the PDB, reveal crystallographic differences for the number of water molecules present in the second hydration site. While compound 4 has a single water molecule, compound 5 has been resolved with a full triad of water molecules.



(a) (b)

Figure 5.10: Crystal structures for compound 4 (a) and compound 5 (b) showing the different hydration patterns observed crystallographically. In these images the compounds, shown in a side orientation to allow for a clearer differentiation of both hydration sites observed, are shown in liquorice representation, while crystallographic water molecules are represented using a red sphere representation for the oxygens only. For clarity, only key resiudes are shown in line representation. The pictures have been created using VMD[235].

Figures 5.11 and 5.12 on pages 132 to 133, graphically illustrate the findings of JAWS as well as GCMC simulations for compound 5 in the brequinar and non-brequinar binding mode respectively, using a grid representation for the water clusters detected. The general procedure to generate grid maps and the B values chosen were identical to the protocol detailed in the last section for compounds 3 and 4.

Again JAWS grid maps appear to correspond to GCMC at high values of Adam's B factor, while upon lowering B more and more waters leave the system. However, for compound 5 also the GCMC maps appear more difficult to interpret, as they seem less detailed especially for the second hydration cluster with waters B, C and D. We

have decided to look at all snapshots and found that the triad is present for the non-brequinar binding mode, but not for the brequinar binding mode. Additionally, we have subsequently selected various sets of coordinates for waters B, C and D as explained earlier, including the coordinates used for the study of compound 4 in the brequinar binding mode. However, all runs performed did not change the fact that for compound 5 in the brequinar binding mode, the triad seems to be reduced to two water molecules. The energies using coordinates that correspond to a representative snapshot of the GCMC simulations are shown in table 5.1 on page 120, indicating that water C is according to JAWS phase two not part of a triad.

Given that this water molecule is in fact not in direct contact with the ligand, this result is surprising. However, the attachment of 4 fluorine atoms, as compared to 2 fluorine atoms in compound 4, leads to subtle differences in sampling the torsional angle between the biphenyl moiety in the simulations, which translates to the amide moiety of the inhibitor. The effect is that water B slightly reorientates, again compared to compound 4, and as a consequence water C becomes more unfavourable and potentially loses ideal H-bonding conditions with waters B and possibly water D.

As outlined in the introduction of this chapter the most rigorous approach one could apply to test these predictions is to validate water C, by using the rigorous double decoupling method. Here we annihilate water C in the brequinar binding mode using a set of coordinates from our GCMC simulations at a value of 10 for Adams B factor. We achieve this by performing a set of simulations, where we decouple the water molecule from its protein binding site. For decoupling the water from bulk solvent we make use of the free energy estimates published by our group [122], i.e. 6.4 kcal mol^{-1}.

As described earlier, during the decoupling process from the protein binding site, we employ a hardwall constraint to guarantee reversibility. This constraint prevents atoms occupying any region within the constrained area, as the water molecule is slowly decoupled from its environment. The hard-wall potential was centred on the oxygen atom of the water molecule to be annihilated. Different radii were tried and a radius of 1.4 Å was found appropriate.

To calculate the decoupling energy of a water molecule from a protein binding site, we calculate $\Delta G_{trans}(water, site)$ and $\Delta G_{constr}(ideal, site)$, as explained in the in-

troduction of this chapter. The decoupling energy then comprises of

$$\Delta G_{dec} = \Delta G_{trans}(water, site) + \Delta G_{constr}(ideal, site) - RT\ ln\ \frac{\sigma_{RS}}{\sigma_R \sigma_S} + P^0(V_R - V_{RS}) \quad (5.13)$$

For the current setup using a 1.4 Å radius hardwall ΔG_{constr} becomes -0.57 kcal mol^{-1}, R is the gas constant, T the temperature, σ_{RS}, σ_R and σ_S are the symmetry numbers of the complex, the protein and the substrate respectively; since water has a symmetry number of 2 this term equals -0.4 kcal mol^{-1}; P^0 is the standard pressure and $V_R - V_S$ represents the volume change of the system when the substrate is decoupled from the protein in a constant pressure simulation; at normal pressures this term may be neglected. Finally, the absolute free energy of annihilating a water molecule is obtained as the difference of decoupling the water from bulk solvent and from the protein, i.e. 6.4 minus ΔG_{dec}.

For the simulation of $\Delta G_{constr}(ideal, site)$ for annihilating water C, we first gradually switch off the electrostatic interactions and then, in a second simulation, we turn off the Lennard-Jones contributions between the water and the remainder of the system. The simulations are carried out at 298 K and P = 1 atm with our in-house software ProtoMS[47] and using the scoop setup described earlier in this chapter. Free energies were calculated with the RETI method[40], and non-bonded interactions were evaluated up to 10 Å while a feathering function was employed to gradually switch the interactions to zero for the last 0.5 Å. Each system was then equilibrated for 10 M MC moves where only solvent molecules were allowed to move, followed by 20M moves of general equilibration, i.e. solutes, solvents and protein was allowed to move. For the solutes all angles and torsions were sampled with the exception of rings, and the protein sidechains were allowed to move while the backbone was kept flexible. For validating purposes we also used an entirely flexible protein, yielding very similar results compared to the simulations applying a rigid backbone. For each simulation 16 evenly spaced values of λ were used and data was collected for 50M moves.

The energies associated with turning off the electrostatic interactions of the water with its protein environment lie at 0.48 $\pm$ 0.1 kcal mol^{-1}, while subsequently turning off the Lennard-Jones contributions stands at 8.4 $\pm$ 0.2 kcal mol^{-1}. These estimates are from 2 independent simulations using a different random number seed and the errors given correspond to the mean unsigned errors. Hence, the decoupling energy for the protein binding site stands at 8.71 kcal mol^{-1}, which results in an absolute

free energy of annihilation of this water molecule of -2.31 $\pm$ 0.05 kcal mol^{-1} which is in good agreement with our JAWS findings.

In summary, according to the GCMC and JAWS simulations, a change in hydration pattern is observed for compound 5 for the brequinar binding mode, while no change is observed for the non-brequinar binding mode. This change affects the triad of water molecules, thereby reducing them to two.

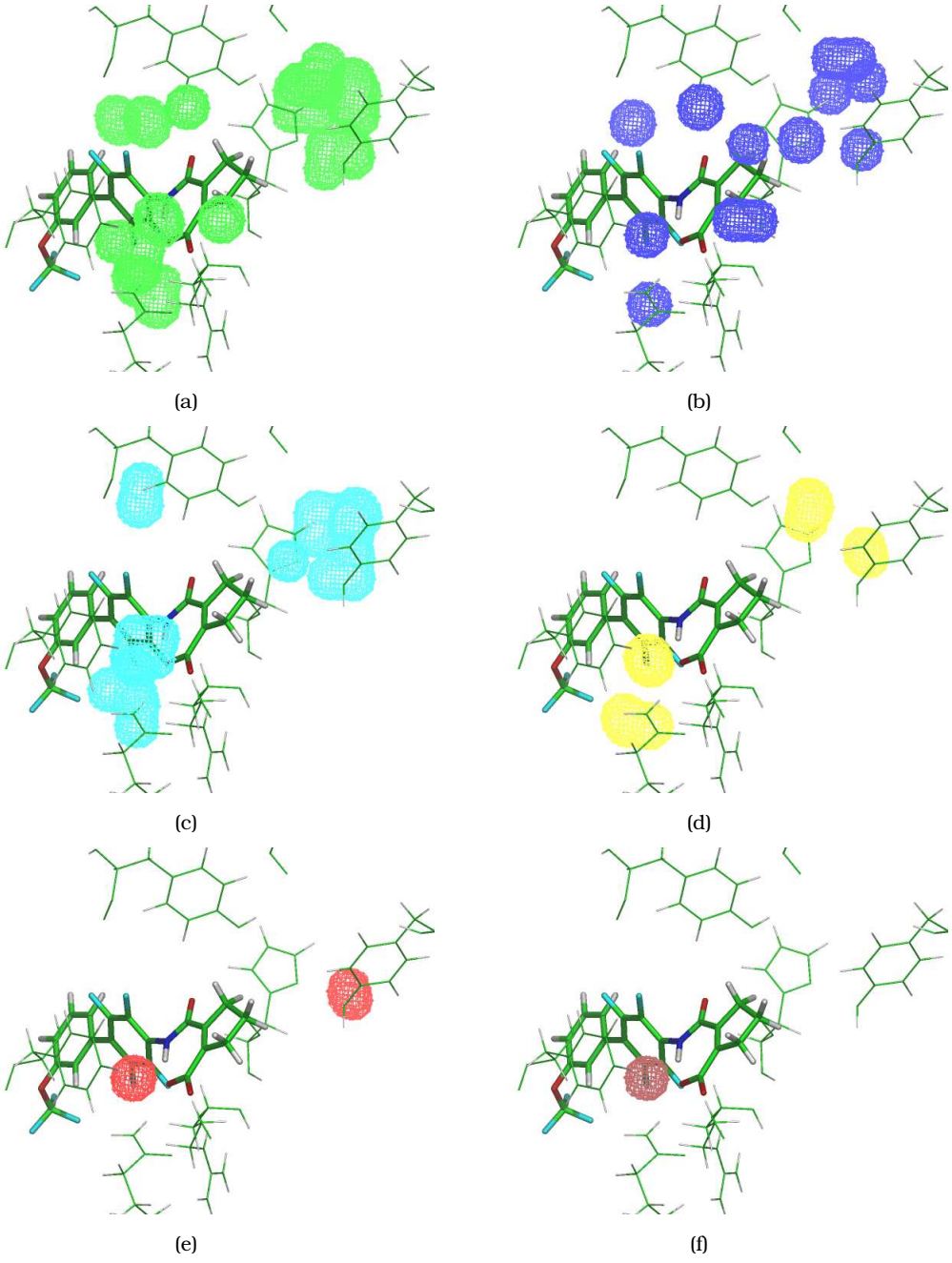


Figure 5.11: Hydration pattern for compound 5 in the brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). Only the inhibitor molecules, shown in stick representation, and key residues, shown in line representation, are shown for clarity. All figures have been created using OpenAstexViewer[234].

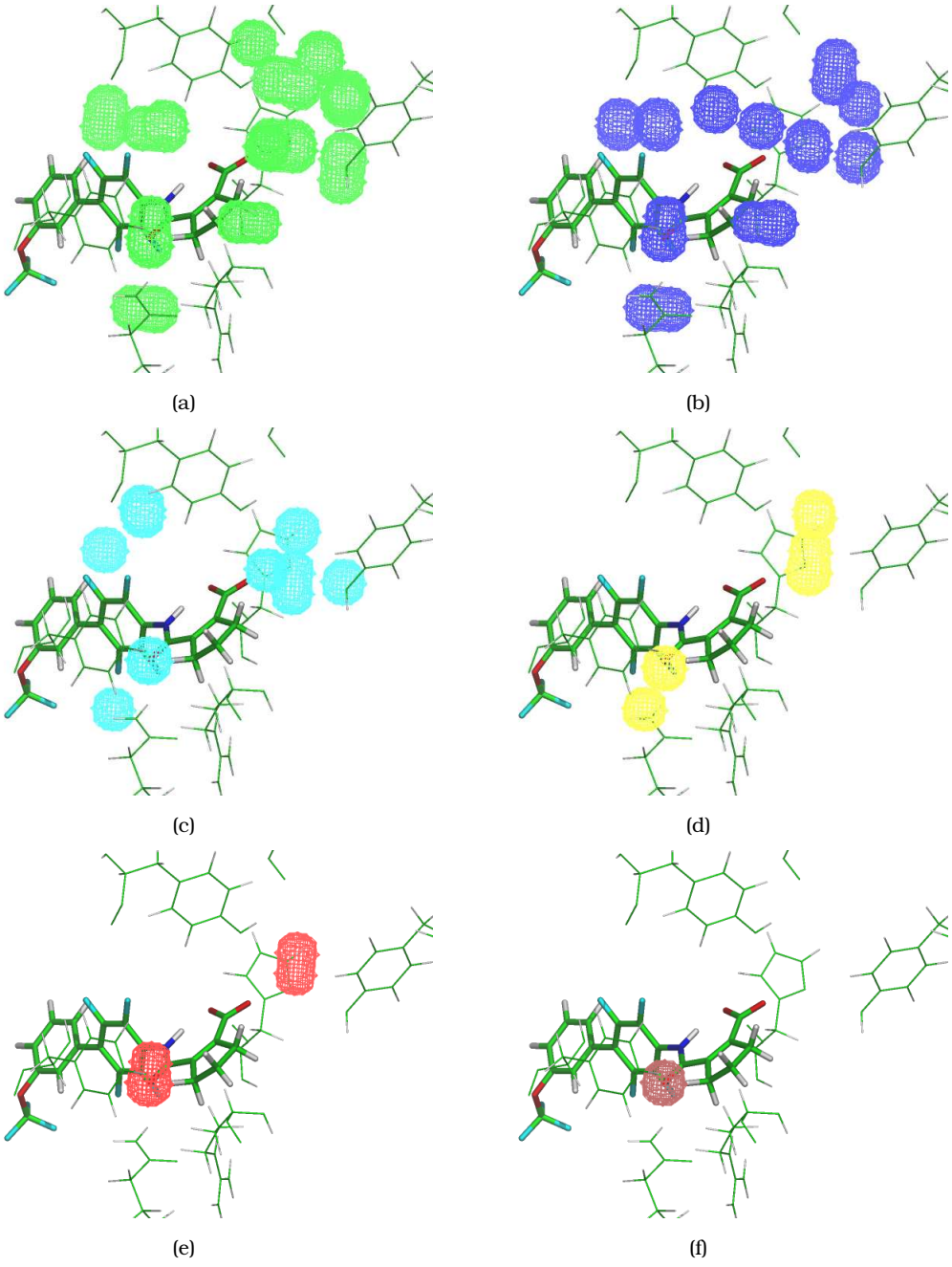
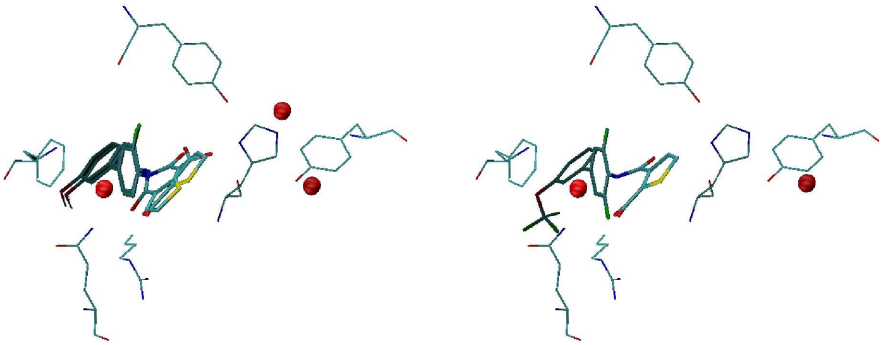


Figure 5.12: Hydration pattern for compound 5 in the non-brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). Only the inhibitor molecules, shown in stick representation, and key residues, shown in line representation, are shown for clarity. All figures have been created using OpenAstexViewer[234].

5.6.4 Detailed results for compounds 6 and 7

The crystallographic structure of compounds 6 and 7, PDB codes 2FPV and 2FPY respectively, are shown in figure 5.13. Baumgartner et al. proposed that the water molecule bridging the interaction between the inhibitor and a nearby TYR residue is present in compound 6 but not in compound 7[48]. However, inspection of figure 5.13, created from the original structures deposited in the PDB, reveals crystallographic differences for the number of water molecules present in the second hydration site, but they are not consistent with results presented in the paper for compound 6[48]. The water bridging these interactions is supposed to be water B, but it is missing in 2FPV; instead water C is present, and neither are present in 2FPY, i.e. compound 7.



(a) (b)

Figure 5.13: Crystal structures for compound 6 (a) and compound 7 (b) showing the different hydration patterns observed crystallographically. In these images the compounds, all shown in a side orientation to allow for a clearer differentiation of both hydration sites observed, are shown in liquorice representation, while crystallographic water molecules are represented using a red sphere representation for the oxygens only. For clarity, only key resiudes are shown in line representation. The pictures have been created using VMD[235].

Figures 5.14, 5.15, 5.16 and 5.17, on pages 138 to 139, graphically illustrate the findings of JAWS as well as GCMC simulations for compounds 6 and 7 in the brequinar and non-brequinar binding mode respectively, using a grid representation for the water densities detected. The general procedure to generate grid maps and the B values chosen were identical to the protocol detailed for compounds 3 and 4.

Again JAWS corresponds to GCMC at high values of Adam's B factor, while upon lowering B more and more waters leave the system. GCMC maps appear straightforward to interpret, and would suggest a change for the solvent exposed hydration

cluster, i.e. water B, C and D if we were to change the binding mode, while no change is observed when staying within one binding mode.

More specifically, for the brequinar binding mode of compounds 6 and 7 the triad is reduced to waters C and D, and thus reflects the structure deposited in the PDB for compound 6. The subsequent selection of various sets of coordinates for waters B, C and D, including the coordinates deposited in 2FPV, for the JAWS phase two algorithm, clearly confirms the presence of waters C and D, while water B is highly disfavoured, with the estimated free energies of binding for these waters given in table 5.1 on page 120.

An additional site was initially proposed using JAWS and GCMC for the brequinar binding mode. This site is in close proximity to water A and the inhibitor's carboxylate moiety. In the following, this water is called water A2. However, the on and off states for water A2 for both compounds 6 and 7, are insufficiently sampled, and in fact this water molecule is hardly turned off in all simulations. The electrostatic attraction of the formal charge present in the inhibitors may be mainly responsible for not allowing this water to be turned off. To aid sampling we have therefore increased the bias used in JAWS stage two, i.e. 6.4 kcal mol^{-1}, to +8 and +10, which subsequently allowed the sampling of on and off states. The calculated free energy of binding for this water molecule stands at approximately -2.0 kcal mol^{-1} for both compounds, indicating that there is an additional water site for compounds 6 and 7 in the brequinar binding mode.

For the non-brequinar binding mode, the hydration pattern is identical to all other compounds in the Baumgartner series, and hence no changes in hydration have been observed for this mode, allowing perturbations within the non-brequinar binding mode to be carried out readily without incorporating the annihilation or creation of waters. As for the results of the brequinar binding mode, the results obtained are shown in table 5.1.

In summary, waters A, A2, C and D define the hydration network according to JAWS and GCMC for compounds 6 and 7 in the brequinar binding mode complexed to DHODH, while for the non-brequinar binding mode waters A, B, C and D are present.

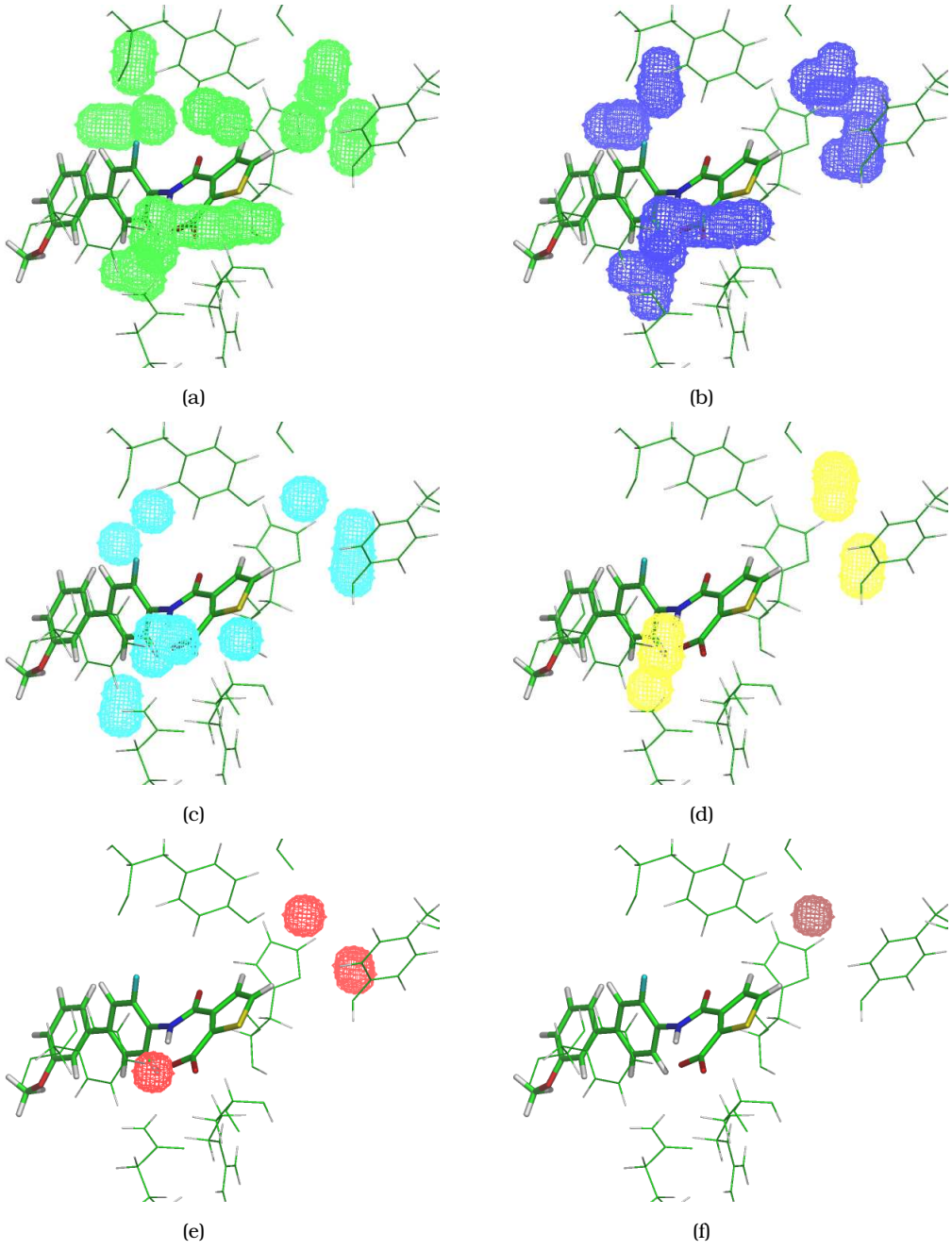


Figure 5.14: Hydration pattern for compound 6 in the brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). All figures have been created using OpenAstexViewer[234].

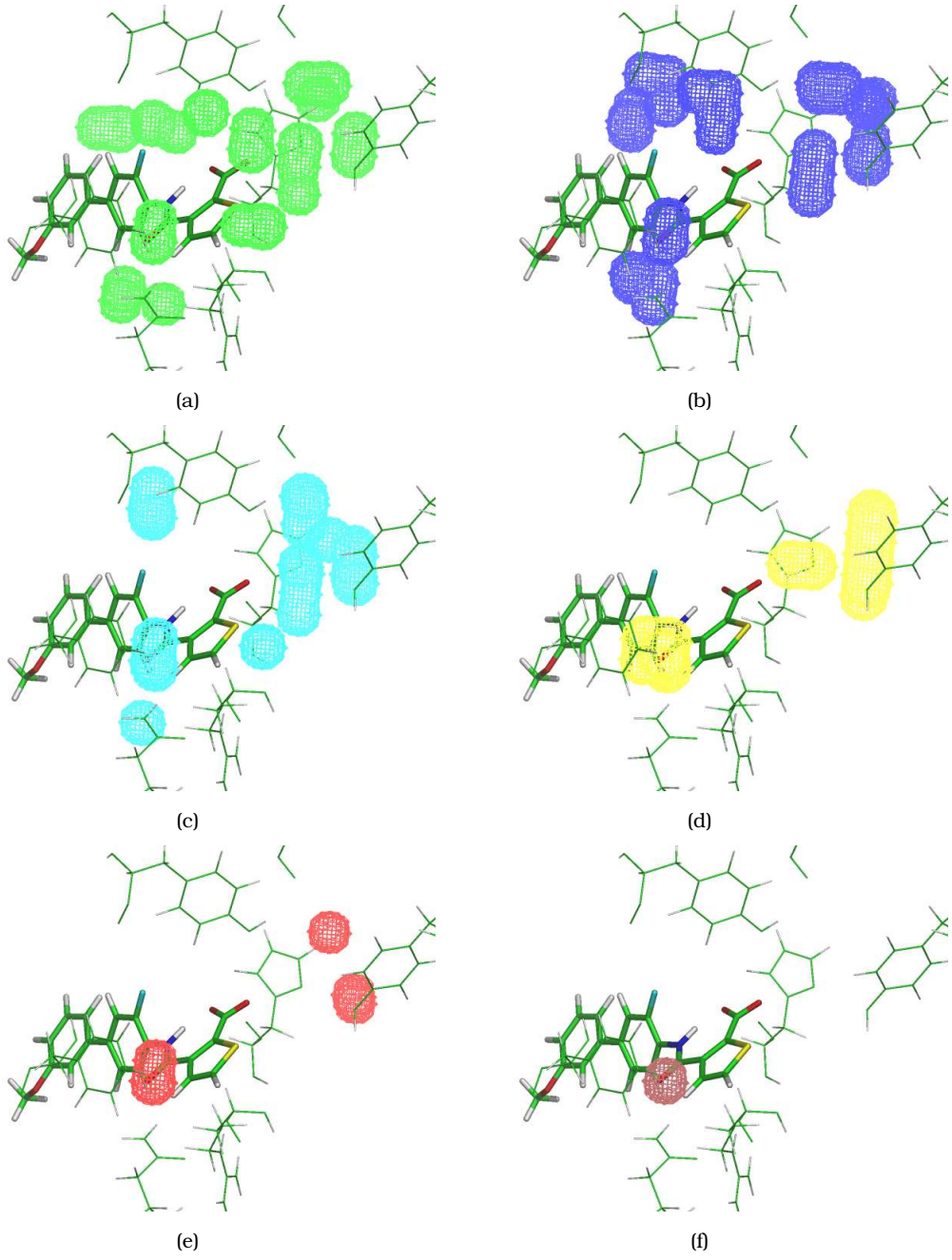


Figure 5.15: Hydration pattern for compound 6 in the non-brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). All figures have been created using OpenAstexViewer[234].

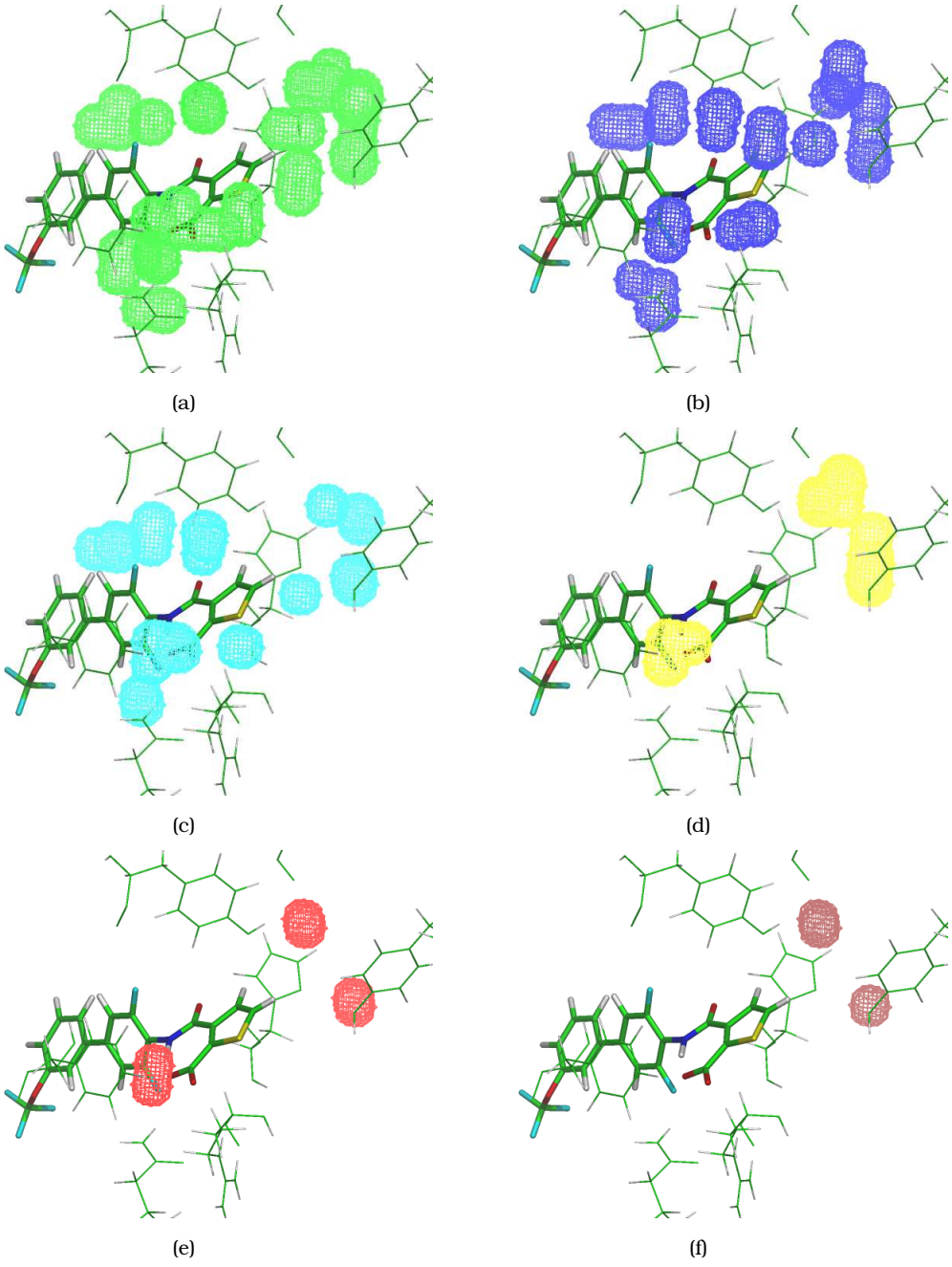


Figure 5.16: Hydration pattern for compound 7 in the brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f). All figures have been created using OpenAstexViewer[234].

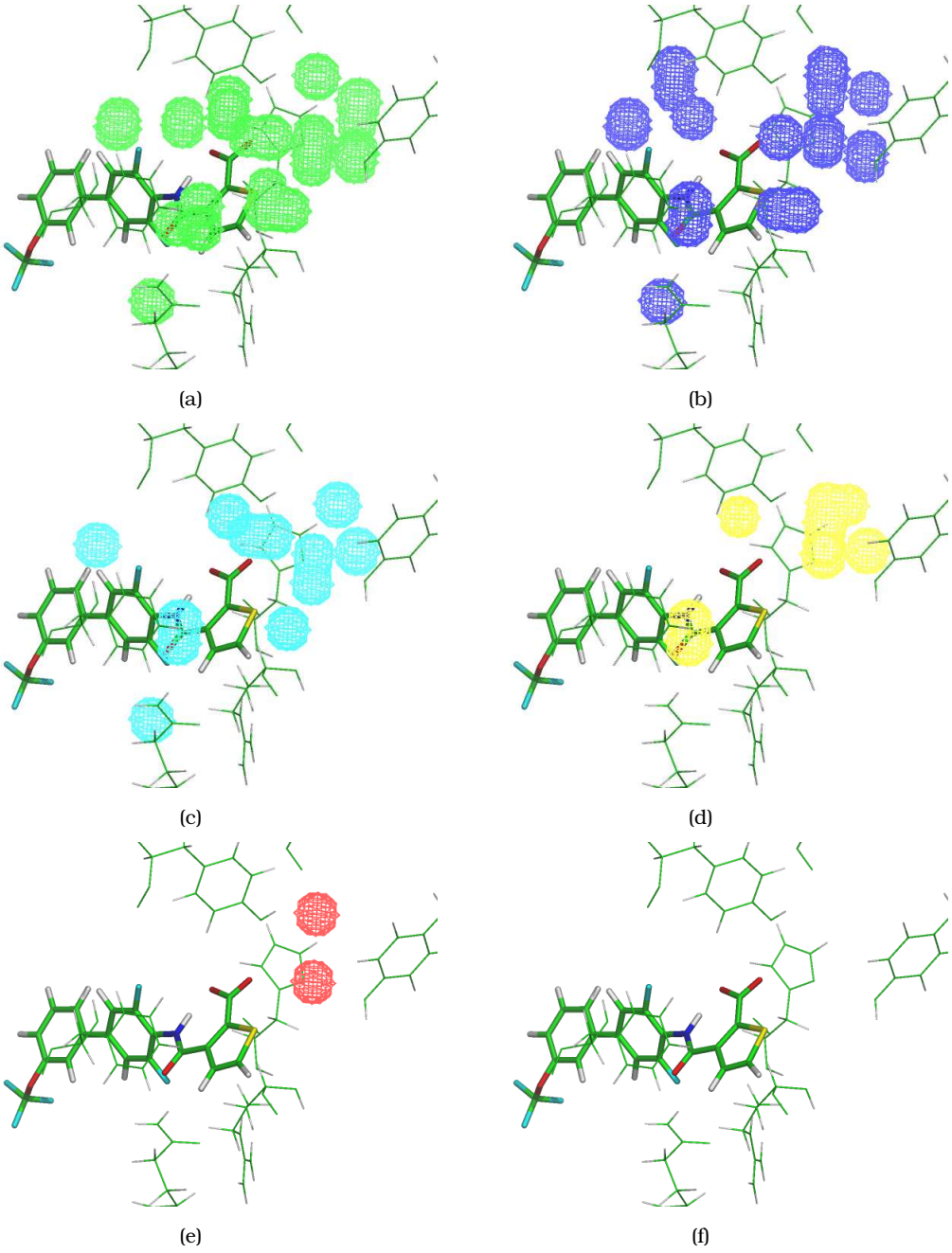


Figure 5.17: Hydration pattern for compound 7 in the non-brequinar mode using the JAWS method at a contour level of 0.2 (a), as well as GCMC simulations at contour levels between 0.2 and 0.4 with B values of 10 (b), 0 (c), -10 (d), -13 (e) and -17 (f).

5.7 Conclusions

The analysis of the water distributions for DHODH complexed to the Baumgartner series of inhibitors[48] was the subject of this chapter. A modified JAWS approach as well as GCMC simulations were used to aid in this task, while the rigorous double decoupling approach was used to ascertain critical waters.

We find that for the entire series in the non-brequinar binding mode, no change in hydration is observed. The single water A occurs consistently in all compounds, which is also true for all compounds in the brequinar binding mode, and a second cluster, consisting of waters B, C, and D, appears in all compounds for the non-brequinar binding mode.

For the brequinar binding mode, we find subtle differences in hydration. Compounds 3 and 4 show a full triad of waters B, C, and D, while for compounds 5, 6 and 7 the triad is reduced to two water molecules. These are waters B and D for compound 5, and waters C and D for compounds 6 and 7. For compounds 6 and 7 we find an additional water, i.e. water A2, that is weakly bound to the inhibitor's carboxylate moiety and may interact with nearby protein residues.

The definition of these hydration states, together with our previous testing of force field parameters, now allows us to define the end states for the Baumgartner series complexed to DHODH. These calculations aim to predict the binding modes and calculate the free energies of binding of the inhibitors and will be presented in the next chapter.

6

The prediction of binding modes and free energies of binding

Chapter 3 introduced in detail the Baumgartner series of inhibitors complexed to DHODH[48]. To capture well the problems associated with the thermodynamic end states, the hydration network of each of the inhibitors and each binding mode has been investigated in chapter 5. This provided us with more details on how to set up perturbations to study the free energies of binding and the prediction of binding modes. To allow the selection of an optimal force field for the simulations, we have assessed the accuracy of several combinations of force fields and partial atomic charge models. This was described in detail in chapter 4. The AM1BCC[49] atomic charge model in conjunction with GAFF[136] was then selected to describe the inhibitors, the cofactor and the natural substrate for subsequent simulations as it appeared to provide the most reasonable balance in terms of accuracy and ease of generating the required parameters by ANTECHAMBER[164].

The perturbations attempted in this chapter aim to calculate the free energies of binding for the brequinar and the non-brequinar binding modes, as well as to confirm the experimentally observed binding modes. These include:

- The perturbation of compounds 3 to 4, compounds 4 to 5, and compounds 6 to 7 within either binding mode using the single topology scheme.

- The perturbation of compounds 3, 4, 5, 6 and 7 from the brequinar to the non-brequinar binding mode using the dual topology scheme.

In the next sections we introduce the system setup, followed by the simulation protocol. We then conclude with the results obtained and close this thesis with the chapter 7, concluding on the achievements of this work.

6.1 System setup

The PDB structure 2FPV[48] of the Baumgartner series of human DHODH was selected a starting point for setup, as it offers the highest level of resolution available within the series.

Hydrogen atoms have not been resolved by the crystallographer and were added to the protein using WHAT-IF[91], Reduce[230] and Molprobity[231]. The inspection of the resulting structure, revealed marginal differences in the protonation states of the protein residues obtained. Thus, together with a visual inspection of all critical residues, and taking on board crystallographic evidence, i.e. the protonation state of HIS56 due to the presence of a stabilizing water molecule[48], a consensus was achieved.

The proteins were setup with the AMBER99 force field[69] , the inhibitors, the co-factor FMN and the natural substrate ORO were setup with the GAFF force field[136], and the atomic partial charges were derived using the AM1BCC method[49], as implemented in the software antechamber[164].

The AM1BCC atomic charges for symmetric sites on the inhibitors, the cofactor and the natural substrate were subsequently averaged to prevent artificially favourable rotameric states. Owing to its intramolecular hydrogen bond that may be present in all the inhibitors, and indeed was observed to be strong for all subsequent simulations, a formal charge of -1 was assigned to all inhibitor molecules, a formal charge of -2 was assigned for the FMN molecule and a formal charge of -1 was assigned to ORO, all of which were protonated using the PRODRG server[232].

In principle, two different sets of AM1BCC charges could be used for the inhibitors, i.e. one that has been derived using the energy minimised inhibitor in the brequinar binding mode, and another set that has been derived using the energy minimised inhibitor in the non-brequinar binding mode. In principle, the consistent

force field approach and the charges used for a molecule should be independent of the conformation of the inhibitor, and hence derived charges for different conformations should have little effect on the computed free energies, and indeed the charges themselves. Potentially, if two conformations have different charge sets and are able to interconvert, then the calculated free energies could be completely erroneous. The molecule could adopt the incorrect conformation for a particular charge set.

To make sure it is reasonable to use one set of charges for each inhibitor in both binding modes, we have run preliminary free energy simulations using RETI[40], where the charges of compound 4 were perturbed from the charges generated for the brequinar binding mode to the ones obtained for using the non-brequinar binding mode in the aqueous phase and in vacuum. These calculations have confirmed that both charge sets are nearly isoenergetic, as the resulting relative free energies were below 0.3 kcal mol^{-1}. Therefore we have decided to use the charges generated for the brequinar binding modes for all inhibitors.

To avoid steric clashes, the protein complexed to compound 6 in the brequinar binding mode was energy minimized using the Sander module of AMBER8[200] and a generalized Born force field (the igb keyword was set to 1). The backbone of the energy minimised protein was kept rigid for subsequent Monte Carlo simulations, which were conducted with our in-house software ProtoMS[47]. To reduce the computational cost, only the protein residues that have one heavy atom within 15 Å of any heavy atom of compound 6 in the brequinar binding mode were retained, resulting in a protein scoop consisting of 205 residues.

All crystallograhic water molecules have been removed from the minimised protein, which was subsequently hydrated by a sphere of TIP4P water molecules[195] of 22 Å radius and centred on the geometric centre of the inhibitor. Once solvated, the crystallographic water molecules within 22 Å of the inhibitor have been reinserted, while TIP4P waters within 2.5 Å of any of these crystallographic waters have been removed. However, each solvating sphere of TIP4P water molecules was subsequently adapted to match the desired hydration network defined in the previous chapter. To prevent evaporation of water molecules at the boundary of the sphere, a half-harmonic potential with a force constant of 1.5 kcal mol^{-1} $Å^{-1}$ was applied to water molecules whose oxygen atom distance was greater than 22 Å. A similar sphere of TIP4P waters was applied to solvate the inhibitors for the free state.

For the purpose of validating the cutoff sensitivity for non-bonded interactions, we have created further scoops by retaining residues within 10, 12, 20 and 25 Å of the center of geometry of the inhibitor. The perturbation of compound 6 into 7 in the non-brequinar binding mode was used as a model to study these effects. After 200M moves of general equilibration, statistics were collected for 70M moves, where the move ratios were identical to the ones presented further below for the single topology calculations. Non-bonded interactions were subject of a residue-based cutoff and the calculated relative free energies of binding were 3.9, 3.3, 6.8, 7.2 and 6.7 kcal mol^{-1} for cutoff values of 10, 12, 15, 20 and 25 Å respectively. Hence, it was decided to use a 15 Å cutoff in subsequent simulations, as calculated free energies did not show any significant dependency after 15 Å.

Biological affinities of the inhibitors, measured in consistent manner and given in with their $IC_{50}s$, were converted to binding free energies via the Cheng-Prusoff equation[233] and by assuming that the ratios of dissociation constants behave similar to the ratios of $IC_{50}s$.

6.2 Monte Carlo simulation protocol

Relative binding free energies for each inhibitor and each binding mode have been calculated using RETI[40]. The free energy gradients required have been calculated using a finite difference scheme where the free energies were obtained with the Zwanzig equation[11] at 16 values of the evenly spaced λ parameter, and the integral was estimated by trapezoidal numerical integration[6]. The finite difference, i.e. $\Delta\lambda$, was set to 0.001.

Perturbations that did not involve a change in binding mode were carried out using the single topology scheme described in chapter 2, while for a change in binding mode the dual topology paradigm was used. The single topology method linearly scales the force field parameters to match either end state of a simulation. For example, if we want to perturb ligand 3 into ligand 4 in the brequinar binding mode, then the force field would describe ligand 3 at $\lambda = 0.0$ and ligand 4 at $\lambda = 1.0$ and a transition matrix is used that allows the geometric conversion of ligand 3 into 4. In the dual topology scheme, both end states are defined separately, and λ scales the interaction with the environment. For example, if we want to perturb ligand 3

from the brequinar into the non-brequinar binding mode, then, at $\lambda = 0.0$, ligand 3 in the brequinar mode would be a fully interacting molecule, while ligand 3 in the non-brequinar mode would not experience any interactions with the environment, i.e. the ligand is in the gas phase.

To avoid the Lennard-Jones endpoint singularity, the dual topology approach requires the soft-scaling of intermolecular interactions[78]. The perturbation of ligand 4 from the brequinar into the non-brequinar binding mode has dealt as a model system to define appropriate scaling factors. In principle, the calculated free energies should not be affected by the choice of these parameters. Indeed, the calculated free energies were rather similar for a range of softcore parameters, but at the same time we found a serious dependency of these parameters on the particular system simulated. For example, a value of 1.5 for the term α in equation 2.57 was found appropriate for all compounds of the Baumgartner series, but not for compounds 6 and 7. Here we have used a scaling factor of 1.75 for the Lennard-Jones interactions and no scaling was applied for the electrostatic interactions.

In all simulations, including the simulations to define an appropriate set of softcore parameters, the bond angles and torsions for the side chains of residues within 15 Å of any heavy atom of the inhibitor and all the bond angles and torsions were sampled during the simulation, with the exception of rings. The bond lengths of the protein and ligand were constrained and all simulations were carried out at a temperature of 298 K.

Solvent moves were attempted with a probability of 85 %, protein sidechain moves with a probability of 9 % and solute move with a probability of 6 %. In the unbound state, solvent moves were attempted 95 % of the time with the remainder being solute moves. Replica exchange moves were attempted every 200K moves. The solvent was equilibrated for 200M configurations to remove any repulsive contact with the solute(s). The system was then equilibrated in one end state corresponding to that of compound 6 in the brequinar binding mode for 50M further moves where solute, protein, and solvent moves were attempted. The resulting configuration was distributed over 16 evenly spaced values of the coupling parameter λ.

For simulations using the single topology scheme, the pre-equilibrated system was equilibrated again for 30M moves (to adapt the system to the different compounds and binding modes in subsequent simulations) before statistics were collected for 16

evenly spaced values of the coupling parameter λ. The amount of data collection using the single topology approach varied, and was subject to obtaining stable averages of intermolecular energies as well as the total energy of the system. However, for all results presented here data was collected for at least 200M moves for each value of λ.

For simulations using the dual topology scheme, the system was equilibrated for 80M moves before statistics were collected. As for the single topology calculations, energy averages were observed during the simulations and the results reported here for the dual topology calculations attempted at least 300M moves.

6.3 Results

The experimental relative binding free energies, converted from the experimental IC_{50} data using the Cheng-Prusoff relationship[233], for the perturbation of compound 3 into compound 4 and compound 4 into 5, with experimental $IC_{50}s$ of 280 to 33 nM and 33 to 7 nM , stand at -1.3 and -0.8 kcal mol^{-1} respectively. The structures of the compounds are shown in figure 6.1. In this figure the moiety that is rotated by 180° is shown in green for compound 3.

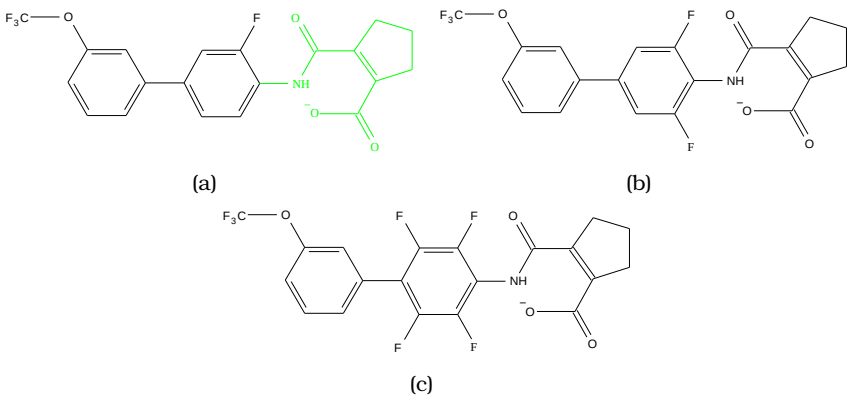


Figure 6.1: Compounds 3 (a), 4 (b) and 5 (c) of the DHODH series developed by Baumgartner et al.[48].

If we imagine that compounds 3 and 4 only exist in a brequinar binding mode, then the relative free energy of binding of compounds 3 and 4 can be calculated using the standard thermodynamic cycle, typical for single topology calculations. The same is true for the perturbation of compound 3 into 4 in the non-brequinar binding mode.

Using RETI MC simulations we calculate the free energy of perturbing compound 3 into 4 in the water phase, and the same perturbation in the protein binding site. The relative free energy of binding for ligand 3 to 4 is then obtained as the difference (protein minus water).

If we want to use free energy simulations to assess whether compound 3 is likely to bind in a certain binding mode, or indeed adopts both binding modes we use RETI MC simulations to calculate the free energy of perturbing of compound 3 in the brequinar binding mode into the non-brequinar binding mode in the water phase, and complexed to the protein. The relative free energy of binding for compound 3 between each binding mode is then obtained as their difference.

The prediction of binding modes is routinely performed by docking programs, and docked poses are usually scored using an appropriate scoring function. However, detailed experimental and theoretical studies on molecules that adopt different binding modes, or molecules able to adapt dual binding modes in one crystal structure, are rare[236]. Furthermore, attempts to reproduce the observed binding poses for the Baumgartner series using the docking program ProPose[237] failed[48]. Therefore it is too early to assess the accuracy of binding mode calculations. Still, first findings of rigorous free energy protocols suggest that binding modes may be assigned when the free energy difference is greater than 1 kcal mol^{-1} [46], while values below might render both binding modes possible.

These individual perturbations may now be combined to allow the calculation of the relative free energy of each binding mode as well as the relative binding free energies for compounds 3 and 4.

No change in hydration states of both, compounds 3 and 4 in either binding mode was observed in chapter 5, and thus a single topology approach may be applied for pertrbing 3 into 4. The solvating sphere of TIP4P water molecules[195], used to represent the solvent sphere, contained all crystallographic waters, including waters A, B, C and D that were investigated previously.

Preliminary results for the binding mode perturbations using different scaling parameters to soften the intermolecular Coulombic and Lennard-Jones interactions, indicated a considerable degree of noise in these calculations, and no set of parameters could be identified to lower the error on the computed free energies. For example, using no scaling for the Coulombic interactions and a scaling factor of 1.5 for

the Lennard-Jones interactions, a set of softcore parameters showing the smoothest gradients for the perturbation of compound 4 from brequinar into the non-brequinar binding mode, yielded a standard deviation of 3.14 kcal mol^{-1} and a standard error of 1.57 kcal mol^{-1} over 3 independent simulations. The average relative free energy can therefore not be determined with certainty, as the error alone would be enough to change the prediction of a possible binding mode. Figure 6.2 shows the recorded gradients for these simulations with respect to λ.

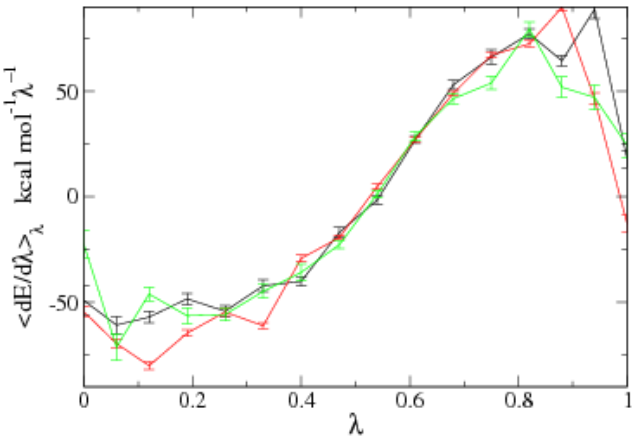


Figure 6.2: Free energy gradients collected during the perturbation of compound 4 from the brequinar to the non-brequinar binding mode in the bound state for three independent runs shown together with the error bars. The x-axis shows the λ coordinate while the y-axis shows the gradients.

The simultaneous decoupling of the intermolecular Coulombic and Lennard-Jones interactions could be a potential reason for the large error associated for these simulations. Additionally, the formal charge on the inhibitor may cause a larger error, but examples of free energy studies involving charged species in protein binding sites are not common[238]. To verify this, and eventually alleviate this problem, a more complex thermodynamic cycle was devised. The devised scheme is to split the dual topology perturbation into a step-wise approach:

1. Using single topology, we uncharge the molecule in the brequinar binding mode for both legs of the perturbation for the part of the inhibitor that is involved in the change of binding mode, i.e. the part of compound 3, for example, that appears green in figure 6.1, i.e. step 1. Since this effectively involves turning off a formal charge for all the inhibitors in the Baumgartner series it will be

energetically highly disfavoured for both the free and the bound state of the perturbation.

2. Using dual topology, we calculate the relative free energy for switching the binding mode using the now partly uncharged molecule for both the free as well as the bound leg, i.e. step 2. Since the molecule, formally, does not carry a charge, the noise in the dual topology legs could be lower.

3. Again using single topology, we recharge the molecule in the non-brequinar binding mode, i.e. step 3. This step will be highly favourable for both the free and the bound leg.

Because free energy is a state function, this splitting performed in the bound and the free state will again yield the relative free energy for the binding modes perturbed. For these simulations we apply the same amount of moves as described in the protocol section of this chapter. The results for these perturbations for compounds 3, 4, 5, 6 and 7 are given in table 6.1.

The steps that involve the charging or uncharging of parts of the compounds yield very large free enery changes. This is primarily down to the magnitude of the intermolecular interactions of the inhibitor with the protein and solvent, while the intramolecular Coulombic terms are high too. However, the results obtained are still precise, and much less noise is apparent compared to previous runs, that attempted to simultaneously turn off both, Lennard-Jones and Coulombic terms. Ideally, the intermolecular Coulombic terms could be turned off separately from the intramolecular Coulombic terms, and thus providing us with smaller values for the charging and un- or recharging steps. This could potentially results in more accurate numbers. However, this feature is not present in the current version of ProtoMS[47] and was not implemented as part of this project.

Most importantly, the errors on the gradients using the dual topology approach are reasonable, and so are the errors for the charging and uncharging steps. The square root of the sum of the squared mean unsigned errors over all 3 steps making up the binding mode perturbations, stand at 0.3 and 0.5 for compounds 3 and 4. Thus, providing the runs yield well converged averages, then the protocol provides consistent answers. Figure 6.3 shows the gradients recorded for the binding mode switch of the partly uncharged molecule. Although the general trend of the gradients

Cmpd.	Step1 (a)	Step1 (b)	Step2 (a)	Step2 (b)	Step3 (a)	Step3 (b)
3	+183.0	109.4	-0.5	-0.5	-185.0	-108.5
	+182.6	109.4	-1.4	-0.5	-184.9	-109.1
	+182.2	109.3	-1.0	-0.5	-185.3	-109.6
Difference	+73.2 ± 0.25		-0.5 ± 0.60		-75.9 ± 0.64	
4	+177.2	+99.8	+0.4	+0.1	-177.8	-98.7
	+176.4	+99.7	+0.6	+0.1	-178.0	-98.7
	+176.8	+99.7	+0.5	+0.1	-177.9	-98.7
Difference	+77.0 ± 0.28		0.4 ± 0.10		-79.2 ± 0.43	
5	+178.5	+103.8	-0.8	-0.6	-179.4	-103.9
	+177.6	+103.6	-0.9	-0.9	-179.5	-103.5
	+178.1	+104.0	-0.8	-0.8	-180.0	-104.3
Difference	+74.3 ± 0.38		0.00 ± 0.21		-75.7 ± 0.57	
6	+196.9	+104.6	+2.7	+0.8	-179.8	-102.3
	+197.2	+104.9	+3.5	+0.9	-179.5	-102.7
	+197.0	+104.2	+2.9	+0.8	-180.3	-101.8
Difference	+92.4 ± 0.29		+2.30 ± 0.20		-77.6 ± 0.57	
7	+169.6	+94.6	+6.5	+1.0	-158.4	-91.6
	+170.2	+94.0	+7.1	+0.9	-157.2	-91.7
	+171.0		+7.9		-157.7	
Difference	+76.0 ± 0.5		+6.2 ± 0.41		-66.3 ± 0.57	

Table 6.1: Calculated free energies for binding mode perturbations going from brequinar to non-brequinar. Steps 1, 2 and 3 correspond to uncharging in the brequinar binding mode using single topology, perturbing the binding mode of the partly uncharged species from brequinar to non-brequinar using dual topology and finally recharging the compounds in the non-brequinar binding mode using single topology respectively, for both the bound legs, i.e. indicated with (a) in table, and the free leg, indicated with (b). All figures are given in kcal mol^{-1} and the error estimate given is the standard error of the mean for the three independent runs.

has not changed, the accompanying errors on the gradients are considerably lower and stand at 0.9 kcal mol^{-1}.

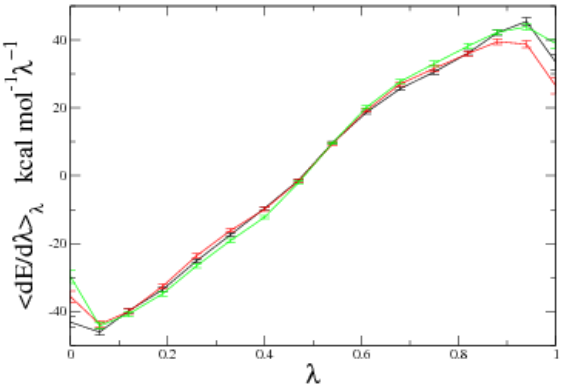


Figure 6.3: Free energy gradients collected during the perturbation of compound 4 from the brequinar to the non-brequinar binding mode in the bound state applying the devised and step-wise approach for three independent runs shown together with the error bars. The x-axis shows the λ coordinate while the y-axis shows the gradients.

Thermodynamic cycles for the perturbation of ligand 3 to 4 are shown in figure 6.4. Overall, the cycles close within an error of 0.9 kcal mol^{-1} for the perturbation of compound 3 into 4, including the free energy between the different binding modes. According to these figures compound 3 prefers the non-brequinar binding mode by 3.2 kcal mol^{-1}, which is in agreement with the experimental findings of Baumgartner et al.[48]. Compound 4, which could not be clearly resolved by experiment, may indeed adopt both binding modes, as the calculated free energy between the binding modes stands at -1.8 kcal mol^{-1}. This somewhat negative free energy change would, by itself, speak for the non-brequinar binding mode, but upon inclusion of hysteresis effects of 0.9 kcal mol^{-1} we have to consider both modes of binding for compound 4. Since free energy is a state function, we can connect any end states, and as such the thermodynamic cycle in figure 6.4(a) may be distilled into a thermodynamic cycle for the free and the bound leg. This separation, shown in 6.4(b) and 6.4(c) for the bound and free leg respectively, is useful, as failure to close the cycles indicates where hysteresis effects are dominant. While the bound legs close within 0.5 kcal mol^{-1}, the error in the free legs is somewhat higher with 1.0 kcal mol^{-1}. This result is surprising, as the free legs are usually easier to converge compared to the bound states. Reconciling that all the inhibitors of the Baumgartner series show an intramolecular

hydrogen bond between the carboxylate and the amide hydrogen, gives further room for speculation, as sampling of the free legs may be artificially restrained by the force field to conserve this non-bonded interaction. Indeed, the acceptance rates upon inclusion of the sampling of the central dihedral that would allow the two conformations to interconvert, drop sharply below 5 %, while they are reasonable with 25 to 35 % when this dihedral is not sampled in the simulations. Assuming that ligand reorganisation effects are significant for the Baumgartner series, would lead one to believe that although the inhibitors are strained upon binding to the protein, they may sample their central torsion considerably more than the simulations attempted here have achieved. Therefore, if this assumption is true, then it may be correct to refer to the free energies associated with the free legs as confined free energies, and appropriate sampling of these could potentially change the predictions.

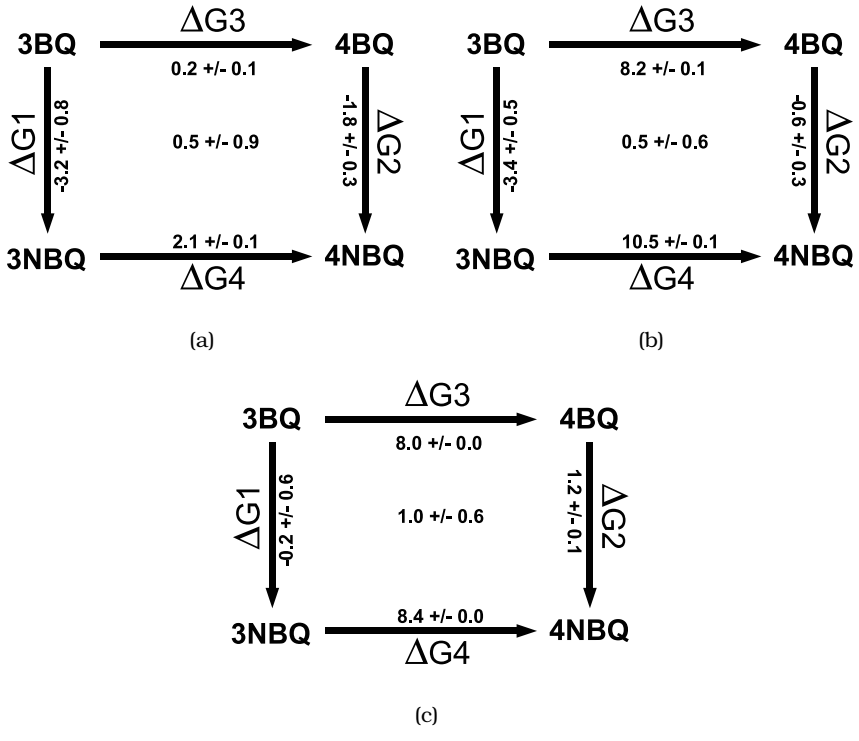


Figure 6.4: Thermodynamic cycle for the perturbation of compound 3 to 4. The binding free energy cycles is shown in (a), and the distilled cycles for the bound state in (b) and the free state in (c). All figures are given in kcal mol^{-1} and the centered figures indicate the cycle closure. The error estimate on each leg of the cycle represents the root square of the sum of the squared mean unsigned errors.

The experimental free energy of binding between compounds 3 and 4 lies at -1.3

kcal mol^{-1} and the current protocol using the thermodynamic cycle 6.4(a) allows for different perturbation pathways to be evaluated to estimate the free energy of binding. Moreover, to account for the dual binding mode behaviour of compound 4 for example, the relative free energies of binding for each mode need to be combined to produce the overall free energy of binding via [101]

$$\Delta\Delta G = -RT\ ln\Big[exp\Big(-\Delta\Delta G_{Mode1}/RT\Big)\ +\ exp\Big(-\Delta\Delta G_{Mode2}/RT\Big)\Big]\ +\ RT\ ln2 \quad (6.1)$$

where R is the ideal gas constant, T is 298 K, and the $\Delta\Delta G$'s are the relative free energies of binding for the two binding modes, i.e. Mode1 and Mode2 in the equation. The second term in equation 6.1 penalises the computed free energies of binding of the unsymmetrical ligands by RT ln 2, because they are relative to the binding modes. Thus, when the relative free energies of binding between 2 binding modes differs by more than 2 kcal mol^{-1}, then the free energy is essentially that of the more favourable binding mode plus RT ln 2. Alternatively, if the relative free energies of binding of the two binding modes are the same, then the RT ln 2 penalty is removed.

Hence, the calculation of the free energies of binding for the perturbation of compound 3 into 4, yields free energies of 2.1 and 3.4 kcal mol^{-1} when using the pathways consisting of the perturbation of 3 into 4 in the non-brequinar binding mode and 3 into 4 from the non-brequinar binding mode into the brequinar binding mode. Here the difference in free energies is less than 2 kcal mol^{-1} and hence the penalty of RT ln 2 may be omitted. The free energy change calculated using this approach then stands at 2.03 kcal mol^{-1}.

Finally, the calculated relative free energies are not accurate and thus do not agree well with experiment, while they are still precise as not only the thermodynamic cycle in 6.4(a) closes within 0.5 but also the free and the bound state close within 1.0 and 0.5 kcal mol^{-1} respectively. If we were to assume that in this thermodynamic cycle perturbation approach we benefit from a fortuitous cancellation of errors, then the accuracy issues seen here may be attributed, in part, to the force field used. If the calculation of the relative free energy of hydration of fluorobenzene is reflective for this error, then it may be at least in the range of approximately 1.0 kcal mol^{-1}.

Thermodynamic cycles for the perturbation of ligand 4 to 5 are shown in figure 6.5. Overall, the cycle closes within an error of 1.5 kcal mol^{-1} for the perturbation of compound 4 into 5, including the free energy between the different binding

modes. Similar to compound 4, compound 5 may adapt a dual binding mode, which is in agreement with the experimental findings of Baumgartner et al.[48], as the calculated free energy between the binding modes for compound 5 stands at -1.4 kcal mol^{-1}. This somewhat negative free energy change would, by itself, speak for the non-brequinar binding mode, but upon inclusion of hysteresis effects of 1.5 kcal mol^{-1} we have to consider both modes of binding for compound 4. Distilling the thermodynamic cycle in figure 6.5(a) into a thermodynamic cycle for the free and the bound leg again show that hysteresis effects for the bound state may be smaller than for the free state. These thermodynamic cycles are shown in 6.5(b) and 6.5(c) for the bound and free leg respectively. While the bound leg closes within 0.1 kcal mol^{-1}, the error in the free legs is somewhat higher with 1.6 kcal mol^{-1}. This result is surprising, as the free legs are usually easier to converge compared to the bound states and reconciling that all the inhibitors of the Baumgartner series show an intramolecular hydrogen bond between the carbohydrate and the amide hydrogen, gives further room for speculation, as sampling of the free legs may be artificially restrained by the force field to conserve this non-bonded interaction. Hence similar conclusions as for the perturbation of compound 3 into 4 apply.

The perturbation of compound 4 into 5 requires, apart for the perturbations presented so far, the annihilation of a water molecule. Since waters A, B, C and D are required to match the predicted end state for compound 4, the simulations were carried out used a scoop containing waters A, B, C and D, while water C was subsequently annihilated using the double decoupling approach. This has been described in chapter 5, and the calculated free energy of annihilation stands at -2.31 kcal mol^{-1}. Therefore, if we would incorporate the annihilation of a water molecule in cycle 6.5(a) for the single topology approach within the brequinar binding mode, then $\Delta G3$ standing at -4.1 kcal mol^{-1} would instead rise to -6.4. However, to fully integrate the annihilation of water molecules, additional simulations would be required.

Calculating the free energies of binding for the perturbation of compound 4 into 5, yields free energies of -4.1 and -5.2 kcal mol^{-1} for the perturbation carried out in the brequinar and non-brequinar binding mode respectively. As for the previous perturbation, the difference in free energies is less than 2 kcal mol^{-1} and hence the penalty of RT ln 2 may be omitted. The calculated free energy change then stands at -5.20 kcal mol^{-1}.

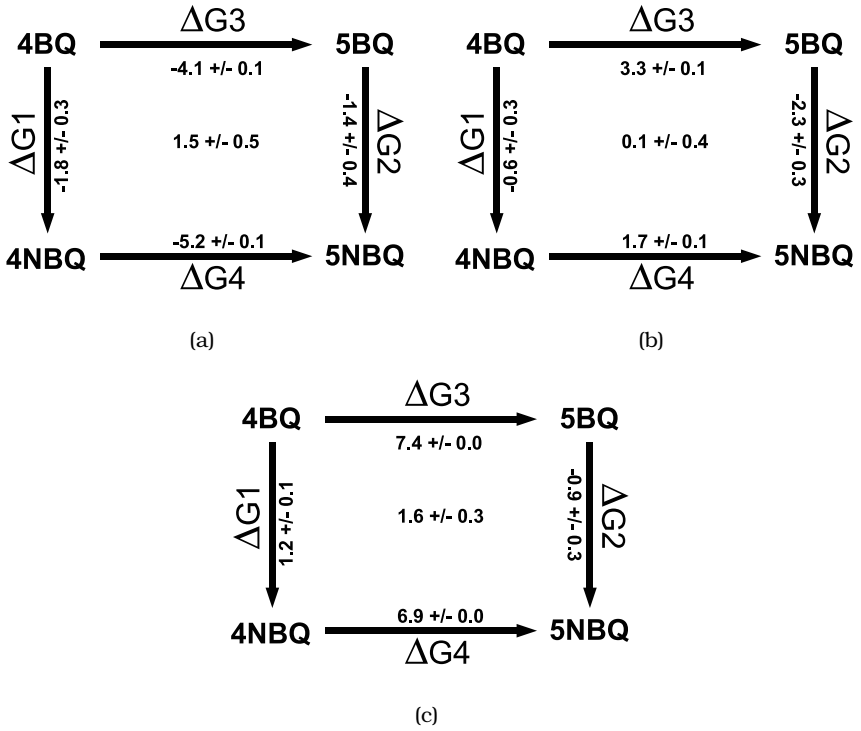
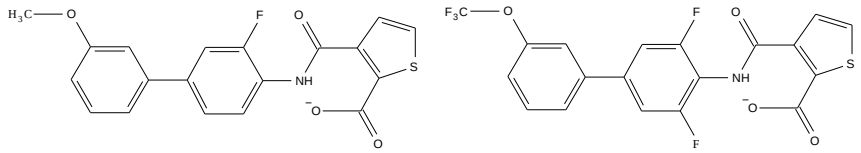


Figure 6.5: Thermodynamic cycle for the perturbation of compound 4 to 5. (a) shows the relative binding free energy cycle, and the distilled cycles for the bound state are shown in (b) and for the free state in (c). All figures are given in kcal mol^{-1} and the centered figures give the cycle closure. The error estimate on each leg of the cycle represents the root square of the sum of the squared mean unsigned errors.

The calculated relative free energies are not accurate and thus do not agree well with experiment, while they are still precise as the thermodynamic cycle in 6.5(a) closes within 1.5, and the distilled cycles for the bound and free state close within 0.1 and 1.6 kcal mol^{-1} respectively. As previously, the potential cancellation of errors in this thermodynamic cycle perturbation approach could reveal the systematic errors in the force field. In this case the accuracy issues seen here may reflective for growing fluorine atoms, and if the error in the calculation of the relative hydration free energy of fluorobenzene is reflective for this error, then it may be in the range of 4.0 kcal mol^{-1}.

The experimental relative binding free energy, converted from the experimental IC_{50} data using the Cheng-Prusoff relationship[233], for the perturbation of compound 6 into compound 7, with experimental $IC_{50}s$ of 44 to 2 nM respectively, stands at

-1.8 kcal mol^{-1}. The structures of the compounds are shown in figure 6.6.



(a) (b)

Figure 6.6: Compounds 6 (a) and 7 (b) of the DHODH series developed by Baumgartner et al.[48].

A significant change in hydration network of both, compounds 6 and 7 in the brequinar binding mode was observed in chapter 5. For both compounds in the brequinar binding mode, the water triad is reduced to two water molecules, i.e. water B is energetically unfavourable, while an additional water molecule has been identified by GCMC and JAWS analysis. This additional water, i.e. water A2, is located close to the carboxylate group of the inhibitor and becomes unfavourable in the non-brequinar binding mode. The non-brequinar binding mode is thus identical to compounds 3 and 4, i.e. waters A, B, C and D are present. The solvating sphere of TIP4P water molecules[195], used to represent the solvent sphere, contained all crystallographic waters, including waters A, B, C and D that were investigated previously.

Initial results for the binding mode perturbations, using the same scaling parameters to soften the intermolecular Coulombic and Lennard-Jones interactions as for the previous perturbations, led to sampling artefacts, and energies could not be evaluated for these simulations. It appears, that even though there is a high degree of similarity between the compounds in the Baumgartner series, the previous scaling parameters were inadequate, as conformational changes observed for the protein and the inhibitor resulted in the inhibitor finally leaving its proposed binding pocket. Therefore, as explained in the simulation protocol section of this chapter, a new set of softcore parameters has been adapted for this perturbation, yielding stable free energy gradients.

Thermodynamic cycles for the perturbation of ligand 6 to 7 are shown in figure 6.7. Overall, the cycle closes within an error of 1.6 kcal mol^{-1} for the perturbation of compound 6 into 7, including the free energy between the different binding modes. Providing these calculations are well converged, then both compounds have a clear

preference for the brequinar binding mode, while the non-brequinar binding mode seems highly unfavourable. This qualitative result is in agreement with the findings of Baumgartner et al. for compound 7, while for compound 6 the electron densities did not allow a unique assignment of a binding mode. According to the thermodynamic cycle 6.7(a), compounds 6 and 7 prefer the brequinar binding mode over the non-brequinar binding mode by 17.1 and 15.9 kcal mol^{-1} respectively. However, these estimates do not incorporate the free energies associated with creating and annihilating waters A2 and B respectively. Including these energetic contributions, as estimated by JAWS stage 2, would lower the free energy between the brequinar and the non-brequinar binding mode for compounds 6 and 7 respectively. However, these need to be confirmed by a double decoupling procedure before they can be integrated into the binding free energy cycle. Providing JAWS is accurate, which we could demonstrate for water B in compound 5 in the brequinar binding mode, then the free energy of annihilation for waters B will lie in the region of -5.0 to -6.0 kcal mol^{-1}.

Finally, the calculated relative free energy for the perturbation of compound 6 into 7 stands at 9.6 kcal mol^{-1}, providing the inhibitors adopt the brequinar orientation only. This is not in agreement with experiment, but the thermodynamic cycle in 6.7(a) closes within 1.6, as do the free and the bound states close within 0.7 and 0.5 kcal mol^{-1} respectively, indicating precise calculations.

6.4 Conclusions

The prediction of the binding free energies and the prediction of binding modes using rigorous free energy methods was the subject of this chapter. This is illustrated on the Baumgartner series of inhibitors complexed to DHODH. While the structural modifications on the inhibitors appear minor, they seem to have big effects on the binding mode behaviour, i.e. the introduction of fluorine atoms leads to dramatically altered binding modes.

Initial attempts, relying on the free energy simulation protocols published previously from our group [18,46], failed to predict the binding modes or the free energies of binding in DHODH. We found that the equilibration times in the system appeared

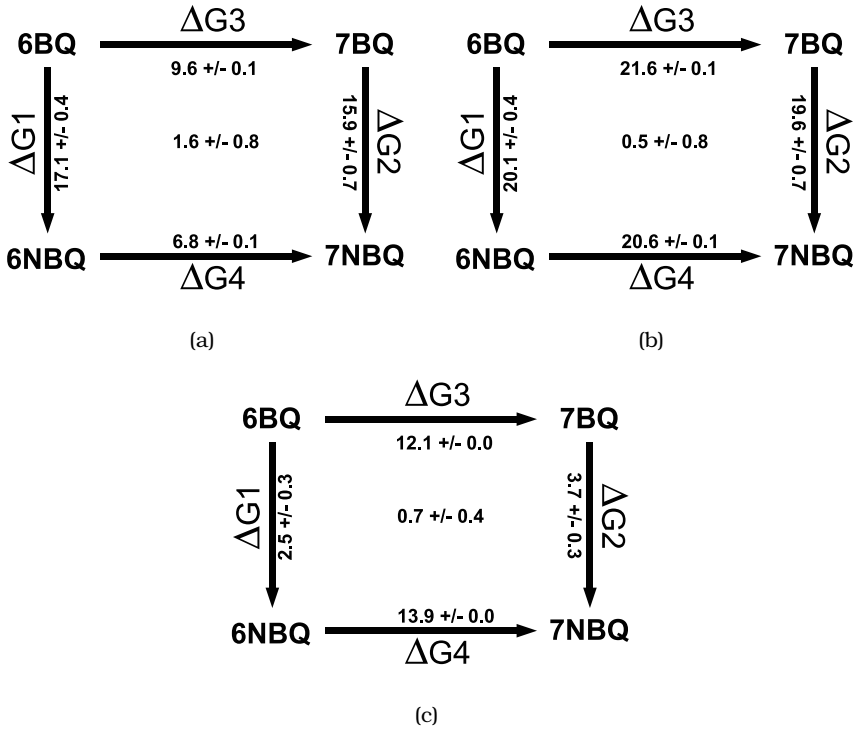


Figure 6.7: Thermodynamic cycle for the perturbation of compound 6 to 7. (a) shows the thermodynamic cycle for the relative binding free energies, and the distilled cycles for the bound state are shown in (b) and the free state in (c). All figures are given in kcal mol^{-1} and the centred figures represent cycle closures. The error estimate on each leg of the cycle represents the root square of the sum of the squared mean unsigned errors.

extraordinarily long, and, as we found later, were mainly due to a constant reorganisation of water molecules. This effect appeared to be pronounced with the size of the protein scoop, and smaller protein scoops result in longer equilibration times. Small scoops in DHODH have the additional consequence that the backbone of the protein becomes increasingly fragmented, and this fragmentation may further increases the time required to equilibrate water molecules appropriately. While this effect may have implications for the dynamics of the system, it did not seem to affect the calculated free energies. Consequently, the testing of different cutoff ranges and scoop sizes led us to define a 15 Å scoop centred on the ligand's centre of geometry.

The sampling of the inhibitors is very difficult, and may be mainly attributed to the strong intramolecular non-bonded interactions. Initial attempts to alleviate this problem were aimed at developing a replica-exchange protocol, that on top of our

RETI protocol introduced exchanges with simulation windows of higher temperature at the physical end states, thus enhancing sampling. However, since the spacing between the temperature windows needs to be small enough to achieve reasonable acceptance rates, this idea had to be buried, as the temperatures required were to high to run these simulations using our computing facilities, i.e. there were too many replicas.

Having defined the more general simulation conditions, i.e. cutoff, scoop sizes, ligand sampling issues, we pursued methods to define rigorously the hydration networks present in DHODH. This was described in chapter 5, and included the validation of the current implementation of JAWS and GCMC in ProtoMS. It is important to stress that these validations were necessary and they have guaranteed that the results presented here are in fact independent of the finely tuned simulation parameters in JAWS and GCMC. For example, different grid spacings and thresholds for θ in JAWS have been investigated in detail, and we are confident that these parameters will not qualitatively change the predicted hydration networks.

Additional problems, mainly established by very noisy gradients, were encountered when we tried to use the dual topology method, as implemented in ProtoMS[32,47], for the change in binding modes. A devised scheme was therefore employed, where the charging/uncharging events and the switch in binding mode were treated separately. The recorded gradients were, as expected, much more stable, and the general trend of the gradients did not change. All these steps finally resulted in a protocol that, judging by the precision of the thermodynamic cycles obtained, allows predictions for the Baumgartner series complexed to DHODH.

Therefore, we have successfully elaborated a complex and detailed simulations protocol, that, in principle, allows the prediction of binding modes. The protocol yields precise free energy estimates, but these are not accurate. However, to further validate this protocol, we are in the process of running more and even longer simulations to possibly lower the error, in particular for the free states. The inhibitors presented here were chosen due to the availability of their crystal structures. To extend the Baumgartner series with compounds of larger range in biological affinities, we are also in the process of applying this protocol to 30 inhibitors that are all based on the inhibitors presented here, for which no crystal structures exist.

7

Conclusions

Structure-based drug discovery models seek to predict receptor-ligand binding free energies from the known or presumed structure of the corresponding complex[6]. The class of docking methods and empirical scoring approaches[58], which are useful in virtual screening applications[59], are now routinely employed in drug-discovery programmes. Although often considered rather inaccurate, docking methods satisfy by generating a first rough dataset that in turn will be evaluated, refined and eventually bio-assayed.

Free energy simulations provide, apart from their potentially much more reliable binding estimate, very detailed information on system dynamics, as properties are being averaged while the system is sampled. This is usually not the case for docking methods, as they often rely on the use of a single conformation.

This research set out with the idea of validating and developing free energy simulation protocols, and to demonstrate their potential usefulness in SBDD. In chapter 3 we define the protein DHODH as a challenging model system for SBDD[48]. This system is challenging for several reasons: despite the high degree of similarity between the inhibitors in the series, they seem to show remarkably altered binding modes upon addition of fluorine atoms. However, at the same time it is not clear whether these altered binding modes are real, as the crystallographic evidence did not allow a more detailed refinement; the refinement in these structures is also complicated through the existence of hydration networks in the protein binding site. This limit in refining the protein-ligand complexes makes it an ideal test case to investigate the performance of rigorous free energy simulations. To develop a novel and reliable

protocol to calculate the relative binding free energies and help in further refining the proposed binding modes, we have performed a detailed and stepwise approach.

To identify a suitable set of force field parameters, we have calculated the relative free energies of hydration for a small, but representative, set of small molecules for DHODH in chapter 4. As reported by other groups, all the force fields employed in this study did perform reasonably well, with GAFF and OPLS-AA performing best. Apart from the classical MM representation of the system, we have also validated a novel QMMM method for the prediction of the relative free energies of hydration and investigated the water structures upon solvating the molecules by using the RDFs obtained during the simulations. Although the QMMM approach performed best for the perturbation of fluorine atoms, it also came with several technical challenges that had to be overcome. Consequently, we have decided to use GAFF in combination with AM1BCC charges for the simulation of DHODH.

In chapter 5 we have employed the JAWS algorithm and GCMC simulations to refine the hydration networks present in DHODH. The results not only suggested a clear change in hydration networks for the different binding modes of the inhibitors, but also within one binding mode, while JAWS and GCMC qualitatively predict the same hydration networks.

In chapter 6 we collect all the results obtained and develop a free energy simulation protocol for the study of DHODH. The thermodynamic cycles obtained close within less than 2 kcal mol^{-1}, and thus demonstrate that the current protocol gives precise estimates. Interestingly, the free legs of the perturbation show higher errors than the bound states and ideally, further simulations should be conducted to lower this error. Despite achieving precision in these simulations and hence allowing us to further refine the binding modes for the Baumgartner series, our predictions are not accurate. Providing that we benefit from the error cancellation when calculating the double free energy differences, this issue may be addressed, in part, to the force field we have used. Therefore it may be appropriate to revisit the simulations using a QMMM approach, and to extend our series with more compounds of wider dynamic range in biological affinities. The combination of using a QMMM method, extending the runtime and the amount of simulations, and including more compounds of wider dynamic range in biological affinities, could then clearly demonstrate that the proto-

col developed not only allows precise but accurate predictions. This will be the aim for our future work.

References

[1] Encyclopedia Britannica. Pharmaceutical industry, 2011. [Online; accessed 10-March-2011]. 1, 2

[2] J. Michel, N. Foloppe, and J. W. Essex. Rigorous free energy calculations in structure-based drug design. *Mol. Inf.*, 29(8-9):570–578, 2010. 1

[3] C. Lim, K. D Jordan, W. Thiel, P. Hobza, and K. Muller-Dethlefs. *Non-covalent interactions: theory and experiment*. RSC Theoretical and Computational Chemistry Series. The Royal Society of Chemistry, 2009. 1

[4] E. C. Hulme and M. A. Trevethick. Ligand binding assays at equilibrium: validation and interpretation. *Brit. J. Pharmacol.*, 161(16):1219–1237, 2010. 1

[5] M. N. Khan and J. Q. A. Findlay, editors. *Ligand-binding assays: development, validation, and implementation in the drug development arena*. Wiley, 2009. 1

[6] C. Chipot and A. Pohorille, editors. *Free energy calculations: theory and applications in chemistry and biology*. Springer Series in Chemical Physics. Springer, 2007. 3, 4, 8, 9, 12, 17, 19, 22, 24, 27, 28, 29, 37, 103, 107, 108, 112, 144, 160

[7] D. Frenkel and B. Smit. *Understanding molecular simulation, second edition: From algorithms to applications (Computational Science Series, Vol 1)*. Academic Press, 2 edition, November 2001. 3, 12, 14, 15, 16, 17, 18, 21

[8] A. E. Mark, W. F. van Gunsteren, and H. J. C. Berendsen. Calculation of relative free energy via indirect pathways. *J. Chem. Phys.*, 94:3808–3816, 1991. 4

[9] B. L. Tembre and J. A. Mc Cammon. Ligand-receptor interactions. *Comput. Chem.*, 8(4):281 – 283, 1984. 4

[10] J. G. Kirkwood. Statistical mechanics of fluid mixtures. *J. Chem. Phys.*, 3:300–313, 1935. 4, 26

[11] R. W. Zwanzig. High-temperature equation of state by a perturbation method. i. nonpolar gases. *J. Chem. Phys.*, 22(8):1420–1427, 1954. 4, 21, 102, 144

[12] W. L. Jorgensen and C. Ravimohan. Monte carlo simulation of differences in free energies of hydration. *J. Chem. Phys.*, 83(6):3050–3054, 1985. 4

[13] C. F. Wong and J. A. McCammon. Dynamics and design of enzymes and inhibitors. *J. Am. Chem. Soc.*, 108(13):3830–3832, 1986. 4

[14] J. Aqvist, C. Medina, and J. E. Samuelsson. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.*, 7(3):385–391, 1994. 5

[15] J. Swanson, R. H. Henchman, and J. A. McCammon. Revisiting free energy calculations: a theoretical connection to MMPBSA and direct calculation of the association free energy. *Biophys. J.*, 86(1):67–74, 2004. 5

[16] J. Michel and J. W. Essex. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aid. Mol. Des.*, 24(8):639–658, 2010. 5

[17] Ajay N. Jain. Scoring functions for protein-ligand docking. *Curr. Prot. Pept. Sci.*, 7(5):407–420, 2006. 5

[18] J. Michel, M. L. Verdonk, and J. W. Essex. Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization? *J. Med. Chem.*, 49(25):7427–7439, 2006. 5, 6, 7, 43, 101, 157

[19] Y. Deng and B. Roux. Calculation of standard binding free energies: aromatic molecules in the t4l lysozyme l99a mutant. *J. Chem. Theory Comput.*, 2(5):1255–1273, 2006. 5, 36, 61

[20] C. R. W. Guimaraes, D. L. Boger, and W. L. Jorgensen. Elucidation of fatty acid amide hydrolase inhibition by potent alpha-ketoheterocycle derivatives from monte carlo simulations. *J. Am. Chem. Soc.*, 127(49):17377–17384, 2005. 5

[21] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. McReynolds, B. K. Shoichet, and K. A. Dill. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, 371(4):1118–1134, 2007. 5

[22] G. Lamoreux and B. Roux. Absolute hydration free energy scale for alkali and halide ions established from simulations with a polarizable force field. *J. Phys. Chem. B*, 110(7):3308–3322, 2006. 6

[23] D. L. Mobley, C. I. Bayly, M. D. Cooper, and K. A. Dill. Predictions of hydration free energies from all-atom molecular dynamics simulations. *J. Phys. Chem. B*, 113(14):4533–4537, 2009. 6

[24] D. L. Mobley, E. Dumont, J. D. Chodera, and K. A. Dill. Comparison of charge models for fixed-charge force fields: small molecule hydration free energies in explicit solvent. *J. Phys. Chem. B*, 111(9):2242–2254, 2007. 6

[25] P. V. Klimovich and D. L. Mobley. Predicting hydration free energies using all-atom molecular dynamics simulations and multiple starting conformations. *J. Comput. Aid. Mol. Des.*, 24(4):307–316, 2010. 6, 42, 66

[26] D. L. Mobley, A. E., C. J. Fennell, and K. A. Dill. Charge asymmetries in the hydration of polar solutes. *J. Phys. Chem. B*, 112(8):2405–2414, 2008. 6

[27] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, and V. S. Pande. Predicting small-molecule solvation free energies: an informal blind test for computational chemistry. *J. Med. Chem.*, 51(4):769–779, 2008. 6, 80

[28] A. Plugatyr and I. M. Svishchev. The hydration of aniline: analysis of spatial distribution functions. *J. Chem. Phys.*, 130(11):114509–114520, 2009. 6, 97

[29] D. L. Mobley and K. A. Dill. Treating entropy and conformational change in implicit solvent simulations of small molecules. *J. Phys. Chem. B*, 112(3):938–946, 2008. 6, 84, 88

[30] F. R. Beierlein, J. Michel, and J. W. Essex. A simple qm/mm approach for capturing polarization effects in protein-ligand binding free energy calculations. *J. Phys. Chem. B*, 115(17):4911–4926, 2011. 6, 97

[31] S. C. Kamerlin, M. Haranczyk, and A. Warshel. Progress in ab initio qm/mm free-energy simulations of electrostatic energies in proteins: accelerated qm/mm studies of pka, redoc reactions and solvation free energies. *J. Phys. Chem. B*, 113(5):1253–1272, 2009. 6

[32] J. Michel, M. L. Verdonk, and J. W. Essex. Protein-ligand complexes: computation of the relative free energy of different scaffolds and binding modes. *J. Chem. Theory Comput.*, 3(5):1645–1655, 2007. 6, 7, 22, 25, 99, 104, 159

[33] C. Oostenbrink and W. F. van Gunsteren. Single-step perturbations to calculate free energy differences from unphysical reference states: Limits on size, flexibility, and character. *J. Comput. Chem.*, 24(14):1730–1739, 2003. 6, 39

[34] J. P. Ulmschneider and W. L. Jorgensen. Polypeptide folding using monte carlo sampling, concerted rotation, and continuum solvation. *J. Am. Chem. Soc.*, 126(6):1849–1857, 2004. 6

[35] J. Hritz and C. Oostenbrink. Efficient free energy calculations for compounds with multiple stable conformations separated by high energy barriers. *J. Phys. Chem. B*, 113(38):12711–12720, 2009. 6, 39

[36] P. A. Valiente, A. G. L, P. R. Batista, E. R. Caffarena, T. Pons, and P. G. Pascutti. New parametrization approaches of the lie method to improve free energy calculations of plmii-inhibitors complexes. *J. Comput. Chem.*, 31:2723–2734, 2010. 6

[37] S.-L. Chen, D.-X. Zhao, and Z.-Z. Yang. An estimation method of binding free energy in terms of abeem/mm and continuum electrostatics fused into lie method. *J. Comput. Chem.*, 32(2):338–348, 2011. 6

[38] K. Wichapong, M. Lawson, S. Pianwanit, S. Kokpol, and W. Sippl. Postprocessing of protein-ligand docking poses using linear response mm-pb/sa: application to wee1 kinase inhibitors. *J. Chem. Inf. Model.*, 50(9):1574–1588, 2010. 6

[39] T. Zhou, D. Huang, and A. Caflisch. Is quantum mechanics necessary for predicting binding free energy? *J. Med. Chem.*, 51(14):4280–4288, 2008. 6, 74

[40] C. J. Woods, J. W. Essex, and M. A. King. The development of replica-exchange-based free-energy methods. *J. Phys. Chem. B*, 107(49):13703–13710, 2003. 6, 27, 28, 39, 78, 117, 130, 143, 144

[41] J. W. Essex and W. L. Jorgensen. An empirical boundary potential for water droplet simulations. *J. Comput. Chem.*, 16(8):951–972, 1995. 6

[42] R. H. Henchman and J. W. Essex. Generation of opls-like charges from molecular electrostatic potential using restraints. *J. Comput. Chem.*, 20(5):483–498, 1999. 6

[43] S. E. Murdock, K. Tai, M. H. Ng, S. Johnston, B. Wu, H. Fangohr, C. A. Laughton, J. W Essex, and M. S. P. Sansom. Quality assurance for biomolecular simulations. *J. Chem. Theory Comput.*, 2(6):1477–1481, 2006. 6

[44] J. Michel, R. D. Taylor, and J. W. Essex. The parameterization and validation of generalized born models using the pairwise descreening approximation. *J. Comput. Chem.*, 25(14):1760–1770, 2004. 6

[45] J. Michel, R. D. Taylor, and J. W. Essex. Efficient generalized born models for monte carlo simulations. *J. Chem. Theory Comput.*, 2(3):732–739, 2006. 6, 77

[46] J. Michel and J. W. Essex. Hit identification and binding mode predictions by rigorous free energy simulations. *J. Med. Chem.*, 51(21):6654–6664, 2008. 7, 36, 43, 99, 147, 157

[47] C. Woods and J. Michel. Protoms2.1, a fortran program for monte carlo simulations of chemical systems. `http://www.protoms.org/`, 2005. 7, 83, 87, 130, 143, 149, 159

[48] R. Baumgartner, M. Walloschek, M. Kralik, A. Gotschlich, S. Tasler, J. Mies, and J. Leban. Dual binding mode of a novel series of dhodh inhibitors. *J. Med. Chem.*, 49(4):1239–1247, Feb 2006. 7, 44, 45, 46, 48, 49, 50, 54, 55, 56, 57, 58, 59, 60, 98, 99, 100, 101, 110, 111, 117, 118, 121, 122, 128, 134, 140, 141, 142, 146, 147, 151, 154, 156, 160

[49] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. am1-bcc model: I. method. *J. Comput. Chem.*, 21(2):132–146, 2000. 7, 80, 82, 88, 91, 110, 141, 142

[50] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, 1993. 7, 80, 88, 92

[51] G. A. Kaminski and W. L. Jorgensen. A quantum mechanical and molecular mechanical method based on cm1a charges: applications to solvent effects on organic equilibria and reactions. *J. Phys. Chem. B*, 102(10):1787–1796, 1998. 7, 74, 82, 88, 92

[52] C. J. Woods, F. R. Manby, and A. J. Mulholland. An efficient method for the calculation of quantum mechanics/molecular mechanics free energies. *J. Chem. Phys.*, 128:14109–1417, 1007. 7, 76, 77, 79, 92

[53] H. Resat and M. Mezei. Grand canonical monte carlo simulation of water positions in crystal hydrates. *J. Am. Chem. Soc.*, 116(16):7451–7452, 1994. 7

[54] J. Michel, J. Tirado-Rives, and W. L. Jorgensen. Prediction of the water content in protein binding sites. *J. Phys. Chem. B*, 113(40):13337–13346, 2009. 7

[55] E. Gallicchio and R. M. Levy. Recent theoretical and computational advances for modeling protein-ligand binding affinities. *Adv. Prot. Chem. Struct. Biol.*, 85:27–80, 2011. 8

[56] W. L. Jorgensen. The many roles of computation in drug discovery. *Science*, 303(5665):1813–1818, 2004. 8

[57] O. Guvench and A. D. MacKerell. Computational evaluation of protein-small molecule binding. *Curr. Opin. Struct. Biol.*, 19(1):56–61, 2009. 8

[58] N. Brooijmans and I. D. Kuntz. Molecular recognition and docking algoritms. *Annu. Rev. Biophys. Biomol. Struct.*, 32:335–373, 2003. 8, 160

[59] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, 2004. 8, 160

[60] M. K. Gilson and H. X. Zhou. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.*, 36(1):21–42, 2007. 8

[61] Y. Deng and B. Roux. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B*, 113:2234–2246, 2009. 8, 31, 39

[62] 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. Struc. Biol.*, 21(2):150–160, 2011. 8, 34, 35, 37, 38, 43

[63] A. R. Leach. *Molecular modelling: principles and applications (2nd Edition)*. Prentice Hall, 2 edition, 2001. 9, 12, 14, 19, 80, 88

[64] D. A. McQuarrie. *Statistical mechanics*. Harper and Row, New York, USA, 1976. 9, 10, 16, 19

[65] P. Atkins and J. De Paula. *Physical chemistry*. Oxford University Press, 8rev ed edition, 2006. 11, 19, 67, 100

[66] H. M. Senn and W. Thiel. Qm/mm methods for biomolecular systems. *Angew. Chem. Int. Edit.*, 48:1198–1229, 2009. 12, 42, 68, 69, 70, 71, 73, 74

[67] W. L. Jorgensen and J. Tirado-Rives. The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110(6):1657–1666, 1988. 12, 43, 67, 78, 81, 88, 89

[68] W. L Jorgensen, D. S. Maxwell, and J. Tirado-Rives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996. 12, 43, 67, 78, 81, 88, 89

[69] J. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074, 2000. 12, 43, 67, 78, 81, 89, 110, 142

[70] A. D. MacKerell, D. Bashford, D. Bellott, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102(18):3586–3616, 1998. 12, 43, 67, 78

[71] S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. de Vries. The martini force field: a coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111(27):7812–7824, 2007. 12

[72] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953. 15, 84

[73] W. L. Jorgensen. Perspective on 'equation of state calculations by fast computing machines'. *Theor. Chem. Acc.*, 103:225–227, 2000. 15

[74] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–144, 1997. 17, 27

[75] A. F. Voter. Classically exact overlayer dynamics: diffusion of rhodium clusters on rh(100). *Phys. Rev. B*, 34:6819–6829, 1986. 18

[76] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23:327–332, 1977. 19

[77] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem, Phys. Lett.*, 222(6):529–539, 1994. 26

[78] M. Zacharias, T. P. Straatsma, and J. A. McCammon. Separation-shifted scaling, a new scaling method for lennard-jones interactions in thermodynamic integration. *J. Chem. Phys.*, 100:9025–9032, 1994. 26, 145

[79] C. Chatfield and A. J. Collins. *Introduction to multivariate analysis*. Chapman and Hall, 1st ed edition, 1980. 29

[80] T. Steinbrecher, D. L. Mobley, and D. A. Case. Nonlinear scaling schemes for lennard-jones interactions in free energy calculations. *J. Chem. Phys.*, 127:214108–214121, 2007. 31

[81] M.R. Reddy and M.D. Erion. *Free energy calculations in rational drug design*. Springer, 2010. 31

[82] J. Michel and J. W. Essex. Prediction of protein-ligand binding affinity by free energy simulations: assumptions, pitfalls abd expectations. *J. Comput. Aid. Mol. Des.*, 24:639–658, 2010. 31

[83] M. R. Shirts, D. L. Mobley, and S. P. Brown. *Free energy calculations in structure-based drug design*. Cambridge University Press, 2010. 31

[84] C. D. Christ, A. E. Mark, and W. F. van Gunsteren. Basic ingredients of free energy calculations: a review. *J. Comput. Chem.*, 31:1569–1582, 2010. 31

[85] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. Meyer Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 112:535–542, 1977. 32, 49, 110

[86] R. W. W. Hooft, G. Vriend, C. Sander, and E. E. Abola. Errors in protein structures. *Nature*, 381:272–272, 1996. 32, 33

[87] A. M. Davis, S. A. St-Gallay, and G. J. Kleywegt. Limitations and lessons in the use of x-ray structural information in drug design. *Drug. Discov. Today*, 13(19-20):831–841, 2008. 33, 101

[88] S. Daopin, K. A. Piez, Y. Ogawa, and D. R. Davies. Crystal structure of transforming growth factor-beta 2: an unusual fold for the superfamily. *Science*, 257:369–373, 1992. 33

[89] M. P. Schlunegger and M. G. Gruetter. An unusual feature revealed by the crystal structure at 2.2 angstrom resolution of human transforming growth factor-beta 2. *Nature*, 358:430–434, 2003. 33

[90] S. Daopin, D. R. Davies, M. P. Schlunegger, and M. G. Gruetter. Comparison of two crystal structures of tgf-beta 2: the accuracy of refined protein structures. *Acta Crystallogr. Sect. D*, 50:85–92, 1994. 33

[91] G. Vriend. WHAT IF: A molecular modeling and drug design program. *J. Mol. Graphics*, 8(1):52–56, 1990. 33, 110, 142

[92] G. Klebe, M. Bhm, F. Dullweber, U. Graedler, H. Gohlke, and M. Hendlich. *Molecular modelling and prediction of bioactivity.* KLUWER Academic /Plenum Publ., 2000. 33

[93] E. J. Sacchettini, J. C. Kromminga, and J. I. Gordon. Escherichia coli-derived rat intestinal fatty acid binding protein with bound myristate at 1.5 angstrom resolution and i-fabp arg106 to gln with bound oleate at 1.74 å resolution. *J. Biol. Chem.*, 268(35):26375–26385, 1993. 34

[94] M. L. Verdonk, J. C. Cole, and R. Taylor. Superstar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.*, 289(4):1093–1108, 1999. 34

[95] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727 – 748, 1997. 34

[96] M. G. Jakoby, K. R. Miller, J. J. Toner, A. Bauman, L. Cheng, E. Li, and D. P. Cistola. Ligand-protein electrostatic interactions govern the specificity of retinol- and fatty acid-binding proteins. *Biochemistry*, 32(3):872–878, 1993. 34

[97] J. Thompson, N. Winter, D. Terwey, J. Bratt, and L. Banaszak. The crystal structure of the liver fatty acid-binding protein: a complex with two bound oleates. *J. Biol. Chem.*, 272:7140–7150, 1997. 34

[98] S. Boyce, D. L. Mobley, G. Rocklin, A. Graves, K. A. Dill, and B. K. Shoichet. Predicting ligand binding affinity with alchemical free energy methods in a polar model binding site. *J. Mol. Biol.*, 394:747–763, 2009. 35, 36, 37

[99] A. P. Graves, D. M. Shivakumar, S. E. Boyce, M. P. Jacobson, D. A. Case, and B. K. Shoichet. Rescoring docking hit lists for model binding sites: predictions and experimental testing. *J. Mol. Biol.*, 377:914–934, 2008. 35

[100] D. L. Mobley, A. P. Graves, J. D. Chodera, A. C. Reynolds, B. K. Shoichet, and K. A. Dill. Predicting absolute ligand binding free energies to a simple model site. *J. Mol. Biol.*, 371:1118–1134, 2007. 35, 36

[101] D. L. Mobley, J. D. Chodera, and K. A. Dill. On the use of orientational restraints and symmetrie corrections in alchemical free energy calculations. *J. Chem. Phys.*, 125:0849021–08490216, 2006. 35, 36, 39, 153

[102] D. L. Mobley and K. A. Dill. Binding of small-molecule ligands to proteins: 'what you see' is not always 'what you get'. *Struct. Fold. Design*, 17:489–498, 2009. 36

[103] E. Gallicchio, M. Lapelosa, and R. M. Levy. Binding energy distribution analysis method (bedam) for estimation of protein-ligand binding affinities. *J. Chem. Theory Comput.*, 6:2961–2977, 2010. 36, 39

[104] A. Aleksandrov, D. Thompson, and T. Simonson. Alchemical free energy simulations for biological complexes: powerful but temperamental. *J. Mol. Recognit.*, 23:117–127, 2009. 36

[105] P. Weinkam, F. E. Romesberg, and P. G. Wolynes. Chemical frustration in the protein folding landscape: grand canonical ensemble simulations of cytochrome c. *Biochemistry*, 48(11):2394–2402, 2009. 36

[106] D. L. Mobley, J. D. Chodera, and K. A. Dill. Confine-and-release method: obtaining correct binding free energies in the presence of protein conformational change. *J. Chem. Theory Comput.*, 3:1231–1235, 2007. 36, 38, 40

[107] S. P. Brown, S. W. Muchmore, and P. J. Hajduk. Healthy skepticism: assessing realistic model performance. *Drug Discov. Today*, 14:420–427, 2009. 37

[108] X. Kong and C. L. Brooks. Lambda-dynamics - a new approach to free energy calculations. *J. Chem. Phys.*, 105:2414–2423, 1996. 39, 105

[109] W. Jiang, M. Hodoscek, and B. Roux. Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics. *J. Chem. Theory Comput.*, 5:2583–2588, 2009. 39

[110] C. A. Chang, W. Chen, and M. K. Gilson. Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. USA*, 104:1534–1539, 2007. 39, 41

[111] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.*, 35:539–546, 2010. 39

[112] M. Lapelosa, G. F. Arnold, E. Gallicchio, and R. M. Levy. Antigenic characteristics of rhinovirus chimeras designed in silico for enhanced presentation of hiv-1 gp41 epitopes. *J. Mol. Biol.*, 397:752–766, 2010. 40

[113] M. Karplus. Dynamical aspects of molecular recognition. *J. Mol. Recognit.*, 23:102–104, 2010. 40

[114] W. Jiang and B. Roux. Free energy perturbation hamiltonian replica-exchange molecular dynamics for absolute ligand binding free energy calculations. *J. Chem. Theory Comput.*, 6:2559–2565, 2010. 40

[115] E. Perola and P. S. Charifson. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.*, 47:2499–2510, 2004. 40

[116] C. Gao, S. M. Park, and H. A. Stern. Accounting for ligand conformational restriction in calculations of protein-ligand binding affinities. *Biophys. J.*, 98:901–910, 2010. 41

[117] H. Okumura, E. Gallicchio, and R. M. Levy. Conformational populations of ligand-sized molecules by replica exchange molecular dynamics and temperature reweighting. *J. Comput. Chem.*, 31:1357–1367, 2010. 41

[118] Y. Lu, R. Wang, C.-Y. Yang, and S. Wang. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein-ligand complexes. *J. Chem. Inf. Model.*, 47:668–675, 2007. 41

[119] Z. Li and T. Lazaridis. Water at biomolecular binding interfaces. *Phys. Chem. Chem. Phys.*, 7(9):573–581, 2007. 41, 101

[120] H. Yu and S. W. Rick. Free energies and entropies of water molecules at the inhibitor-protein interface of dna gyrase. *J. Am. Chem. Soc.*, 131:6608–6613, 2009. 41

[121] R. Baron, J. P. Setny, and A. J. McCammon. Water in cavity-ligand recognition. *J. Am. Chem. Soc.*, 132:12091–12097, 2010. 41

[122] C. Barillari, J. Taylor, R. Viner, and J. W. Essex. Classification of water molecules in protein binding sites. *J. Am. Chem. Soc.*, 129(9):2577–2587, 2007. 41, 103, 129

[123] J. Michel, J. Tirado-Rives, and W. L. Jorgensen. Prediction of the water content in protein binding sites. *J. Phys. Chem. B*, 113:13337–13346, 2009. 41, 99, 101, 103, 105, 110, 112, 113, 114, 118

[124] D. Hamelberg and J. A. McCammon. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: double-decoupling method. *J. Am. Chem. Soc.*, 126(24):7683–7689, 2004. 41

[125] J. Michel, J. Tirado-Rives, and W. L. Jorgensen. Energetics of displacing water molecules from protein binding sites: consequences for ligand optimization. *J. Am. Chem. Soc.*, 131:15403–15411, 2009. 41, 62

[126] Y. Deng and B. Roux. Computation of binding free energy with molecular dynamics and grand canonical monte carlo simulations. *J. Chem. Phys.*, 128:115103–115110, 2008. 41

[127] X. Ge and B. Roux. Absolute binding free energy calculations of sparsomycin analogs to the bacterial ribosome. *J. Phys. Chem. B*, 114:9525–9539, 2010. 41

[128] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65:712–725, 2006. 42

[129] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D.E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins*, 78:1950–1958, 2010. 42, 67

[130] W. C. Swope, H. W. Horn, and J. E. Rice. Accounting for polarization when using fixed charge force fields. ii. method and application for computing effect of polarization cost on free energy of hydration. *J. Phys. Chem. B*, 114:8631–8645, 2010. 42

[131] M. R. Shirts and V. S. Pande. Solvation free energies of amino acid side chains for common molecular mechanics water models. *J. Chem. Phys.*, 122:134508–134512, 2005. 42, 96

[132] D. L. Mobley, C. I. Bayly, M. D. Cooper, M. R. Shirts, and K. A. Dill. Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. *J. Chem. Theory Comput.*, 5:350–358, 2009. 42, 43, 80, 96, 97

[133] R. B. Best and G. Hummer. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J. Phys. Chem. B*, 113:9004–9015, 2009. 42

[134] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schneiders, I. Haque, D. L. Mobley, D. S. Lambrecht, and R. A. DiStasio. Current status of the amoeba polarizable force field. *J. Phys. Chem. B*, 114:2549–2564, 2010. 42, 68

[135] C. M. Baker, P. E. M. Lopes, X. Zhu, B. Roux, J. Alexander, and D. MacKerell. Accurate calculation of hydration free energies using pair-specific lennard-jones parameters in the charmm drude polarizable force field. *J. Chem. Theory Comput.*, 6:1181–1198, 2010. 42

[136] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–1174, 2004. 43, 67, 78, 79, 82, 88, 110, 141, 142

[137] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv. Prot. Chem.*, 66:27–85, 2003. 43

[138] L. D. Schuler, X. Daura, and W. F. van Gunsteren. An improved gromos96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.*, 18:1205–1218, 2001. 43

[139] M. E. Jones. Pyrimidine nucleotide biosynthesis in animals: genes, enzymes, and regulation of ump biosynthesis. *Ann. Rev. Biochem.*, 49:253–279, 1980. 44, 45

[140] O. Bjrnberg, P. Rowland, S. Larsen, and K. F. Jensen. Active site of dihydroorotate dehydrogenase a from lactococcus lactis investigated by chemical modification and mutagenesis. *Biochemistry*, 36(51):16197–16205, 1997. 45

[141] M. Nagy, F. Lacroute, and D. Thomas. Divergent evolution of pyrimidine biosynthesis between anaerobic and aerobic yeasts. *Proc. Natl. Acad. Sci. USA*, 89(19):8966–8970, 1992. 45

[142] E. A. Neidhardt, S. R. Punreddy, J. E. McLean, L. Hedstrom, and T. H. Grossman. Expression and characterization of e. coli-produced soluble, functional human dihydroorotate dehydrogenase: a potential target for immunosuppression. *J. Mol. Microbiol. Biotechnol.*, 1(1):183–188, 1999. 45

[143] V. Hines and M. Johnston. Analysis of the kinetic mechanism of the bovine liver mitochondrial dihydroorotate dehydrogenase. *Biochemistry*, 28(3):1222–1226, 1989. 45

[144] F. S Nielsen, P. S. Andersen, and K. F. Jensen. The b form of dihydroorotate dehydrogenase from lactococcus lactis consists of two different subunits, encoded by the pyrdb and pyrk genes, and contains fmn, fad, and [fes] redox centers. *J. Biol. Chem.*, 271(46):29359–29365, 1996. 45

[145] R. L. Fagan, M. N. Nelson, P. M. Pagano, and B. A. Palfey. Mechanism of flavin reduction in class 2 dihydrooroate dehydrogenases. *Biochemistry*, 45(50):14926–14932, 2006. 45

[146] R. A. Pascal and C. T. Walsh. Mechanistic studies with deuterated dihydroorotates on the dihydroorotate oxidase from crithidia fasciculate. *Biochemistry*, 23:2745–2752, 1984. 45

[147] P. Nandi, G. H. Kingsley, and D. L. Scott. Disease-modifying antirheumatic drugs other than methotrexate in rheumatoid arthritis and seronegative arthritis. *Curr. Opin. Rheumatol.*, 20(3):251–256, 2008. 45

[148] K. L. McCance and S. E. Huether. *Pathophysiology*. Elsevier Mosby, 5th edition edition, 2008. 46

[149] J. A. Singh, R. Christensen, G. A. Wells, M. E. Suarez-Almazor, R. Buchbinder, M. A. Lopez-Olivo, G. E. Tanjong, and P. Tugwell. Biologics for rheumatoid arthritis: an overview of cochrane reviews. *Cochrane Db. Syst. Rev.*, 2009. 46

[150] U. Mueller-Ladner, T. Pap, R. E. Gay, M. Neidhart, and S. Gay. Mechanisms of disease: the molecular and cellular basis of joint destruction in rheumatoid arthritis. *Nat. Clin. Pract. Rheum.*, 1:102–110, 2005. 47

[151] S. Sanders and V. Harisdangkul. Leflunomide for the treatment of rheumatoid arthritis and autoimmunity. *Am. J. Med. Sci.*, 323(4):190–193, 2002. 46

[152] P. Pinto and M. Dougados. Leflunomide in clinical practice. *Acta Reumatolog. Portug.*, 31(3):215–224, 2006. 46

[153] G. J. Kleywegt, M. R. Harris, J. Y. Zou, T. C. Taylor, A. Waehlby, and T. A. Jones. The uppsala electron-density server. *Acta Crystallogr. D Bio.l Crystallogr.*, 60:2240–2249, 2004. 49

[154] S. Liu, E. A. Neidhardt, T. H. Grossman, T. Ocain, and J. Clardy. Structure of human dihydroorotate dehydrogenase in complex with antiproliferative agents. *Structure*, 8:25–33, 2000. 49

[155] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.2r1. August 2010. 51, 52, 98, 100

[156] U. Ryde, L. Olsen, and K. Nilsson. Quantum chemical geometry optimisations in proteins using crystallographic raw data. *J. Comp. Chem.*, 23:1058–1070, 2002. 62

[157] B. K. Shoichet, A. R. Leach, and I. D. Kuntz. Ligand solvation in molecular docking. *Proteins*, 34(1):4–16, 1999. 64

[158] C. Kalyanaraman, K. Bernacki, and M. P. Jacobson. Virtual screening against highly charged active sites: Identifying substrates of alpha beta barrel enzymes. *Biochemistry US*, 44(6):2059–2071, 2005. 64

[159] A. Ben Naim. *Molecular theory of solutions*. Oxford university press, 1984. 65, 93, 102, 103

[160] M. R. Shirts, J. W. Pitera, W. C. Swope, and V. S. Pande. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.*, 119(11):5740–5761, 2003. 66, 91, 96

[161] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995. 67, 89

[162] P. A. Kollman. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Acc. Chem. Res.*, 29(10):461–469, 1996. 67, 89

[163] M. Udier-Blagovic, P. Morales de T., S.A. Pearlman, and W.L. Jorgensen. Accuracy of free energies of hydration from cm1 and cm3 atomic charges. *J. Comput. Chem.*, 25:1322–1332, 2004. 67, 82

[164] J. Wang, W. Wang, P. A. Kollman, and D. A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.*, 25(2):247–260, 2006. 67, 79, 82, 110, 141, 142

[165] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D.S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. Clark, M. E. Johnson, and T. Head-Gordon. Current status of the amoeba polarizable force field. *J. Phys. Chem. B*, 114(8):2549–2564, 2010. 68

[166] A. Warshel and M. Levitt. Theoretical studies of enzymatic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103:227–249, 1976. 68

[167] F. Liao, Y. Huang, J. Ge, W. Zheng, K. Tedsree, P. Collier, X. Hong, and S. C. Tsang. Morphology-dependent interactions of zno with cu nanoparticles at the materials' interface in selective hydrogenation of co2 to ch3oh. *Angew. Chem. Int. Ed.*, 50:2162–2165, 2011. 68

[168] J. Sauer and M. Sierka. Combining quantum mechanics and interatomic potential functions in ab initio studies of extended systems. *J. Comput. Chem.*, 21:1470–1493, 2000. 68

[169] Y. Shiota, K. Suzuki, and K. Yoshizawa. Qm/mm study on the catalytic mechanism of benzene hydroxylation over fe-zsm-5. *Organometallics*, 25(13):3118–3123, 2006. 68

[170] O. Acevedo and W. L. Jorgensen. Cope elimination: elucidation of solvent effects from qm/mm simulations. *J. Am. Chem. Soc.*, 128(18):6141–6146, 2006. 68

[171] A. N. Alexandrova and W. L. Jorgensen. Why urea eliminates ammonia rather than hydrolyzes in aqueous solution. *J. Phys. Chem. B*, 111(4):720–730, 2007. 68

[172] A. Heyden, H. Lin, and D. G. Truhlar. Adaptive partitioning in combined quantum mechanical and molecular mechanical calculations of potential energy functions for multiscale simulations. *J. Phys. Chem. B*, 111(9):2231–2241, 2007. 69

[173] F. Maseras and K. Morokuma. Imomm: A new integrated ab initio + molecular mechanics geometry optimization scheme of equilibrium structures and transition states. *J. Comput. Chem.*, 16:1170–1179, 1995. 69

[174] S. Humbel, S. Sieber, and K. Morokuma. The imomo method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and sn2 reaction: Rcl+cl-. *J. Chem. Phys.*, 105:472065–472074, 1996. 70

[175] M. Svensson, S. Humbel, R. D. J. Froese, T. Matsubara, and K. Sieber, S. Morokuma. Oniom: A multi-layered integrated mo + mm method for geometry

optimizations and single point energy predictions. a test for diels-alder reactions and pt(p(t-bu)3)2 + h2 oxidative addition. *J. Phys. Chem.*, 100(50):19357–19363, 1996. 70

[176] D. Bakowies and W. Thiel. Hybrid models for combined quantum mechanical and molecular mechanical approaches. *J. Phys. Chem.*, 100:10580–10594, 1996. 70

[177] I. Antes and W. Thiel. *On the treatment of link atoms in hybrid methods*, chapter 5, pages 50–65. 2004. 70

[178] M. A. Thompson and G. K. Schenter. Excited states of the bacteriochlorophyll b dimer of rhodopseudomonas viridis: A qm/mm study of the photosynthetic reaction center that includes mm polarization. *J. Phys. Chem.*, 99(17):6374–6386, 1995. 72

[179] M. A. Thompson. Qm/mmpol: A consistent model for solute/solvent polarization. application to the aqueous solvation and spectroscopy of formaldehyde, acetaldehyde, and acetone. *J. Phys. Chem.*, 100(34), 1996. 72

[180] J. Gao and K. Byun. Solvent effects on the n .fwdarw. .pi.* transition of pyrimidine in aqueous solution. *Theoret. Chem. Acc.*, 96(3):151–156, 1997. 72

[181] Y. Zhang, H. Lin, and D. G. Truhlar. Self-consistent polarization of the boundary in the redistributed charge and dipole scheme for combined quantum-mechanical and molecular-mechanical calculations. *J. Chem. Theory Comput.*, 3(4):1378–1398, 2007. 72

[182] K. Nam, J. Gao, and D. M. York. An efficient linear-scaling ewald method for long-range electrostatic interactions in combined qm/mm calculations. *J. Chem. Theory Comput.*, 1(1):2–13, 2005. 72

[183] A. R. Dinner, X. Lopez, and M. Karplus. A charge-scaling method to treat solvent in qm/mm simulations. *Theoret. Chem. Acc.*, 109(3):118–124, 2003. 72

[184] R. B. Murphy, D. M. Philipp, and R. A. Friesner. A mixed quantum mechanics/molecular mechanics (qm/mm) method for large-scale modeling of chemistry in protein environments. *J. Comput. Chem.*, 21:1442–1457, 2000. 73

[185] M. Freindorf, Y. Shao, T. R. Furlani, and J. Kong. Lennard-jones parameters for the combined qm/mm method using the b3lyp/6-31g*/amber potential. *J. Comput. Chem.*, 26:1270–1278, 2005. 73

[186] E. Rosta, M. Klahn, and A. Warshel. Towards accurate ab initio qm/mm calculations of free-energy profiles of enzymatic reactions. *J. Phys. Chem. B*, 110(6):2934–2941, 2006. 74

[187] T. H. Rod and U. Ryde. Accurate qm/mm free energy calculations of enzyme reactions: Methylation by catechol o-methyltransferase. *J. Chem. Theory Comput.*, 1(6):1240–1251, 2005. 74, 75

[188] M. Strajbl, G. Hong, and A. Warshel. Ab initio qm/mm simulation with proper sampling: 'first principle' calculations of the free energy of the autodissociation of water in aqueous solution. *J. Phys. Chem. B*, 106(51):13333–13343, 2002. 74, 77

[189] R. Iftimie, D. Salahub, and J. Schofield. An efficient monte carlo method for calculating ab initio transition state theory reaction rates in solution. *J. Chem. Phys.*, 119:11285–11298, 2003. 74, 75, 76, 77, 78

[190] P. Bandyopadhyay. Accelerating quantum mechanical/molecular mechanical sampling using pure molecular mechanical potential as an importance function: The case of effective fragment potential. *J. Chem. Phys.*, 122:91102–91106, 2005. 74

[191] M. R. Reddy, U. C. Singh, and M. D. Erion. Use of a qm/mm-based fep method to evaluate the anomalous hydration behavior of simple alkyl amines and amides: Application to the design of fbpase inhibitors for the treatment of type-2 diabetes. *J. Am. Chem. Soc.*, 133(21):8059–8061, 2011. 75

[192] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. 75

[193] B. Hetenyi, K. Bernacki, and B. J. Berne. Multiple 'time step' monte carlo. *J. Chem. Phys.*, 117:8203–8208, 2002. 77

[194] K. Bernacki, B. Hetenyi, and B. J. Berne. Multiple 'time step' monte carlo simulations: Application to charged systems with ewald summation. *J. Chem. Phys.*, 121(44):1755195–1755202, 2004. 77

[195] 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.*, 79(2):926–935, 1983. 80, 83, 91, 111, 143, 147, 156

[196] W. L. Jorgensen and J. Tirado-Rives. Molecular modeling of organic and biomolecular systems using boss and mcpro. *J. Comput. Chem.*, 26(16):1689–1700, 2005. 81

[197] J. W. Storer, D. J. Giesen, C. J. Cramer, and D. G. Truhlar. Class iv charge models: a new semiempirical approach in quantum chemistry. *J. Comput. Aided Mol. Des.* 81

[198] J. D. Thompson, C. J. Cramer, and D. G. J. Truhlar. Parameterization of charge model 3 for am1, pm3, blyp, and b3lyp. *J. Comput. Chem.*, 24:1291–1301, 2003. 81

[199] G. Schaftenaar and J. H. Noordik. Molden: a pre- and post-processing program for molecular and electronic structures. *J. Comput. Aid. Mol. Des.*, 14:123–134, 2000. 82

[200] D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, B. Wang, D. A. Pearlman, M. Crowley, S. J. Brozell, R. E. Duke, R. Luo, K. B. Merz, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J. W. Caldwell, W. S. Ross, and P. A. Kollman. Amber 8. the scripps research institute, la jolla. 2004. 82, 111, 143

[201] E. Vanquelef, S. Simon, G. Marquant, E. Garcia, G. Klimerak, J. C. Delepine, P. Cieplak, and F.-Y. Dupradeau. R.e.d. server: a web service for deriving resp and esp charges and building force field libraries for new molecules and molecular fragments. *Nucl. Acids Res.*, 39:511–517, 2011. 82

[202] C. J. Cramer, G. C. Lynch, and D. G. Truhlar. Amsol, quantum chemistry program exchange program no. 606, qcpe, 1992. 82

[203] L. Martinez, R. Andrade, E. G. Birgin, and J. M. Martinez. Packmol: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.*, 30:2157–2164, 2009. 83

[204] C. Woods. Sire, advanced molecular simulation framework. `http://www.siremol.org/`, 2005. 87, 93, 97

[205] P. J. Knowles and N. C. Handy. A determinant based full configuration interaction program. 54:75–83, 1989. 87

[206] J.S. Bader and D. Chandler. Computer simulation study of the mean forces between ferrous and ferric ions in water. *J. Phys. Chem.*, 96:6423–6427, 1992. 89

[207] R. C. Rizzo and W. L. Jorgensen. OPLS All-Atom model for amines: Resolution of the amine hydration problem. *J. Am. Chem. Soc.*, 121(20):4827–4836, 1999. 92

[208] J. E. Ladbury. Just add water! the effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chem. Biol.*, 3(12):973–980, 1996. 101

[209] B. C. Roberts and R. L. Mancera. Ligand-protein docking with water molecules. *J. Chem. Inf. Model.*, 48(2):397–408, 2008. 101

[210] M. L. Verdonk, G. Chessari, J. C. Cole, M. J. Hartshorn, C. M. Murray, J. W. Nissink, R. D. Taylor, and R. Taylor. Modeling water molecules in protein-ligand docking using gold. *J. Med. Chem.*, 48(20):6504–6515, 2005. 101

[211] Carugo O. and Bordo D. How many water molecules can be detected by protein crystallography? *Acta. Crystallogr. D. Biol. Crystallogr.*, 55:479–483, 1999. 101

[212] P. J. Goodford. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.*, 28(7):849–857, 1985. 101

[213] J. L. Gelp, S. G. Kalko, X. Barril, J. Cirera, X. de La Cruz, F. J. Luque, and M. Orozco. Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins*, 45(4):428–437, 2001. 101

[214] R. Malham, S. Johnstone, R. J. Bingham, E. Barratt, S. E. Phillips, C. A. Laughton, and S. W. Homans. Strong solute-solute dispersive interactions in a protein-ligand complex. *J. Am Chem Soc.*, 127(48):17061–17067, 2005. 101

[215] Z. Li and T. Lazaridis. The effect of water displacement on binding thermodynamics: concanavalin a. *J Phys Chem B.*, 109(1):662–670, 2005. 101

[216] T. Young, R. Abel, B. Kim, B. J. Berne, and R. A. Friesner. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. USA*, 104(3):808–813, 2007. 101

[217] R. Abel, T. Young, R. Farid, B. J. Berne, and R. A. Friesner. Role of the active-site solvent in the thermodynamics of factor xa ligand binding. *J. Am. Chem. Soc.*, 130(9):2817–2831, 2008. 101

[218] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.*, 72(3):1047–1069, 1997. 102, 104

[219] A. Ben Naim and Y. Marcus. Solvation thermodynamics of non-ionic solutes. *J. Chem. Phys.*, 81:2016–2027, 1984. 103

[220] S. Boresch, F. Tettinger, M. Leitgeb, and M. Karplus. Absolute binding free energies: a quantitative approach for their calculation. *J. Phys. Chem. B*, 107:9535–9551, 2003. 104

[221] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free energy estimation: Umbrella sampling. *J. Comput. Phys.*, 23:187–199, 1977. 105

[222] H. Resat and M. Mezei. Grand canonical monte carlo simulation of water positions in crystal hydrates. *J. Am. Chem. Soc.*, 116(16):7451–7452, 1994. 107

[223] D. J. Adams. Chemical potential of hard-sphere fluids by monte carlo methods. *Mol. Phys.*, 28:1241–1252, 1974. 107

[224] D. J. Adams. Grand canonical ensemble monte carlo for a lennard-jones fluid. *Mol. Phys.*, 29:307–311, 1975. 107, 108

[225] H. J. Woo, A. R. Dinner, and B. Roux. Grand canonical monte carlo simulations of water in protein environments. *J. Chem. Phys.*, 121(13):6392–6400, 2004. 108, 109

[226] M. Mezei. A cavity-biased monte carlo method for the computer simulation of fluids. *Mol. Phys.*, 40:901–906, 1980. 108, 109

[227] M. Mezei. Grand-canonical ensemble monte carlo simulation of dense fluids: lennard-jones, soft spheres and water. *Mol. Phys.*, 61:565–582, 1987. 108, 109

[228] D. H. L. Yau, S. Y. Liem, and K-Yu Chan. A contact cavity-biased method for grand canonical monte carlo simulations. *J. Chem. Phys.*, 101(9):7918–7924, 1994. 109

[229] M. Mezei and F. Guarnieri. Simulated annealing of chemical potential: a general procedure for locating bound waters. application to the study of the differential hydration propensities of the major and minor grooves of dna. *J. Am. Chem. Soc.*, 118(35):8493–8494, 1996. 109

[230] J.M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285(4):1735 – 1747, 1999. 110, 142

[231] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, and D. C. Richardson. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nuc. Acids Res.*, 35(Web Server issue):W375–383, 2007. 110, 142

[232] A. W. Schuttelkopf and D. M. F. van Aalten. Prodrg - a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallogr.*, pages 1355–1363, 2004. 111, 142

[233] Cheng Y.-C. and W. H. Prusoff. Relationship between the inhibition constant (ki) and the concentration of inhibitor which causes 50 per cent inhibition (i50) of an enzymatic reaction. *Biochem. Pharmacol.*, 22(23):3099 – 3108, 1973. 112, 144, 146, 155

[234] Michael J. Hartshorn. Astexviewer: a visualisation aid for structure-based drug design. *J. Comput. Aid. Mol. Des.*, 16:871–881, 2002. 113, 120, 124, 125, 126, 127, 132, 133, 136, 137, 138

[235] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *J. Mol. Graphics*, 14:33–38, 1996. 118, 128, 134

[236] K. M. Holtz, B. Stec, J. K. Myers, S. M. Antonelli, T. S. Widlanski, and E. R. Kantrowitz. Alternate modes of binding in two crystal structures of alkaline phosphatase-inhibitor complexes. *K. Mol. Graphics Modell.*, 9:907–915, 2000. 147

[237] M. Seifert, F. Schmitt, T. Herz, and B. Kramer. ProPose: a docking engine based on a fully configurable proteinligand interaction model. *J. Mol. Model.*, 10(5-6):342–357, 2004. 147

[238] S. Z. Banba and C. L. B. Guo. Efficient sampling of ligand orientations and conformations in free energy calculations using the l-dynamics method. *J. Chem. Phys.*, 104:6903–6910, 2000. 148

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other British or foreign examination board.

The thesis work was conducted under the supervision of Prof. Jonathan W. Essex at the University of Southampton.

Southampton, October 2011