University of Southampton

The Use of Free Energy Simulations as Scoring Functions

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

Julien MICHEL
School of Chemistry
University of Southampton

August 2006

Supervisor : Dr. J. W. Essex
Adviser : Prof. G. A. Attard
Industrial Supervisors : Dr. M. L. Verdonk and Dr. R. D. Taylor

Abstract

UNIVERSITY OF SOUTHAMPTON
FACULTY OF SCIENCE
CHEMISTRY

Doctor of Philosophy

THE USE OF FREE ENERGY SIMULATIONS AS SCORING FUNCTIONS

by Julien Michel

The combination of theories of implicit solvation, derived from the laws of classical electrostatics, with theories of free energy calculation, derived from the principles of statistical thermodynamics, is investigated. The aim of the investigation is to develop an efficient protocol for the prediction of the binding free energy of protein-ligand complexes. The Generalised Born Surface Area (GBSA) theory of implicit solvation and the Replica Exchange Thermodynamic Integration (RETI) method were selected for this work. A set of optimum parameters were derived for a GBSA force field, compatible with the General Amber Force Field (GAFF). The resulting implicit models of water were validated by assessing their ability to reproduce the salient features of the potentials of mean force for the association of several small molecules in solution. A protocol that combines efficiently GBSA potential energy function evaluation with Monte Carlo sampling was devised and validated by calculating the relative binding free energy of selected protein-ligand complexes. The implicit solvent free energy calculation protocol was then applied to determine the relative binding free energies of a set of congeneric inhibitors to two different proteins, cyclooxygenase-2 and neuraminidase. The method was found to perform as well or better than established binding free energy calculation protocols, while converging free energy estimates faster. Established protocols can typically only calculate relative free energies between structurally similar compounds. A methodology was devised to permit the calculation of the relative free energy of structurally different compounds. The method extends thus the scope of free energy calculations. It is expected that the combination of the two methodologies will allow free energy calculations to be applied to a wider variety of problems, of direct relevance to the pharmaceutical industry.

“I stand at the seashore, alone, and start to think. There are the rushing waves ... mountains of molecules, each stupidly minding its own business ... trillions apart ... yet forming white surf in unison.

Ages on ages ... before any eyes could see ... year after year ... thunderously pounding the shore as now. For whom, for what ? ... on a dead planet, with no life to entertain.

Never at rest ... tortured by energy ... wasted prodigiously by the sun ... poured into space. A mite makes the sea roar.

Deep in the sea, all molecules repeat the patterns of one another till complex new ones are formed. They make others like themselves ... and a new dance starts.

Growing in size and complexity ... living things, masses of atoms, DNA, protein ... dancing a pattern ever more intricate.

Out of the cradle onto the dry land ... here it is standing ... atoms with consciousness ... matter with curiosity.

Stands at the sea ... wonders at wondering ... I ... a universe of atoms ... an atom in the universe. ”

Richard P. Feynman

Acknowledgements

It is customary to open this section by acknowledging the support of your supervisor. I am indeed very grateful for Jon's gentle guidance that kept me out of "dead-ends", kept me focused on the "big picture" and yet provided me with a lot of independence. If I feel confident in carrying out independent research projects today or in the future, this will be in large part thanks to his influence.

I cannot quantify the impact of Dr. Christopher Woods on this project, but it is clear that, had he not written ProtoMC during his PhD, and patiently taught me the basics of computer simulations and programming, the outcomes of my research would have been different.

At times, a PhD can be quite stressful and I feel fortunate to have completed it in a friendly environment. Thanks to all the current and former members of the group and especially Sebastien Fouchet, Justine Taylor, Luca Fenu and Ben Cossins.

Theoretical research can be frustrating and I have sometimes wondered if my work is of any use ! I am grateful to Astex Therapeutics for funding this research and suggesting me protein-ligand systems that matter. Thanks to Dr Richard D. Taylor, Dr Marcel L. Verdonk and Dr. Christopher W. Murray for giving me an idea of what it is like to carry out research in a private company.

Sometimes you need to get out of the lab...And I am thankful for the memorable moments I spent with my friends in England, especially Nikos, Niall and Amaury.

My family has always been very supportive and encouraged me to leave France to find the right opportunities, even if it now means that I do not see them as often as I would like to. Merci papa et maman.

I kept the last paragraph for you Maria. Meeting you was a blessing and ever since then I feel you gave my life a higher meaning. I sense your presence behind everything I do and I cannot see purpose without you. Thank you.

Contents

1 Computer aided drug design **1**

1.1 Introduction 1

1.2 Docking and scoring in structure based drug design 2

1.3 Key notions of statistical mechanics 5

1.3.1 A brief definition 5

1.3.2 Concepts and postulates of statistical mechanics 5

1.3.3 The Boltzmann distribution 6

1.4 Classical potentials 9

1.5 Sampling methods 11

1.5.1 Metropolis Monte Carlo 11

1.5.2 Monte Carlo moves 13

1.5.3 Molecular dynamics 16

1.6 Rigorous free energy calculation methods 17

1.6.1 The importance of the free energy 17

1.6.2 Absolute free energy calculation 17

1.6.3 Free energy perturbation 19

1.6.4 Thermodynamic integration 20

1.6.5 Replica exchange thermodynamic integration 21

1.6.6 One-step multiple perturbed states 22

1.6.7 Slow and fast growth 23

1.7 Calculating errors in free energy simulations 24

1.8 Approximate Free Energy Methods 26

1.8.1 Linear interaction energy 26

1.8.2 MM/PBSA 27

1.9 Continuum solvation 28
1.9.1 The Born equation 29
1.9.2 Electrostatics and the generalised Born model 31
1.9.3 The apolar component of solvation 34
1.10 Conclusion and outline of the thesis 35

2 Parameterisation and validation of a generalised Born surface area model of water 37
2.1 Introduction 37
2.2 Selecting a force field 38
2.3 Representing water 39
2.3.1 Construction of a dataset 40
2.3.2 The adjustable parameters of a GBSA model 41
2.3.3 Deriving a set of optimum parameters with a genetic algorithm 43
2.3.4 Comparison with Poisson Boltzmann calculations 47
2.3.5 Behaviour of the parameterised models in potential of mean force calculations 51
2.3.6 Cause of errors in the PMFs 68
2.4 Conclusion 70

3 Efficient generalised Born models for Monte Carlo simulations 72
3.1 Introduction 72
3.2 Generalised Born in a Monte Carlo Simulation 72
3.3 Implementing a generalised Born force field suitable for free energy calculations 73
3.4 Selecting a test system and setup 74
3.5 Approximated generalised Born Potential 77
3.6 Simplified Sampling Potential 81
3.6.1 Theory 81
3.6.2 Application to a GBSA model 84
3.7 Conclusion 86

4 Application to a protein-ligand system : cyclooxygenase-2 90
4.1 Introduction 90

4.2 Presentation of the system 91
4.3 System setup and simulation protocols 94
4.4 Explicit solvent simulations results 96
4.5 Generalised Born simulations results 102
4.6 Influence of protein flexibility 113
4.7 Computational cost and convergence 116
4.8 Comparison with empirical scoring functions 119
4.9 Conclusion 121

5 Application to a protein-ligand system : neuraminidase 123
5.1 Introduction 123
5.2 Presentation of the system 123
5.3 System setup and simulation protocols 127
5.4 Explicit solvent simulations results 129
5.5 Generalised Born simulations results 142
5.6 Influence of protein flexibility 153
5.7 Importance of configurational averaging 156
5.8 Computational cost and convergence 157
5.9 Comparison with empirical scoring functions 159
5.10 Conclusion 161

6 Alternative pathways in free energy calculations 164
6.1 Introduction 164
6.2 Single and dual topology methods 164
6.3 Softening the intermolecular interactions 168
6.4 Solvation free energy calculations 170
6.4.1 Relative solvation free energy of ethane and methanol . . 171
6.4.2 Relative solvation free energy of benzene, ethane and methanol 181
6.5 Binding free energy calculations 184
6.5.1 Relative binding free energy of celecoxib analogues . . . 184
6.5.2 Relative binding free energy of diclofenac and celecoxib . 196
6.6 Conclusion 202

7 Concluding remarks 204

A Solving the integrals of chemical problems 213

A.1 The curse of dimensionality 213

A.2 Importance Sampling 216

A.3 Markov Chains 218

A.3.1 Definition 218

A.3.2 Detailed Balance 220

A.3.3 Performance of a Markov chain 221

A.3.4 The Metropolis Monte Carlo algorithm 222

A.4 The connection with molecular simulations 223

Chapter 1

Computer aided drug design

> "Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry.... if mathematical analysis should ever hold a prominent place in chemistry – an aberration which is happily almost impossible – it would occasion a rapid and widespread degeneration of that science."
>
> Auguste Comte

1.1 Introduction

This work aims at developing novel computer methodologies that allow for the reliable and accurate calculation of the relative affinities of a range of ligands to a protein. Even with modern computational resources, present day techniques are time consuming and limited in scope. The general availability of a fast, reliable, and accurate computational method to predict binding free energies would be of significant assistance to the pharmaceutical industry.

1.2 Docking and scoring in structure based drug design

Modern drug discovery makes regular use of computational methodologies to assist in the development of new drugs.[1]

Invaluable insights can be obtained from the analysis of the crystallographic or NMR-derived structures that describe the interactions between a small molecule inhibiting an important enzyme. The wealth of information thus obtained is used to direct the efforts of synthetic chemists towards the design of better drugs, more quickly.[2–5] It is unfortunate that obtaining a well resolved X-ray structure of a protein-ligand complex often represents a significant effort, even with the advent of tools that can help to automate such undertaking.[3,6,7] Furthermore, one would ideally like to be able to inspect how thousands of different small molecules could bind to a particular target. The docking methodology has been developed to address such issue. Given the three dimensional structure of the binding site of a particular target, is it possible to use a computational method to predict the orientation a small molecule would adopt inside ?

Because of the potential rewards, extensive research has been conducted in this area over the last two decades[8–11] and several methodologies to solve such problem have been proposed.[12–17] A good docking algorithm must possess two essential ingredients to be successful. First, it must be equipped with a good search strategy. Drug-like molecules often contain several rotatable bonds and in order to determine the optimum configuration the ligand should adopt when bound to a protein, all of the most likely configurations must be enumerated quickly. The problem is further complicated if the flexibility of the protein is to be considered. Often, docking algorithms do not consider protein flexibility because it increases hugely the search space and does not permit the consideration of several thousands ligands within a reasonable amount of time and computational resource. Because the search strategy amounts to an optimisation problem, a wide variety of techniques, often borrowed from operational research, are employed by different docking programs.

To analyse the configurations generated by the search strategy, it is necessary

to rank the proposed configurations according to their plausibility. To achieve this aim, scoring functions are used to evaluate a score which typically depends on a range of physical descriptors and energetic terms that describe the main features of protein-ligand binding.[11,18] A wide variety of scoring functions exists in the literature.[19–26] Typical scoring functions can be classified in three main categories: empirical, knowledge-based and physical (based on a molecular mechanics force field) scoring functions. An example of each of those is given below.

. The GOLD scoring function[21] consists of three terms, a hydrogen bonding term, a van der Waals term, and an internal energy term. The van der Waals interaction energy uses a 4-8 Lennard-Jones potential. The hydrogen bonding term is based on empirical values for the strength of hydrogen bonds between different atom types. The hydrogen bond energy is weighted based on the angle and the bond length between the donor and acceptor. The van der Waals term account for hydrophobic interactions between the ligand and the protein. The internal energy term consider the energy necessary for the ligand to adopt its configuration in the binding site, which may differ from the one it would adopt freely in solution. The total energy is a weighted sum of the three terms, making this a semi-empirical scoring function.

. Knowledge-based potentials are derived using observed frequencies of atom-atom interactions in known structures of protein-ligand complexes. If these frequencies are converted into free energies using Boltzmann distributions, the potentials are generally called potentials of mean force (PMF). The main difference with empirical potentials is that no binding data are needed, which has the advantage that relatively large training sets can be easily devised. Another advantage is that PMF based scoring functions include implicitly, in principle, all the forces that play a role in complex formation. A disadvantage is that rather large and well balanced datasets may be necessary to reflect the diversity of protein binding sites and ligand functional groups. An example of a PMF scoring function is DrugScore.[24]

. Some earlier scoring functions use force fields derived from the field of molecular simulations. The AutoDOCK scoring function uses the Coulomb

and van der Waals terms of force field functions.[13] To account for the screening effect of the solvent on electrostatic interactions, a distance-dependent dielectric constant is used. Internal ligand energies and entropic terms are completely ignored.

While a rigorous comparison of different docking program is often difficult,[27] it is generally accepted that modern docking methods obtain successfully the correct binding mode of a ligand about 70% of the time, making docking a valuable tool in the search for new inhibitors of a protein.

Given a cheap computational method that finds the right orientation of a small molecule in the binding site of a protein, it is possible to generate rapidly thousands of dockings. The typical scoring functions discussed above seem able to discriminate reasonably between compounds that would bind very poorly or not at all and those that would bind.[18] Unfortunately, most of these methods fall short of expectations when one tries to order a series of known binders according to their potency.[28] It is perhaps not surprising when one realises that to correctly determine the binding mode of a small molecule, one usually needs to satisfy a set of hydrogen bonds constraints and hydrophobic contacts. The correct evaluation of a binding free energy is however, much more sensitive to the underlying potential energy surface and the proper consideration of solvation and entropic effects, which are precisely what most common scoring functions lack. If a cheap computational method that could robustly predict the binding energy of a series of related inhibitors existed, it would become possible to perform lead optimisation *in silico*. Such an ability would reduce the time and effort necessary to produce a new drug by cutting down the number of compounds that have to be synthetised and tested for activity in a laboratory.

Much is known about the physical principles of ligand-protein interactions and there exists a formalism to calculate binding free energies from first principles. The purpose of the remainder of this chapter will be to introduce the necessary background to understand free energy calculations and discuss how they can be applied in practice to drug design.

1.3 Key notions of statistical mechanics

1.3.1 A brief definition

The science of thermodynamics arose from the realisation that material systems could be described in terms of a small number of parameters that were related to each other by simple laws. For instance a simplified form of the ideal gas law that relates the product of the pressure and volume of a gas to its temperature was proposed by Boyle as early as 1661.[1] In fact, the bulk of modern thermodynamics was developed without ever having to conjure the atomistic picture of matter to mind. However, in the nineteenth century, a number of physicists were growing uneasy with the empirical foundations of thermodynamic. Gibbs wrote in the introduction to his book *Elementary principles in statistical mechanics*.[29]

> "The laws of thermodynamics, as empirically determined, express the approximate and probable behavior of systems of a great number of particles, or, more precisely, they express the laws of mechanics for such systems as they appear to beings who have not the fineness of perception to enable them to appreciate quantities of the order of magnitude of those which relate to single particles, and who cannot repeat their experiments often enough to obtain any but the most probable results."

Thus statistical mechanics can be thought of as a branch of physics which tries to explain the laws of thermodynamics from the mechanical properties of collections of molecules. Because it relies on a probabilistic picture the term statistical is adequate. Much of the early developments in statistical mechanics are due to Clausius, Maxwell and Boltzmann.

1.3.2 Concepts and postulates of statistical mechanics

Statistical Mechanics take the view that the macroscopic properties (i.e, observables such as volume, compressibility...) of a system arise naturally from the mi-

[1] It appears the law was formulated in reply to criticisms from the Jesuit Franciscus Linus to the work of Boyle and Hooke (inventor of the microscope), published in 1660 under the title *New Experiments Physico-Mechanical*. The law, which was formulated as an hypothesis then (even though it was backed up by several experiments), was included in the second edition of this book and published in 1662.

croscopic (i.e atomistic) behaviour of that system. We now consider that the system of interest is a collection of N particles in a box. At any instant, each particle has a given momentum and occupies a point in space. The set of 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. Under a given set of conditions (for example, constant volume of the box and constant temperature), the collection of particles naturally adopt different set of positions/momenta through time and equivalently, follows a time trajectory in phase space. Instead of focusing on the time evolution of a trajectory in phase space, it is possible to imagine that the collection of microstates the system can adopt naturally forms an ensemble. In equilibrium, the microstates in that ensemble are distributed according to a probability density $\pi(\Gamma)$. The two important postulates of statistical mechanics can now be formulated.[29]

1. *Postulate of equal a priori probabilities:* This postulate states that two microstates i,j that have the same energy are equally probable and therefore $\pi_i = \pi_j$.

2. *Postulate of ergodicity:* This postulate states that the time evolution of a trajectory in phase space is such that one is guaranteed to visit eventually all the states which have a non-zero probability of existence. This postulate means that the time average of a property equals the ensemble average of that property at equilibrium.

1.3.3 The Boltzmann distribution

Under these conditions, it is possible to derive an expression for the probability density π for a particular ensemble. In the rest of this section 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. Similar derivations leading to different expressions can be obtained for different ensembles.

The probability distribution for the NVT ensemble is[30]

$$\pi_{NVT}(i) = \frac{1}{Q_{(NVT)}} \exp(-\beta E_i) \tag{1.1}$$

E_i is the energy of state i, β is equal to $\frac{1}{kT}$ where T is the temperature and k the Boltzmann constant. The exponential term is known as the Boltzmann factor and represent the weight of the state in that ensemble. Q_{NVT} is a normalisation constant called the partition function. π_{NVT} is often referred to as the Boltzmann distribution. When dealing with systems with a finite number of states, the partition function Q_{NVT} is simply the sum of the Boltzmann factor of each state, i.e

$$Q_{NVT} = \sum_i \exp(-\beta E_i) \tag{1.2}$$

In the limit of very large number of states, equation 1.2 can be replaced by an integral and it is then more appropriate to consider the phase space $\Gamma = (p^N, r^N)$ as a continuum and write under the conditions of the classical approximation[30]

$$Q_{NVT} = \frac{1}{N!}\frac{1}{h^{3N}} \int\int dp^N dr^N \exp(-\beta E(p^N, r^N)) \tag{1.3}$$

The term in $\frac{1}{N!}$ is necessary when dealing with indistinguishable particles as in this case two configurations that differs only by the exchange of coordinates/momenta between two particles correspond to only one real configuration. This term must be adjusted if the system is a mixture of different particles. The other term involves the Planck constant h and is of quantum mechanical origin. In the rest of this document we will occasionally define Q using either equation 1.2 or 1.3. We assume that any future conclusions drawn from Q applies to either form unless otherwise noted.

The connection between a macroscopic observable A_{obs} and its microscopic value $A(\Gamma)$ can be made through the following relationship:

$$A_{obs} = \langle A_{ens} \rangle = \frac{1}{Q_{NVT}} \int\int dp^N dr^N A(p^N, r^N) \exp(-\beta E(p^N, r^N)) \tag{1.4}$$

Equation 1.4 states that the ensemble average $\langle A_{ens} \rangle$ is equal to the macroscopic observable A_{obs}. Note how this ensemble average is calculated by integrating over all the positions and momenta that the set of N particles can adopt.

The partition function in equation 1.3 is often used in a simplified form. The Energy $E(p^N, r^N)$ can be separated into a kinetic part $K(p^N)$ and a potential part

$U(r^N)$. The kinetic part is also called the ideal part because a system where the only energy term is of kinetic origin would be an ideal gas. The potential part is called the excess part by reference to thermodynamics where deviations from an ideal system are attributed to 'excess' terms.

$$\begin{aligned} Q_{NVT} &= \frac{1}{N!}\frac{1}{h^{3N}}\int\int \exp\Big(-\beta(U(r^N)+K(p^N))\Big)\,dp^N\,dr^N \\ &= \frac{1}{N!}\frac{1}{h^{3N}}\int\int \exp\Big(-\beta U(r^N)\Big)\exp\Big(-\beta K(p^N)\Big)\,dp^N\,dr^N \\ &= \frac{1}{N!}\frac{1}{h^{3N}}\int \exp\Big(-\beta K(p^N)\Big)\,dp^N\int \exp\Big(-\beta U(r^N)\Big)\,dr^N \\ &= Q_{NVTid}.Q_{NVTexcess} \end{aligned} \tag{1.5}$$

The ideal part can be evaluated analytically.

$$Q_{NVTid} = \frac{V^N}{N!\Lambda^{3N}} \text{ where } \Lambda = (h^2/2\pi m k_B T)^{\frac{1}{2}}, \tag{1.6}$$

where Λ is the thermal de Broglie wavelength , m the mass of each particle and V the volume of the system.

$Q_{NVTexcess}$ is often written as:

$$Q_{NVTexcess} = \frac{1}{N!}\frac{1}{h^{3N}}Z_{N,NVT} \tag{1.7}$$

where $Z_{N,NVT}$ is the configurational integral. When dealing with the excess part it is common to drop out the first two terms and focus on the configurational integral. With the following simplifications, when one is interested in the ensemble average of a property that depends only on the coordinates, the momentum contributions in equation 1.4 can be safely ignored and the calculation of $\langle A_{ens}\rangle$ simplifies to:

$$\langle A_{ens}\rangle = \frac{\int dr^N A(r^N)\exp(-\beta U(r^N))}{Z_{N,NVT}} \tag{1.8}$$

Equation 1.8 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 postulates of statistical mechanics, this ensemble average is taken to be equal to the macroscopic

(thermodynamic) value of A, this equation provides us with a mean to derive thermodynamic properties *ab initio*.

1.4 Classical potentials

Central to the evaluation of any thermodynamic property A is the potential energy function $U(r^n)$ present in equation 1.8. The best available potentials belong undoubtedly to the realm of quantum mechanics.[31] Unfortunately, with present day computer technology they are impractical for the routine simulation of biomolecular systems. Heroic first principle simulations of liquid water have appeared in the literature,[32] but it is likely that quantum chemistry will not be the method of choice for simulations of large systems before a long time. Instead, so called classical force fields are often used to model the interactions between atoms. These methods relies on simple functional forms and sets of parameters empirically adjusted to reproduce the experimental or quantum chemical properties of molecules. There are many force fields that may be used to represent biomolecules, for example OPLS,[33] AMBER,[34] MM3[35] and CHARMM22.[36] Most of these force fields share similar functional forms and differ in their empirical parameters and the means used to derive them.

The functional form of the total potential energy, U_{total}, in the AMBER force field is as follows:

$$U_{total} = U_{bond} + U_{angle} + U_{dihedral} + U_{non-bonded}. \tag{1.9}$$

The bond and angle contributions are described by harmonic potentials and account for all the interactions between directly bonded (1-2) or directly angled (1-3) atoms:

$$U_{bond} = \sum_{bonds} K_b(r - r_{eq})^2 \tag{1.10}$$

$$U_{angle} = \sum_{angles} K_\theta(\theta - \theta_{eq})^2 \tag{1.11}$$

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

The torsional term is used to model interactions between pairs of (1-4) atoms and is computed as:

$$U_{dihedral} = \sum_{dihedrals} A_n(1+\cos(n\phi-\delta)), \tag{1.12}$$

where ϕ is the dihedral angle, n is the multiplicity (which gives the number of minimum points in the function as the torsion angle changes from 0 to 2π), δ is the phase angle and A_n is the force constant.

Finally, the non-bonded energy is composed of an electrostatic and a Lennard-Jones term:

$$U_{non-bonded} = \sum_i \sum_{j>i} \left\{ \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\}, \tag{1.13}$$

where the sum is over all atom pairs i, j. The q_i are the partial atomic charges, ε_{ij} and σ_{ij} are the Lennard-Jones well-depth energy and collision-diameter parameters, ε_0 is the permittivity of free space and r_{ij} is the inter-atomic distance. The non-bonded term is also applied to 1-4 atoms, but the magnitude of the interactions is reduced by adopting a scaling factor. For the AMBER all atom force field the coulombic interactions between 1-4 atoms are scaled by 0.833 and the Lennard-Jones interactions by 0.5.

The total energy of the system is taken as the sum over all inter- and intramolecular terms. Because the evaluation of the energy between all pairs of atoms can be time consuming in large system, the intermolecular terms are normally truncated such that interactions between atoms separated by more than a cut-off distance are ignored.[37,38] This cutoff 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 cutoff distance, then all of the pair-pair interactions between the two groups are ignored.[38] This truncation of the non-bonded terms can lead to discontinuities in the potential energies and forces associated with the interaction.[38] To overcome this problem, the non-bonded terms, $E^{nb}(r)$, may be scaled by multiplying by a *switching function*, $S(r)$,[38]

$$E^{nb'}(r) = S(r) \times E^{nb}(r), \tag{1.14}$$

where r is the distance between atoms. The aim of the switching function is to preserve the nature of the non-bonded interaction at low r, while gradually smoothing the energy to zero by the cutoff distance. This switching function may be applied over the entire range of distances, or only for a short range of distances before the cutoff,[38]

$$\begin{aligned} E^{nb'}(r) &= E^{nb}(r) \ \text{ for } \ r < r_{feather} \\ E^{nb'}(r) &= S(r) \times E^{nb}(r) \ \text{ for } \ r_{feather} \leq r \leq r_{cut} \\ E^{nb'}(r) &= 0 \ \text{ for } \ r > r_{cut}, \end{aligned} \tag{1.15}$$

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

When using group-based cutoffs, it is important to ensure that the switching function has the same value for each pair of atoms between the interacting groups.[38] This may be achieved by calculating a single value of the switching function for the interacting groups, and multiplying it by the total non-bonded interaction energy between the groups.

1.5 Sampling methods

Once a potential U deemed sufficiently accurate to reproduce closely the thermodynamic properties of interest is available, there remains the need to adopt a method to generate an ensemble of configurations that will be employed to estimate the configurational integral 1.8.

1.5.1 Metropolis Monte Carlo

The Metropolis Monte Carlo method was developed in 1953.[39] Behind its apparent simplicity lies many advanced mathematical concepts, and the curious reader, desirous of understanding why the Metropolis method really works is referred to appendix A. The algorithm is listed below.

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 $\chi = (\pi_j/\pi_i)$
4. If the move is accepted set i = j, otherwise i = i
5. Accumulate any property of interest A(i)
6. Return to 1 or terminate after a number of iterations

An important property that the algorithm must obey is the principle of detailed balance or microscopic reversibility (see appendix A).

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{1.16}$$

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

$$\begin{aligned} \pi_i Q_{ij} &= \pi_j Q_{ji} \\ \pi_i p_{ij} \alpha_{ij} &= \pi_j p_{ji} \alpha_{ji} \\ \pi_i p_{ij} \frac{\pi_j}{\pi_i} &= \pi_j p_{ji} \\ p_{ij} &= p_{ji} \end{aligned} \tag{1.17}$$

And we see that detailed balance is respected if 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.

Suppose we want to use Metropolis Sampling to sample from the Boltzmann distribution, then the acceptance test will be

$$\begin{aligned} \frac{\pi_{j,NVT}}{\pi_{i,NVT}} &= \frac{\exp(-\beta U_j)/Z_{N,NVT}}{\exp(-\beta U_i)/Z_{N,NVT}} \\ &= \exp(-\beta(U_j - U_i)) \end{aligned} \tag{1.18}$$

and it follows that we do not need to know the normalisation factor $Z_{N,NVT}$ which is fortunate as it is usually not possible to determine this parameter.

In a computer 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.

1.5.2 Monte Carlo moves

Standard Monte Carlo moves

Metropolis sampling often relies on the assumption of detailed balance. Perhaps the simplest way to require that the transition matrix probabilities p_{ij} and p_{ji} are equal is to select a trial state j randomly. Because the vast majority of the 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 formed by performing a small alteration to state i, the reasoning being that if state i is a member of the ensemble π, then a state j that is similar to state i has a reasonable probability to be part of that ensemble as well. In practice this is often done by picking randomly one particle in the system and performing a random translational/rotational displacement of that particle. If the particle has any internal degrees of freedom, then these degrees of freedom can be randomly modified as well. The magnitude of these modifications is often adjusted according to the moved particle and a rule of thumb is that the overall acceptance rate should be set to about 40 %. An important caveat is that alterations to the maximum range of the displacements should not be made while statistics are collected as this would violate detailed balance.

Biased moves

When applying the basic Metropolis method, we are often faced with situations where the probability of making a transition to an interesting state ($\pi_j > \pi_i$) is small. We accept the generated state with a probability $\alpha_{ij} = min(1,\chi)$. Because in this situation χ would be less than 1, the overall probability Q_{ij} is equal to p_{ij} which can be quite small. In that case, if a method to detect interesting transitions i to j can be formulated, it would be advantageous to introduce some bias to favour the transition but this usually means that the transition matrix will no longer be

symmetric ($p_{ij} \neq p_{ji}$).

Observe that a general solution to enforce detailed balance within the Metropolis algorithm is to set

$$\begin{aligned} \alpha_{ij} &= min(1,\chi) \\ \alpha_{ji} &= min(1,\frac{1}{\chi}) \\ \chi &= \frac{\pi_j p_{ji}}{\pi_i p_{ij}} \end{aligned} \tag{1.19}$$

In this case, the probabilities p_{ij} and p_{ji} enter in the acceptance test and have to be determined, but note it is now necessary to determine the bias on the reverse move in order to be able to perform the acceptance test. With some moves, it can be difficult to determine the probability of the reverse move and the introduction of bias in a Metropolis Monte Carlo scheme has to be made carefully.

As an illustration, consider the case of preferential sampling.[40] In this type of move, solvent molecules that are close to a solute of interest are moved more often than solvent molecules further apart. The probability that a solvent molecule i is picked is based on $W_i = \frac{1}{r_{is}^k}$ where r_{is} is the distance between the molecule i and the solute s and k is a parameter. The parameter W is normalized for every solvent molecule.

$$W_i^{'} = \frac{W_i}{\sum_{j=1}^{N} W_j} \tag{1.20}$$

A solvent molecule is then selected randomly from the N solvent molecules according to its weight W and displaced by a random amount. Detailed balance must be satisfied by the following relation:

$$\pi_i \frac{1}{N} W_i^{'} min(1,\chi) = \pi_j \frac{1}{N} W_j^{'} min(1,\frac{1}{\chi}) \tag{1.21}$$

which is solved for

$$\chi = \frac{\pi_j W_j^{'}}{\pi_i W_i^{'}} \tag{1.22}$$

The weight $W_j^{'}$ is not known when the move was initiated, but it is readily available once the solvent molecule has been displaced and the new distance r_{js} can be calculated. Other examples of biased moves encountered in molecular simulations are cavity biased insertion of molecules (for simulations in the grand canonical ensemble),[41] force biased moves[42] or configurational biased moves.[43]

Generalised ensembles

As mentioned previously, the potential energy surface of chemical systems of interest is often found to be very frustrated, with many minima and barriers that pose a challenge to standard sampling methods. The biased moves described in the previous section can be very effective to solve a sampling problem, but they can typically be applied only to specific systems. The method of Parallel Tempering[44–46] takes a different approach and increases sampling of the entire system by forming a generalised ensemble over temperature.[44] The method works by running a set of simulations of a given system at different temperatures. The individual simulations are also called replicas and the method is referred to as Replica Exchange. Periodically, Parallel Tempering moves are attempted between different replicas. If the move is accepted, the replicas exchange their temperature and the simulations proceed normally until the next attempted move. The Parallel Tempering acceptance test is designed such that each simulation is forming a correct NVT or NPT ensemble. For instance, in a NVT simulation a replica i at inverse temperature β_A should exchange with replica j at inverse temperature β_B with probability

$$\exp\left[\left(\beta_B-\beta_A\right)\left(E_B(j)-E_A(i)\right)\right] \geq rand(0,1), \tag{1.23}$$

With Parallel Tempering, a low temperature configuration can be taken into a high temperature simulation, undergo a large configurational change and then 'cool down' back to its 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. Another drawback is the necessity to run several simulations at high temperature when only one ensemble at room temperature may be of practical interest.

1.5.3 Molecular dynamics

Another commonly used method to generate ensemble of thermally relevant states is Molecular Dynamics (MD).[37] In MD simulations, the time evolution of a system of a N atoms placed in a starting configuration is monitored. The forces acting on that system can be calculated with the knowledge of its potential energy function and Newton's laws of motion. Because of the deterministic nature of Newton's laws, this information is sufficient to generate a trajectory of the simulated system over time. Analytical solutions are not practical as they would requires the solution of 3N coupled, second order differential equations. Fortunately, many numerical approaches permit the repeated integration of the forces over small time intervals to yield a trajectory. Because the total energy of the system is conserved by the application of Newton's laws, MD simulations naturally form the NVE ensemble. Algorithms that connect the system to a thermostat or barostat allow the sampling of the NVT or NPT ensemble. 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 simulations).

While both approaches should 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.[47] In a MD simulation, all the degrees of freedom of the system are subject to forces and hence move. It is often necessary to constrain many degrees of freedom using algorithms such as the SHAKE method.[48] 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.

1.6 Rigorous free energy calculation methods

1.6.1 The importance of the free energy

The free energy governs many important thermodynamic phenomena. It is the driving force behind chemical processes. The ability to predict a free energy gives the ability to predict solvation, binding, stability, phase transitions and many other properties. The free energy of a system is directly related to its partition function. In the canonical ensemble the expression for the Helmholtz free energy is simply:

$$A = -k_B T \ln Q_{NVT}. \tag{1.24}$$

In the isothermal-isobaric ensemble, the quantity on the l.h.s of equation 1.24 is the Gibbs free energy. At thermodynamic equilibrium, the free energy is minimised.[29] By measuring the absolute free energy of two comparable systems, it is possible to determine which is the more favoured.

1.6.2 Absolute free energy calculation

For simplicity, we ignore the contribution of the ideal part to the partition function which may be analytically evaluated[30] and hence write the following for the excess free energy:

$$\begin{aligned} A &= -\frac{1}{\beta}\ln Q_{NVT} \\ &= \frac{1}{\beta}\ln(1/Q_{NVT}) \\ &= \frac{1}{\beta}\ln\frac{N!h^{3N}}{\int \exp\Big(-\beta U(r^N)\Big)dr^N} \end{aligned} \tag{1.25}$$

Now we can write

$$1 = \frac{1}{(8\pi^2 V)^N}\int \exp\Big(+\beta U(r^N)\Big)\exp\Big(-\beta U(r^N)\Big)dr^N \tag{1.26}$$

The constant factor in equation 1.26 arise from the integration of 1 over phase space.

Equation 1.26 is inserted into 1.25

$$\begin{aligned} A &= \frac{1}{\beta}\ln\frac{N!h^{3N}}{(8\pi^2V)^N}\frac{\int\exp\Big(+\beta U(r^N)\Big)\exp\Big(-\beta U(r^N)\Big)}{\int\exp\Big(-\beta U(r^N)\Big)}dr^N \\ &= \frac{1}{\beta}\ln\frac{N!h^{3N}}{(8\pi^2V)^N}\int\exp\Big(+\beta U(r^N)\Big)\pi(r^N)dr^N \\ &= \frac{1}{\beta}\ln\frac{N!h^{3N}}{(8\pi^2V)^N}\Big\langle\exp\Big(+\beta U(r^N)\Big)\Big\rangle \end{aligned} \tag{1.27}$$

Equation 1.27 shows that the free energy of a system can be calculated as an ensemble average. The constant factor can be difficult to calculate since it would require the definition of the volume of phase space V. If one is interested in the difference in absolute free energy between two comparable systems, it can be ignored as it only acts to shift the value of the absolute free energy by a constant offset. Unfortunately, when subjected to the techniques developed previously, equation 1.27 exhibits a very poor convergence behaviour. This is because Metropolis Monte Carlo samples states according to the Boltzmann distribution which generates mostly states of low energy. States of high energy are rarely encountered, yet they make large contributions to the ensemble average because of the sign of the exponential. [2]

Because in chemical systems, there are so many high energy states (often corresponding to atomic overlaps) a direct estimation of the free energy A by equation 1.27 is a hopeless task. Furthermore, one could question the merit of this approach. The vast majority of thermodynamic experiments relies on the measurements of equilibrium constants which can be directly related to a *change* of free energy. The author is not familiar with any technique that would measure an absolute free energy. This leads us to the next section which discuss calculation of free energy *differences*.

[2] That high energy states corresponding to configurations of vanishingly small Boltzmann factor make an increasingly large contribution to this ensemble average is a puzzling mathematical fact that begs to be reconciled with a physical interpretation.

1.6.3 Free energy perturbation

Although a direct approach to the calculation of free energies is impractical, Zwanzig has shown that it is possible to calculate the relative free energy of two different systems A and B.[49]

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

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

$$
\begin{aligned}
&= -\frac{1}{\beta}\ln\left[\frac{\int \exp(-\beta U_B(r^N)) \times \exp(-\beta U_A(r^N))\exp(\beta U_A(r^N))\,dr^N}{\int \exp(-\beta U_A(r^N))\,dr^N}\right] \\
&= -\frac{1}{\beta}\ln\left[\frac{\int \exp(-\beta U_A(r^N)) \times \exp\big(-\beta(U_B(r^N) - U_A(r^N)\big)\,dr^N}{\int \exp(-\beta U_A(r^N))\,dr^N}\right] \\
&= -\frac{1}{\beta}\ln\left[\int \frac{\exp(-\beta U_A(r^N))}{Q_A} \times \exp(-\beta \Delta U_{AB}(r^N))\,dr^N\right] \\
&= -\frac{1}{\beta}\ln\left[\int \pi_A(r^N) \times \exp(-\beta \Delta U_{AB}(r^N))\,dr^N\right] \\
&= -\frac{1}{\beta}\ln\big\langle \exp(-\beta \Delta U_{AB}(r^N))\big\rangle_A
\end{aligned}
\tag{1.28}
$$

This equation shows that the relative free energy is the logarithm of the ensemble average of the exponential of the Boltzmann weighted energy difference between the potentials U_A and U_B. In computer simulations, the Zwanzig equation is implemented using the Free Energy Perturbation methodology.[37] A simulation is performed with the potential U_A and at each step i of the Markov chain the quantity $\exp(-\Delta U_{AB}(i)/k_B T)$ is accumulated. The approach sounds simple but in practice there are several pitfalls that must be avoided in order to obtain reliable

results.

One difficulty encountered in applying equation 1.28 to calculate free energy differences is that the potential energy function of systems B and A can be too different. If the low energy regions of B are in portions of phase space corresponding to high energy regions of A, then a simulation run with potential U_A will rarely generate the significant configurations of potential U_B. As a result, the free energy change $\Delta G_{A\rightarrow B}$ is likely to be overestimated. The same situation arises if the potentials are switched and $\Delta G_{B\rightarrow A}$ will be overestimated. Any difference between these two quantities is known as hysteresis. If the hysteresis is large, the calculated free energies will be a poor approximation of the actual quantity.

A simple solution is to multi-stage the calculation. A series of intermediate potentials $U_{P(\lambda)}$ are defined, where $U_{P(0)} = U_A$ and $U_{P(1)} = U_B$. One can then connect state A and B by a set of more similar states and eq 1.28 can be rewritten as a sum of energy differences.

$$G_B - G_A = \Delta G = \sum_{\lambda=0}^{1} -k_B T \ln < \exp(-\Delta U'/k_B T) >_{\lambda_k} \tag{1.29}$$

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

1.6.4 Thermodynamic integration

Thermodynamic Integration (TI) is another established rigorous free energy method.[38] Instead of summing free energy differences between neighbouring values of λ, a set of simulations are run at different λ values. The free energy gradient $(\frac{\partial G}{\partial \lambda})_\lambda$ is estimated at each of these λ values. Once all the free energy gradients are known, they may be integrated to yield the relative free energy change along the λ coordinate.

$$G_{\lambda=1} - G_{\lambda=0} = \int_0^1 \left(\frac{\partial G}{\partial \lambda}\right)_\lambda d\lambda \tag{1.30}$$

This integral can be evaluated numerically, e.g. via the trapezium rule.[38] The free energy gradients themselves may be obtained analytically or numerically. The

ensemble average of the gradient of force field, $\left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda$, is equal to the free energy gradient.[30]

$$\int_0^1 \left(\frac{\partial G}{\partial \lambda} \right)_\lambda d\lambda = \int_0^1 \left\langle \frac{\partial U}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{1.31}$$

The ensemble average of the gradient of the force field can be evaluated by calculating the gradient of each force field term directly with respect to λ. An alternative numerical route is to approximate the gradient, $(\frac{\partial G}{\partial \lambda})_\lambda$, via the finite difference, $(\frac{\Delta G}{\Delta \lambda})_\lambda$.

This free energy difference can be calculated via the Zwanzig equation, with the reference state at λ, and the perturbed state at $\lambda + \Delta\lambda$. This would give a forwards estimate of the free energy gradient. A perturbed state of $\lambda - \Delta\lambda$ yields the backwards estimate. These two estimates should of course be equal if $\Delta\lambda$ were sufficiently small, and the trajectory ran until the Zwanzig equation had converged. This method is normally referred to as Finite Difference Thermodynamic Integration[50] (FDTI).

Over the last decade, several studies of protein-ligand complexes have been published, suggesting that the free energy perturbation or thermodynamic integration methodologies can yield results in good agreement with experimental measurements of binding affinities of protein-ligand complexes.[28,51–58]

1.6.5 Replica exchange thermodynamic integration

A recent interesting development in free energy calculation methods, inspired by generalised ensemble methods, is the Replica Exchange Thermodynamic Integration method (RETI).[59,60] RETI forms a generalised ensemble over the coupling parameter λ which connects two different Hamiltonians in a free energy simulation. To conducts a RETI simulation, a set of replicas that covers the 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. A suitable acceptance test is:

$$\exp\left[\beta\big(E_B(j) - E_B(i) - E_A(j) + E_A(i)\big)\right] \geq rand(0,1). \tag{1.32}$$

Unlike Parallel Tempering, a RETI simulation can be performed at no extra cost since all the simulations are already needed in standard simulations. Neigbouring replicas tends to exchange with higher probabilities than in PT simulations as the systems tends to be more similar over a change of λ than over a change of temperature. RETI provides enhanced sampling as it allows individual trajectories to jump to distantly related configurations in phase space. In favorable cases, it can allow some replicas to overcome barriers by 'dodging it'. This happens when a replica at λ_i exchanges with another replica running at a λ_j value which does not experience this barrier, performs some local sampling and then exchanges back into the original λ_i value in a region that lies on the 'other side' of the barrier. If every λ value experience a similar high barrier, then the quality of the sampling will not be improved much over standard methods. Calculations of the relative solvation energies of water and methane in water and the binding energy of halides to a Calix[4]pyrrole system have been reported in the literature.[59,60] In each case, RETI performed better than established free energy methods.

1.6.6 One-step multiple perturbed states

A drawback to established free energy simulation protocols is that they require several simulations at different values of the coupling parameter λ to yield a single free energy difference. In the free energy perturbation methodology the relative free energy difference is estimated by perturbing a reference state into the state of interest. It has been proposed by Schafer et al. that a simulation could be run in which the reference state is perturbed into several different states.[61] This allow the simultaneous calculation of the relative binding free energy of several species, and because intermediates values of the coupling parameter λ are not simulated, the methodology can quickly become 10-100 times faster than conventional FEP or TI. In this situation, it is important that the reference state is similar to the perturbed states for the free energy differences to be readily converged. This constraint is relaxed by adopting a non chemical reference state that is designed to have good overlap with the series of perturbed states. This is typically done in conjunction with the use of a softened non-bonded interaction function.[62]

The methodology has been applied with some success to the calculation of a series of structurally similar[63] or different compounds binding to the ligand-binding domain of the estrogen receptor.[64] In a similar study applied to ligands binding to the protein Xa, it was found that the design of an appropriate reference state was difficult.[65] The accurate reproduction of solvation free energies with this methodology required a particular "soft-dipole" energy function.[66]

In general it seems that the one step multiple perturbed states can be very efficient, providing a suitable reference state that will sample all the low energy configurations of all the considered ligands, can be devised *a priori*. In complex systems, typical of the condensed phase, this can prove difficult.

1.6.7 Slow and fast growth

An established free energy method that has been in used for some time is the so called 'slow growth' method. In a slow growth method, the value of λ is slowly increased by a constant amount after a number of MC moves or MD time-steps such that at the start of the simulation, $\lambda = \lambda_0$, and by the end of the simulation, $\lambda = \lambda_1$. If the simulation consists of M steps, then $\delta\lambda$ is given by,[67]

$$\delta\lambda = \frac{\lambda_1 - \lambda_0}{M}. \tag{1.33}$$

Because the system is constantly perturbed, the gradual increase of λ requires work, which can be calculated as:

$$W = \sum_{i=1}^{M} \delta\lambda \left(\frac{\partial E}{\partial \lambda} \right)_{\lambda = \lambda_0 + i\delta\lambda}. \tag{1.34}$$

An advantage of the slow growth approach is that a single simulation is required to obtain a work value. In the limit of an infinitesimal increase of λ, the system would be in thermodynamic equilibrium throughout the whole simulation and the work value obtained could be related directly to an equilibrium free energy difference. However, because the perturbation of the coupling parameter and the number of steps between two subsequent increase of λ is necessary finite in a

computer simulation, the computed work value will always be larger than the free energy change.[38]

$$W \geq \Delta G. \tag{1.35}$$

An interesting elaboration on the slow growth approach is the 'fast growth' method proposed by Jarzinsky[68,69] whose essence is conveyed by equation 1.36:

$$\overline{\exp(-W/k_BT)} = \exp(-\Delta G/k_BT). \tag{1.36}$$

In the equality 1.36, the change in free energy ΔG of a perturbation is shown to be related to the average of the work values W calculated for several fast growth simulations. Crucially, the equality is independent of the rate at which the coupling parameter λ is increased. This means that an equilibrium thermodynamic property can be derived by averaging the work values of several non equilibrium simulations. Because λ is typically increased much more quickly than in a slow-growth simulation, the method has been named fast growth.[70]

The validity of the fast growth method has been initially tested on Lennard Jones fluid,[70] the potential of mean force between a pair of methane molecules in water,[71] the charging of a sodium ion in water,[72] and tested experimentally.[73] Refinements to the fast growth method have proposed recently, aimed mainly at reducing the fluctuations in the distributions of work values to improve the rate of convergence of the average work value.[74,75]

The main potential of the fast growth method seems to lie in its trivial parallelization which would make it very suitable to modern GRID enabled computing technologies. The author is not aware of published work concerning the application of the fast growth method to protein ligand binding free energy calculations.

1.7 Calculating errors in free energy simulations

When using Monte Carlo or molecular dynamics methods to generate configurations and obtain the density of states of the simulated system, errors are introduced because of the necessary finite number of configurations. The ensemble average of a property $< A >$ is usually said to be converged if if does not change significantly

when the number of configurations used to determine it is increased. A useful way to assert such a proposition is to subdivide a simulation of N configurations into K blocks of N/K configurations, calculate $< A >_K$ for each block and then the standard deviation from that distribution of values. In principle, if all the portions of phase space that contribute significantly to $< A >$ have been visited with the right probability in each block, all the values will be similar and the standard deviation low. This method suffers from two important difficulties.[37]

First, the blocks must be long enough to be completely statistically uncorrelated with each other. Monte Carlo or molecular dynamics 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. Second a low standard deviation guarantees by no mean that simulation results have converged to the right answer. If the system is unable to climb local barriers, the simulation may explore thoroughly one local minimum and miss out completely other important regions of phase space. A block analysis will suggest the results are (incorrectly) converged.

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 (e.g, simulate at very high temperature and cool down the system). This method has the obvious drawback that one has now to run several simulations instead of one.

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. Figure 1.1 highlights the 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. The extent by which the cycle closure deviates from this figure is a measure of the lack of convergence.

Unfortunately, none of the approaches discussed here guarantees that the simulation results are converged. It appears 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.

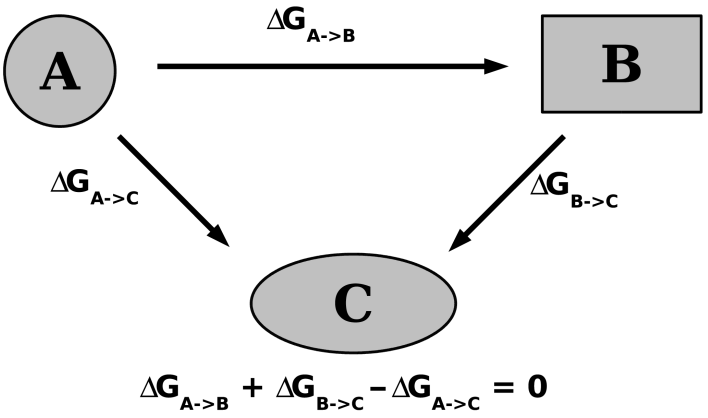


Figure 1.1: The closure of a thermodynamic cycle. While only two simulations are necessary for the calculation of the relative free energy of system B or C with respect to system A, a third simulation that calculates the relative free energy of system C with respect to B allows the closure of a cycle involving A,B and C. Deviations from the theoretical result of 0 are a measure of errors.

1.8 Approximate Free Energy Methods

The methods discussed in the previous sections are rigorous in the sense that they yield free energies in a statistical mechanical sense. Their application can be quite time consuming however and this has prompted the development of several methods that try to approximate the calculation of a binding free energy, of which two such methods will be discussed.

1.8.1 Linear interaction energy

In the Linear Interaction Energy method (LIE)[76] the absolute binding free energy of a ligand to a protein is estimated by running two independent simulations. One is the ligand free in solution and the other is the solvated protein-ligand complex. The absolute binding free energy, ΔG, is then estimated from the simulation average of the difference in electrostatic energy between the ligand and the environment in the two simulations, $\langle \Delta U_{elec} \rangle$, and a similar average for the van der Waals ligand-environment energy, $\langle \Delta U_{vdw} \rangle$,

$$\Delta G = 0.5 \langle \Delta U_{elec} \rangle + \alpha \langle \Delta U_{vdw} \rangle. \tag{1.37}$$

There are some theoretical reasons that can justify the factor 0.5 for the Couloumbic term and they can be found in the Linear response approximation.[76,77] The value of the parameter α was originally determined by a fit to the experimental binding free energy of a series of endothiapepsin inhibitors and a value of 0.161 was thus derived. In subsequent work from different research groups it was shown that in general good results can be obtained but that no universal set of parameters can explain the binding free energies of different protein ligand complexes.[78–83] Furthermore, the factor 0.5 for the coulombic term did not give good agreement with experiment in many cases. As a result it has been proposed to add new terms to equation 1.37 depending on the system under study and to derive appropriate weighting factors empirically. Modern LIE studies can typically be conducted when there is enough experimental data on protein-ligand binding free energies such that sufficiently robust parameters for that system can be derived for an equation derived from 1.37. In that respect, LIE appears more as a knowledge based method than a true *ab initio* free energy prediction method.

1.8.2 MM/PBSA

Another approximate binding free energy calculation method that has received much attention is the Molecular Mechanics/Poisson Boltzmann Surface Area method (MM/PBSA).[84]

$$\Delta G_{bind} = \left\langle \Delta E_{mm} \right\rangle + \Delta G_{solv} - T\Delta S. \tag{1.38}$$

Where $< \Delta E_{mm} >$ is a difference in molecular mechanics energy, ΔG_{solv} is a solvation free energy, T the temperature and ΔS a change in entropy.

In this approach, a series of molecular dynamics trajectory for the solvated protein-ligand complex and the protein and ligand in solution is first generated. The first term in equation 1.38 is taken as the difference in the average molecular mechanics energies between the complex and isolated protein and ligand. A number of snapshots (typically 50-200) are extracted from these molecular dynamics trajectories and subjected to a Poisson Boltzmann Surface Area (PB/SA) calculation. The ligand and protein are then taken apart and their molecular mechanics energies are calculated. The difference in the resulting average energies yield the

second term of equation 1.38. The third term reflects the change in entropy upon binding and is calculated from one snapshot, using normal mode analysis.

To date the MM/PBSA method has been applied to several protein-ligand systems.[85–90] While impressive results have been reported in some cases, large scale validation studies tend to indicate that MM/PBSA is accurate to within 2 to 3 kcal mol^{-1} only.[89] Pearlman has shown that, on a series of congeneric p38 MAP kinase inhibitors, the MM/PBSA was very inaccurate and no better than some common scoring functions.[91] Woo et al. reported MM/PBSA results in error by more than 70 kcal mol^{-1} in the calculation of the absolute binding free energy of a phosphotyrosine peptide pYEEI to the Src homology 2 domain of human Lck.[92]

There are a number of reasons why the MM/PBSA methodology can be expected to fail in some cases. First, the distribution of snapshots taken from the molecular dynamics simulation is small and the resulting 'ensemble' may not reflect accurately the true Boltzmann distribution. Second, the change in solvation free energy is estimated by using these same snapshots and yet one should expect a different distribution of states between an explicit and implicit representation of the solvent. Third, the first two terms of equation 1.38 are taken as the difference of very large numbers (the molecular mechanics energy of a protein ligand system is usually in the order of a few thousand kcal mol^{-1}) with significant fluctuations. Obtaining a binding free energy in the order of minus one to ten kcal mol^{-1} from this protocol can be challenging. Fourth, the last term in equation 1.38 is a poor approximation of the entropy loss, problematic to calculate in many instances, and often simply ignored.

1.9 Continuum solvation

Many interesting biomolecules performs their functions in an aqueous environment. Realistic computer simulations therefore have to consider the effects of the solvent on the solute structure and dynamics. A perhaps obvious approach is to represent the surrounding solvent with a large number of molecules, each interacting according to a potential defined in the previous section. Unfortunately, traditional simulations must include thousands of water molecules to solvate properly

a protein. Complicated methods such as periodic boundary conditions or Ewald summation are also needed to avoid artificial boundary effects.[38] Thus, it is not uncommon that during the simulation, most of the time is spent computing non bonded interactions for solvent molecules, which are usually not the prime interest of the simulation.

A serious alternative to these explicit solvent simulation is to consider the solvent as a high-dielectric continuum interacting with charges that are embedded in solute molecules of lower dielectric. The solute response to the reaction field of the solvent dielectric can then be modelled by applying laws of classical electrostatics.

1. Thousands of solvent molecules do not have to be modelled explicitly, reducing the complexity of the system and the CPU cost.

2. In an explicit solvent simulation, after a solute move, several solvent moves may be needed to reorganise the surrouding solvent molecules while no such problem is observed in implicit solvent simulations. Furthermore, the presence of the explicit solvent molecules may render large conformational changes of the solute much more difficult.

The Poisson Boltzmann (PB) equation is one of the most accurate ways to model these electrostatic interactions.[93] Analytical solutions of the PB equation for solutes of arbitrary shape are not available and are usually obtained by finite-difference or boundary-element numerical methods. Solving the PB equation is quite expensive for large molecules and other more efficient and approximate methods have been proposed.

One of these methods is the generalised Born (GB) approach which we have adopted in our work and is presented in the next sections.

1.9.1 The Born equation

Born has shown[94] that an analytical equation for the electrostatic energy of an isolated ion can be derived from classical electrostatic theory.

Classical electrostatic theory states[95] that the total electrostatic energy in a dielectric medium is defined as:

$$G = \frac{1}{8\pi}\int_V \vec{E}.\vec{D}.dV \tag{1.39}$$

$$\vec{D} = \varepsilon.\vec{E} \tag{1.40}$$

$\vec{E}$ and $\vec{D}$ are the electric field and electric displacement, ε is the dielectric constant of the medium and dV a volume element. $\vec{E}$ can be obtained from Gauss Law:

$$\int_S \vec{E}.d\vec{S} = \frac{Q}{\varepsilon} \tag{1.41}$$

The surface integral on the left is the area integral over any closed surface. Q is the total charge that lies within the space delimited by $\vec{S}$.

For a uniformly charged spherical shell of radius α and interior dielectric ε_{vac} inside and outside, one can obtain:

$$\vec{E}_{int} = 0 \qquad \vec{D}_{int} = 0 \qquad r < \alpha \tag{1.42}$$

$$\vec{E}_{out} = \frac{kq}{\varepsilon_{vac} r^3}.\vec{r} \qquad \vec{D}_{out} = \frac{kq}{r^3}.\vec{r} \qquad r > \alpha \tag{1.43}$$

With q the total charge of the sphere and k the Coulomb constant ($1/(4\pi\varepsilon_0)$).

The total electrostatic energy of the system is:

$$G_{vac} = \frac{1}{8\pi}\int_V \vec{E}.\vec{D}.dV \tag{1.44}$$

$$= \frac{1}{8\pi}[\int_{in} \vec{E}_{in}.\vec{D}_{in}dV + \int_{out} \vec{E}_{out}.\vec{D}_{out}dV] \tag{1.45}$$

$$= \frac{k^2}{8\pi\varepsilon_{vac}}\int_{out} \frac{q^2}{r^4}dV \tag{1.46}$$

By integrating 1.46 from α to ∞ we find:

$$G_{vac} = \frac{q^2k^2}{2\varepsilon_{vac}\alpha} \tag{1.47}$$

If the same spherical system is now considered in a dielectric medium with an interior dielectric of ε_i and and exterior dielectric constant of ε_{solv}, the total electrostatic energy can similarly be shown to be:

$$G_{solv} = \frac{q^2 k^2}{2\varepsilon_{solv}\alpha} \tag{1.48}$$

The electrostatic energy to transfer a spherical charged ion of radius α from a medium of dielectric ε_{vac} to another of dielectric ε_{solv} is the difference between 1.48 and 1.47. This is the Born Equation[94]

$$\Delta G_{born} = \frac{k^2}{2}\left(\frac{1}{\varepsilon_{solv}} - \frac{1}{\varepsilon_{vac}}\right)\frac{q^2}{\alpha} \tag{1.49}$$

1.9.2 Electrostatics and the generalised Born model

The Born model of solvation can be generalised to a molecule of arbitrary shape by treating each atom as a sphere of radius α_i, a charge q_i and interior dielectric ε_i.

If we assume initially that each sphere is separated by a distance large enough so that they appear as point charges to other spheres (e.g single dielectric medium), then the total electrostatic energy of the system is the sum of the Coulombic interaction and the Born solvation energy.

$$G_{tot} = \frac{1}{2}\sum_i \sum_{j\neq i} \frac{q_i q_j}{\varepsilon_{solv} r_{ij}} - \frac{1}{2}\left(\frac{1}{\varepsilon_{vac}} - \frac{1}{\varepsilon_{solv}}\right)\sum_i \frac{q_i^2}{\alpha_i} \tag{1.50}$$

Unfortunately, equation 1.50 is not valid for molecular systems where the radius α_i and the distance r_{ij} are usually too close for the former to be negligible.

Still[96] has shown that by splitting the Coulombic interaction into two terms, one can write equation 1.51

$$G_{tot} = \frac{1}{2}\sum_i \sum_{j\neq i} \frac{q_i q_j}{\varepsilon_{vac} r_{ij}} - \frac{1}{2}\left(\frac{1}{\varepsilon_{vac}} - \frac{1}{\varepsilon_{solv}}\right)\sum_i \sum_{j\neq i} \frac{q_i q_j}{r_{ij}} - \frac{1}{2}\left(\frac{1}{\varepsilon_{vac}} - \frac{1}{\varepsilon_{solv}}\right)\sum_i \frac{q_i^2}{\alpha_i} \tag{1.51}$$

Terms 2 and 3 can then be recombined in a single formula:

$$\Delta G_{tot} = \frac{1}{2}\sum_i \sum_{j \neq i} \frac{q_i q_j}{\varepsilon_{vac} r_{ij}} + \Delta G_{Genborn} \tag{1.52}$$

Where $\Delta G_{Genborn}$ is:

$$\Delta G_{Genborn} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{vac}} - \frac{1}{\varepsilon_{solv}}\right)\sum_i \sum_j \frac{q_i q_j}{\sqrt{r_{ij}^2 + B_i B_j e^{\frac{-r_{ij}^2}{4B_i B_j}}}} \tag{1.53}$$

Where B_i and B_j are Born radii, similar to the quantity α_i in equation 1.49.

This expression reduces to the Born equation for the case of a single spherical ion and gives the Coulomb energy as $r_{ij} \rightarrow \infty$.

It is important to realise that equation 1.53 has no physical basis. It results effectively from an interpolation between different theoretical results: the Born equation, the Onsager dipole energy equation and the Coulomb equation for widely separated charges.

Much of the difficulty with 1.53 consists in computing the Born radii B_i. The Born radius of one atom is affected by neighbouring atoms and is no longer equal to α_i. In the generalised Born formalism, B_i is defined as, the radius that would give the actual electrostatic energy of the molecule-dielectric system by the Born equation if all other atoms of the system were uncharged (only displacing the dielectric). This corresponds to defining a spherically averaged dielectric boundary for atom i (the angular dependence is not taken into account). The evaluation of the integral itself is not straightforward as it depends on the position of all other atoms of the solute with respect to the solvent/solute boundary. B_i can be derived by the Poisson equation but this nullifies the advantage of the model.

In the original paper from Still,[96] the Born radii are computed using a numerical method which can be summarised as:

1. Consider a shell of thickness T_k surrounding the van der Waals surface of atom k.

2. Weight the interior radius (r_k - 0.5T_k) of this shell using the ratio of solvent accessible surface area A_k to the actual surface area.

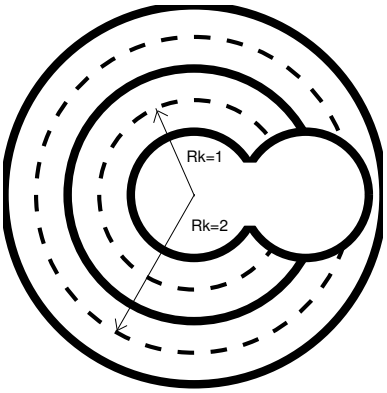


Figure 1.2: Generation of consecutive shells by equation 1.54

3. Repeat the weight for the exterior radius (r_k + $0.5T_k$) and calculate the difference between weighted interior and exterior radii.

4. Sum the difference between weighted interior and exterior radii for a series of concentric shells up to shell M which encompasses the whole of the van der Waals surface of the molecule.

5. For shell M no weight is applied and the radius is simply added to the previous summation term, to obtain an effective Born radius, which is then used in equation 1.53 .

The method is illustrated by figure 1.2. A formal description of this algorithm is given by equation.1.54

$$\frac{1}{B_i} = \sum_{k=1}^{M} \frac{A_k}{4\pi r_k^2} \left[\left(\frac{1}{r_k - 0.5T_k} \right) - \left(\frac{1}{r_k + 0.5T_k} \right) \right] + \frac{1}{r_{M+1} - 0.5T_{M+1}} \qquad (1.54)$$

Because this method is time consuming, analytical approaches that are approximate have also been developed.[97,98] It is hoped that most of the errors arising from these inexact solutions are systematic and can be corrected by empirical terms. In our work we have used the Pairwise Descreening Approximation developed by Hawkins et al[97] to compute Born radii (details of the derivation are not shown).

$$B_i^{-1} = \alpha_i^{-1} - \frac{1}{2}\sum_{j\neq i}[\frac{1}{L_{ij}} - \frac{1}{U_{ij}} + \frac{R_{ij}}{4}(\frac{1}{U_{ij}^2} - \frac{1}{L_{ij}^2}) + \frac{1}{2R_{ij}}ln\frac{L_{ij}}{U_{ij}} + \frac{S_{ij}^2\alpha_j^2}{4R_{ij}}(\frac{1}{L_{ij}^2} - \frac{1}{U_{ij}^2})] \tag{1.55}$$

$$L_{ij} = 1 \text{ if } R_{ij} + S_{ij}\alpha_j \leq \alpha_i$$

$$L_{ij} = \alpha_i \text{ if } R_{ij} - S_{ij}\alpha_j \leq \alpha_i < R_{ij} + S_{ij}\alpha_j$$

$$L_{ij} = R_{ij} - \alpha_j \text{ if } \alpha_i \leq R_{ij} - S_{ij}\alpha_j$$

$$U_{ij} = 1 \text{ if } R_{ij} + S_{ij}\alpha_j \leq \alpha_j$$

$$U_{ij} = R_{ij} + S_{ij}\alpha_j \text{ if } \alpha i < R_{ij} + S_{ij}\alpha_j$$

R_{ij} is the distance between two spheres centred on atoms, α_i the intrinsic Born radius of atom i (or otherwise, the Born radius that would give its solvation energy if it was alone) and S_{ij} a screening factor that scale the Born radius of atom j. This factor was introduced to correct systematic errors in the PDA approximation, because PDA over-estimates the Born radius by not taking into account the fact that two atomic spheres, j, j' can overlap. In this case S_{ij} accounts for the overlapping region of dielectric being displaced twice. This means that scaling factors should have a value between 0 and 1.

The GB equations, or PB for that matter do not tell the whole story of solvation. It is also necessary to take into account other effects that are described in the next section.

1.9.3 The apolar component of solvation

Solvation is not entirely determined by the distribution of charges of the solute inside a cavity.

To insert a solute inside solvent, a cavity the size of the solute has to be created. In the case of water, the hydrogen bonding network is disrupted and solvent molecules have to reorganise and reorient around the solute.

van der Waals forces also play a role as solute atoms are able to establish interactions with solvent atoms. Solvent atoms lie usually far enough away from solute atoms for the van der Waals forces to be predominantly attractive.

In most continuum solvation models, both effects are taken into account with a single dependence on the solvent accessible surface area (SASA) of the solute.[96,99,100]

$$G_{nonpol} = G_{cav} + G_{vdW} = \sum_{k=1}^{N} \sigma_k . SASA_k \tag{1.56}$$

The SASA is the surface that defines the region of space that solvent is excluded from upon insertion of the solute. For that purpose it is assumed that water can be represented as a sphere of radius 1.4 Å , and the SASA is defined by rolling that sphere over the van der Waals surface of the solute. The set of coefficients σ_k are usually empirically derived.

Using the SASA of a solute to model solvation has the drawback that buried atoms do not interact at all with the solvent, which is not the case in explicit solvent simulations.[101] It is also known that equation 1.56 cannot explain the change in solvation free energy for a series a linear alkanes and the solvation free energies of different rotamers of butane or hexane.[102]

Some workers have proposed in the recent years a more complex treatment of the non polar part of solvation by explicitly considering cavitation and dispersive forces separately.[103,104] The methodology has been employed to study the free energy surface of small peptides and appear to yield superior results.[105]

1.10 Conclusion and outline of the thesis

The routine prediction of protein-ligand binding free energies by computer simulation would provide the pharmaceutical industry with a very powerful tool to develop better drugs efficiently. Current routinely employed technologies are based on simplistic empirical functions that are known to suffer from several flaws and are not judged reliable enough to discriminate between different potent binders. By constructing an atomistic model of protein-ligand interactions, the free energy of binding can be derived using the laws of statistical mechanics. Practical solutions of the resulting integral can be obtained by generating ensemble of states representative of the protein-ligand interaction, and this is traditionally achieved by Monte

Carlo or molecular dynamics methods. That collection of states is a function of the potential energy function which mediates interactions between all of the simulated particles. This integral can be further simplified by recasting the absolute free energy calculation problem in the calculation of relative binding free energies, which requires the introduction of a parameter λ to smoothly transform one ligand into another. Even then, the resulting ensemble averages are hard to evaluate efficiently and non rigorous methods try to approximate a binding free energy with simpler approaches. These methods suffer from numerous approximations which limits their accuracy and range of application.

There are several reasons why free energy calculations are not routinely employed by the pharmaceutical industry to predict binding free energies. One of is their high computational cost which drastically limit the number of compounds that can be tested. A second reason is that it is often difficult to consider structurally dissimilar compounds in a single free energy study. This limit the range of systems that can be studied with the existing methodologies.

The main hypothesis that prompted this work can be summarised as follow: *free energy calculations could be made more efficient, and yet accurate, by adopting a simplified treatment of solvation*.

The validity of this assertion is tested in the following chapters. Chapter 2 will discuss a novel parameterisation of a generalised Born model of water, with the aims of adopting this model in protein ligand binding free energy calculations. In chapter 3, algorithmic improvements aimed at increasing the efficiency of generalised Born simulations will be introduced. Validation of the technique is then performed by completing two protein ligand binding free energy studies and the results are reported in chapter 4 and 5. With conclusive answers to the main objective of this work provided, alternative methods aimed at improving the range of molecules that can be studied with free energy techniques will be then considered in chapter 6. The lessons learned throughout this thesis will finally be summarised in chapter 7 to gauge the success of this project.

Chapter 2

Parameterisation and validation of a generalised Born surface area model of water

> "I recognize that many physicists are smarter than I am–most of them theoretical physicists. A lot of smart people have gone into theoretical physics, therefore the field is extremely competitive. I console myself with the thought that although they may be smarter and may be deeper thinkers than I am, I have broader interests than they have."
>
> Linus Pauling

2.1 Introduction

Traditional free energy studies are often faced with two major obstacles. Typical molecular force fields have been developed to simulate small organic or biomolecules. While they provide a reasonable set of parameters for proteins or DNA, it is common that they lack parameters to describe completely the complex organic molecules that are typical of drugs. Recipes that allow the derivation of these missing parameters are known, but their application can be very time-consuming. A major aim of this thesis is to provide a free energy method that allows the simulation of a large number of protein-ligand complexes in a few hours. The ability to

simulate quickly such systems is of little use if it is necessary to previously spend weeks deriving missing force field parameters.

Representing solvent effects in a free energy simulation is also a source of problems. While the most proven method is to consider explicitly a large number of solvent molecules surrounding the solute, it is also the most expensive way to address this issue.

The goal of this chapter is to introduce methods that deal with these two difficulties.

2.2 Selecting a force field

When comparing various modern force fields such as AMBER, OPLS or CHARMM,[33,34,36] it can be difficult to pick one force field that repeatedly outperforms the others. Shirts et al. have shown that the OPLS force field is more accurate than others in reproducing the solvation free energies of analogue of amino acid side chains.[106] However, the derivation of parameters for the OPLS force field is impractical for complex small molecules for which limited or no thermodynamic data is available. The AMBER force field is known to perform reasonably in simulations of proteins and nucleic acids. Recently, the General Amber Force Field (GAFF) has been introduced.[107] GAFF has been designed to be compatible with the AMBER parameter sets and to provide force field parameters for small molecules. The recipe for deriving parameters for the GAFF force field expects atomic partial charges derived using the RESP/HF6-31G** method. In this approach, a quantum mechanical package is used to obtain a map of the electrostatic potential around the solute of interest. Atomic partial charges that can reproduce the quantum mechanical electrostatic potential are then obtained by fitting. Apart from the associated computational cost, special considerations have to be taken to deal with buried atoms which tend to be assigned ill-defined atomic partial charges. Interestingly, the basis set 6-31G** is known to overestimate polarisation in the gas phase by 10-20%. This corresponds roughly to the amount of polarisation the solute is expected to receive when solvated.[34] This rather fortuitous error thus provides partial charges suitable for condensed phase simulations. Because of the complexity and

the cost of the RESP method, a new approach has been proposed by Jakalian et al.[108,109] In the AM1/BCC method, the wavefunction of the solute of interest is obtained with the AM1 semi-empirical Hamiltonian.[110] Atomic partial charges are then extracted from the wavefunction, based on a Mulliken population analysis.[111] The Mulliken charges are known to perform poorly in condensed phase simulations. In the second step of the method, a set of bond charge corrections (bcc) are applied to the Mulliken charges to obtain atomic partial charges that reproduce closely the QM electrostatic potential obtained with the RESP method. The AM1/BCC method has been shown to reproduce well solvation free energies of small molecules and interaction energies of various biologically relevant dimers, and at a fraction of the computational cost of the RESP method.[109] The Antechamber program was developed to automate system setup and is part of the AMBER suite of programs.[112] Antechamber provides file conversion, atom typing facilities, force field parameter assignment, and a number of charge calculation methods, including the AM1/BCC method. The combination of the GAFF force field with the AM1/BCC method provides a route by which force field parameters for a large number of small molecules can be derived easily. The availability of the Antechamber program makes automation of system setup a possibility. These methods were therefore adopted for the work covered in this thesis.

2.3 Representing water

An efficient way to decrease the complexity of the system to be studied by a free energy simulation is to reduce the number of modelled particles. Explicit water molecules can be removed from the system and the influence of the solvent on the solute can be represented through the use of continuum solvation techniques such as the Poisson Boltzmann or generalised Born equation.[93,96] Many studies have pointed out that the generalised Born method is very efficient and yet reasonably accurate. This theory was therefore selected to conduct free energy simulations in a realistic solvent environment.

Reddy et al. have studied the transferability of a parameterised GBSA model to different force fields and concluded that while reasonable behaviour is observed

in every case, re-parameterisation for each force field is necessary to yield optimum performance.[113] At the beginning of this project, there was no parameterised model of GBSA compatible with AM1/BCC partial charges available in the literature. A set of parameters compatible with RESP/HF6-31G** partial charges has been proposed,[114] but was judged inadequate because of the small size of its training set and its lack of parameters for sulphur and halides, necessary to cover the chemical functionalities of many drug-like molecules.

Because of the lack of suitable parameter sets, it was decided to parameterise a generalised Born surface area model compatible with AM1/BCC charges and covering a large number of chemical functions that are commonly encountered in drug-like molecules.

2.3.1 Construction of a dataset

Experimental vacuum to water transfer energies of a wide range of simple organic compounds [115] were gathered. A model of each of these molecules was then built for the AMBER99 force field and the atomic partial charges were derived using the AM1/BCC method.

This set contains a balanced mixture of organic functions encountered within drugs. It was also required that these chemicals would not undergo any conformational change upon solvation. This hypothesis was not rigorously tested but compounds showing possible problems were not included (for instance : 1,2 dichloroethane, triethylamine). It is necessary that solutes in the set keep the same conformation in vacuum and in aqueous phase, otherwise the experimental solvation energy would contains force field terms other than the GBSA term that represent the internal strain of the solute. Taking the influence of these terms into account would render the parameterisation process too complex.

The set was split in two groups, a training set with approximately 75% of the compounds and a validation set. Compounds in the validation set were not used to derive coefficients for our solvation model. This allows us to check the transferability of the derived parameters. Table 2.1 summarises the composition of the dataset.

Table 2.1: Composition of the dataset

Family	Training	Validation
Linear Alkanes	6	2
Branched Alkanes	3	1
Cycloalkanes	4	1
Alkenes	6	2
Alkynes	3	1
Arenes	6	2
Alcohols	7	3
Aldehydes	4	2
Ketones	6	2
Carboxylic acids	4	1
Esthers	6	2
Ethers	6	2
Aliphatic amines	10	3
Aromatic amines	6	2
Nitriles	3	1
Amides	3	1
Thiols	6	2
Multi-functionals	8	3
Halides	18	7
Nitro	4	2
O charged	2	1
N charged	9	3
S charged	3	1
Total	133	47

2.3.2 The adjustable parameters of a GBSA model

The equations 1.51 to 1.56 that underly the GBSA theory of solvation have been introduced in the previous chapter. The application of these equations requires the knowledge of a number of parameters. Because there is no *a priori* way to establish the value of these parameters, they are often adjusted so that the complete set of parameters reproduce experimental data. Alternatively, the behaviour of a GBSA model can be compared to the more rigorous Poisson Boltzmann equation.

The apolar component of solvation modelled by the somewhat simplistic equation 1.56 requires a set of σ_k with k depending on the atom type. These parameters

were treated empirically and were fitted to experimental data.

All generalised Born methods require the calculation of a Born radius. This can be achieved by different algorithms. Most of them require the definition of an "intrinsic" Born radius. This quantity corresponds to the radius a completely isolated atom should adopt in order to reproduce its solvation free energy through the Born equation 1.49. Except for a few exceptions, this quantity cannot presently be determined experimentally; it is therefore common practice to derive these parameters empirically. The Born radius can be expected to be reasonably similar to the van der Waals radius of an atom and in this work the intrinsic Born radius of an atom i was derived by multiplying the van der Waals radius of atom i by an appropriate multiplicative offset. In the AMBER force field polar hydrogen atoms usually have a small or zero van der Waals radius which is not appropriate for the definition of the intrinsic Born radius because it leads to too negative solvation energies. To avoid these problems, polar hydrogen atoms are required to adopt a minimum intrinsic Born radius of 1.10 Å.

Once the intrinsic Born radii are defined, "effective" Born radii (or simply Born radii) can be calculated by various algorithms.[96–98,103,116] In this work, we have adopted the Pairwise Descreening Approximation to calculate this quantity,[97] mainly because of its computational efficiency. In equation 1.55, scaling factors S_{ij} are required to correct systematic errors introduced by the approximation. It is not clear how these factors should be derived. Previous studies have adjusted them against the Born radii calculated by more rigorous methods.[97] Others have treated them as empirical parameters and adjusted them to minimise errors against experimental hydration free energies.[117,118] Since there is uncertainty as to which protocol is the most accurate, both approaches were considered in this study. Models where scaling factors are treated empirically will be denoted PDAexp. A numerical integration scheme described by equation 1.54 was used to adjust the scaling factors and the resulting models will be denoted PDAnum.

While it is tempting to introduce more parameters to increase the fit against experimental data, this practice should be avoided because it can lead to overfitting and poor transferability of the model. Very generic atom types were defined and their number was kept low, in contrast to several existing GBSA parameter-

isation.[103,117,118] In total, 14 adjustable parameters were considered. They break down as follows:

. Six different van der Waals offsets to determine α_i are used. One for sp^2oxygen (Osp2), one for sp^3oxygen (Osp3), one for carboxyl and phosphate oxygen (O2, AMBER atom type O2), one for nitrogen (N), one for nitrogen in amino groups (N23, AMBER atom type N2, N3) and one for the remaining atoms (others).

. Six different scaling factors are used in the PDA, one for hydrogen (H), one for carbon (C), one for nitrogen (N), one for oxygen (O), one for sulphur (S), and one for every halide (X).

. Two surface tensions terms are used, a positive value for hydrogen, carbon and fluoride (γ), and a value of zero for the remaining atoms.

We stress that in the PDAnum series, the scaling factors were adjusted to reproduce the Born radii obtained by numerical integration and not experimental solvation free energies. They should therefore not be considered as empirical parameters in that series.

2.3.3 Deriving a set of optimum parameters with a genetic algorithm

With 8 to 14 empirical parameters, it is clear that the optimum set of parameters can not be derived by systematically varying each one of them. Genetic algorithms are stochastic optimization methods that can avoid becoming trapped in local minima and are thus an appropriate method for problems involving large number of parameters.[119] The term genetic is used because of the similarities between this optimisation method and the natural evolution of genes in living organisms.

In a genetic algorithm, the following steps are applied:

1 Generate a population of individuals. Here an individual is a set of parameters that represent one solution to the function to be optimised. The population is a set of these set of parameters.

2 Randomly pair individuals in the population and create offspring by mixing the parameters of the two individuals. This can be accomplished in many ways, for example by cutting each of the two sets of parameters at one point, merging them and selecting one of the two resulting combination. This operation is called crossover.

3 Select some parameters in the offspring and perturb their value by a random amount. This operation is called mutation.

4 Evaluate the fitness of each individual. This is done by evaluating how well each individual performs.

5 Replace the population of parents by their offspring and return to step [2]. This can be done in a number of way, which should be selected to favour individuals with the best fitness. Alternatively, check for a termination criterion.

After a suitable number of generations, step 2 allows for the optimum set of parameters found in the initial population to emerge as the best individual. Step 3 introduces new diversity in the parameters so that, eventually, the whole parameter space can be covered. Many elaborate genetic algorithms exist, but the vast majority follow the simple steps described here.

To derive a set of best performing parameters, the author wrote a program in C that calculates the solvation energy for each molecule in the dataset according to the generalised Born theory. This program was then linked to a Genetic Algorithm C++ library[120] and an optimisation routine was written.

The genetic algorithm used here is an elitist model.[119] The fitness was defined as the unsigned mean error between experimental and predicted hydration free energies. The parameters are represented as a set of real numbers. Uniform crossover is applied with a probability of 0.7. Gaussian mutations are applied with a probability of 0.015. A difficulty sometimes encountered in the application of genetic algorithms is that a few initial individuals of good fitness rapidly overtake the whole population and prevent efficient coverage of the parameter space. To prevent this phenomenon, linear scaling of the fitness, which reduces the disparity between good and bad performing individuals, and a large population of 500

individuals are used. The genetic algorithms were run for a large number of generations to ensure that the results were converged. The best results were selected for further tests.

For the first parameterisation method, where all the parameters are optimised against experimental free energies of hydration, ΔG_{hyd}, the genetic algorithm returned different sets of parameters that minimize the error between predicted and experimental free energies of hydration. The top three solutions showed some differences in the parameters, suggesting the existence of a broad minimum. A broad minimum has also been observed by Rankin[121] when using a PBSA model, where three parameters were optimised against the experimental ΔG_{hyd} of a dataset of organic molecules. Because of the larger number of parameters in our case, an equivalent study is not possible. As observed by other workers, some scaling factors adopted values greater than unity and others are surprisingly small for some atoms. The first two solutions, were judged very similar and only the first was kept. The third solution, which is only marginally worse at predicting hydration free energies, was also selected for further investigation because of the large scaling factors for oxygen it adopted. In the remaining study, the first solution is denoted PDAexp-1 and the third solution PDAexp-2.

In the second parameterisation method, the PDA is not employed initially and the offsets to the van der Waals radii and the surface tension term are optimised with the use of the finite difference scheme for the calculation of the effective Born radii described in the previous chapter by equation 1.54. The best results of the genetic algorithm were all converged toward the same solution to yield a model with an unsigned mean error of 1.01 kcal mol^{-1}. The scaling factors were then derived to minimize the errors caused by the Pairwise Descreening Approximation. This reproduced the polarisation energies obtained by the accurate computation of Born radii to within 0.30 kcal mol^{-1}. When the numerical integration of the Born radii is replaced by the PDA approximation with this set of scaling factors, the mean unsigned error against the experimental free energies of hydration increased to 1.13 kcal mol^{-1}. Here again, no significant difference was found for the scaling factors between the top solutions and therefore a single set of parameters was kept. This set of parameters will be called PDAnum-1 for the rest of this study.

While computing potentials of mean force (see next sections), it was found that slight modifications of PDAnum-1 increased the performance of the model. The modified version is called PDAnum-2.

Table 2.2: Sets of Offsets to the van der Waals radii for the derived models

Model	Osp^2	Osp^3	O2	N	N23	others
PDAnum-1	1.00	0.66	0.88	0.73	0.87	0.86
PDAnum-2	1.00	0.66	0.85	0.73	0.95	0.86
PDAexp-1	1.00	0.80	0.91	0.88	1.00	0.95
PDAexp-2	1.00	0.77	0.89	0.88	0.80	0.95

Table 2.3: Sets of PDA Scaling factors for the derived models

Model	H	C	N	O	S	X^a
PDAnum-1	0.81	0.77	0.70	0.88	0.84	0.93
PDAnum-2	0.81	0.77	0.70	0.88	0.84	0.93
PDAexp-1	0.63	0.57	0.84	1.12	0.96	1.07
PDAexp-2	0.60	0.59	0.91	1.39	0.97	1.12

a X = F, Cl, Br, I

The offsets to the van der Waals radii and the PDA scaling factors for the four solvation models are summarised in tables 2.2 and 2.3. For PDAexp-1 and PDAexp-2 the surface tension term γ is 0.0078 kcal.mol^{-1}.Å^{-2}. For PDAnum-1 and PDAnum-2 the surface tension term γ is 0.0070 kcal.mol^{-1}.Å^{-2}.

The total average mean unsigned error for each solvation model for calculating the hydration free energy is reported in table 2.4 for the training set and the validation set.

Table 2.4: Mean Error in kcal.mol^{-1} for the derived models

Model	Training set	Validation set
PDAnum-1	1.13	1.17
PDAnum-2	1.13	1.07
PDAexp-1	0.79	0.87
PDAexp-2	0.83	1.01

PDAexp-1 and PDAexp-2 outperform PDAnum-1 or PDAnum-2 on average. This is presumably because the scaling factors have been allowed to compensate for errors other than those caused by the PDA. Since PDAnum-1 differs from the two previous models only in the way the scaling factors have been parameterised, and performs significantly worse on the training set, on first inspection it appears it is better to let the scaling factors compensate for other errors than those caused by the Pairwise Descreening Approximation, as has been done previously.[117,118] This assertion will be examined in the next sections.

2.3.4 Comparison with Poisson Boltzmann calculations

The performance of generalised Born models is often assessed by comparing the electrostatic component of solvation yielded by a GB model to Poisson-Boltzmann (PB) calculations. PB calculations were also run on the dataset of small molecules with the program APBS.[122] We note, however, that when such comparisons have been made, the dielectric boundary in the two models is usually not the same.[116,118] Since PB results are quite sensitive to the definition of the dielectric boundary, comparison between the two methods must be made with care. In this study, we chose to compare the GB models to a PB model derived by Rankin et al. because the PB models reported in that study have been optimised against the hydration free energy of a dataset of small molecules compatible with the AMBER force field.[121] The set of models studied by Rankin consist of three parameters (ρ, D_{in}, α) which were systematically varied within a selected range. ρ is an offset to the AMBER van der Waals radii, D_{in} is the interior dielectric constant and α the surface tension. We chose the model (ρ = 1, D_{in} = 1, α = 0.0070 kcal.mol^{-1}.Å^{-2}) because it gives good agreement against the experimental data used in their study, has the same interior dielectric constant implied by the GB theory, and a surface tension similar to our models. Note that the surface tension is not important because we are only interested in comparing the electrostatic component of solvation between the GB and PB models. In their study, the radius of polar hydrogens is set to 1.0 Å.

The following protocol was used for the PB calculations. The grid size was set

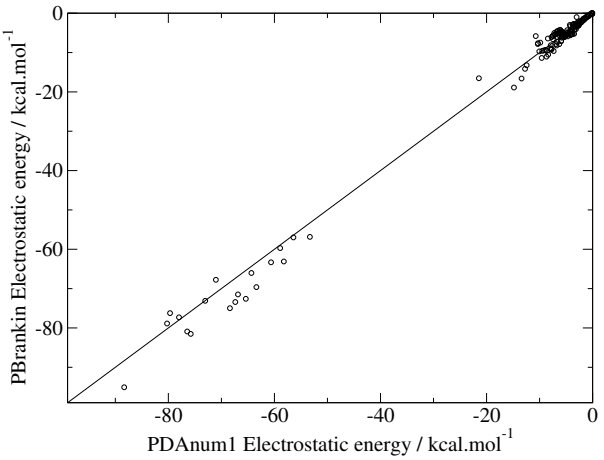


Figure 2.1: Plot of the electrostatic component of solvation ΔG_{pol} computed by Poisson Boltzmann calculations (PBrankin) and generalised Born calculations (PDAnum1).

to 65x65x65 points, the spacing was 0.25 Å, and multiple Debye-Hückel boundary conditions were used. The interior dielectric was set to 1.0, the exterior dielectric set to 78.0. To ensure that the computed electrostatic energies were not sensitive to the protocol, they were calculated for thirty random orientations of each small molecule in the grid. This showed that the energies were accurate to within 0.1 kcal mol^{-1} for neutral compounds, and to within 0.4 kcal mol^{-1} for charged compounds (data not shown).

The effective Born radii computed by the generalised Born models were also compared to the 'perfect' Born radii computed by the PB approach.[123] The 'perfect' Born radius for an atom i in a molecule was derived by solving the PB equation where every atomic partial charge but the one on atom i was set to zero. The 'perfect' Born radius is then readily available through equation 2.1 where ΔG_{pol}^{PB} is the solvation energy derived by the PB calculation.

$$B_i^{perfect} = -\frac{1}{2}\left(\frac{1}{\varepsilon_{vac}} - \frac{1}{\varepsilon_{solv}}\right)\frac{q_i^2}{\Delta G_{pol}^{PB}} \tag{2.1}$$

The electrostatic component of the hydration free energies (ΔG_{pol}) calculated by the four GB models was compared to the values obtained by the Poisson Boltzmann model of Rankin[121] (PBrankin). The average unsigned difference for PDAnum1, PDAnum2, PDAexp1 and PDAexp2 is 1.08 kcal mol^{-1}, 1.14 kcal mol^{-1}, 1.08

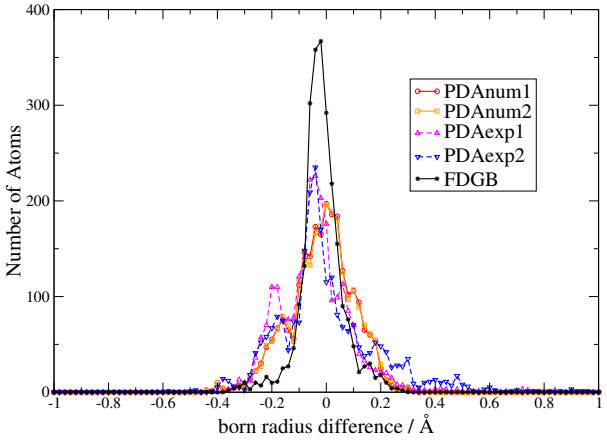


Figure 2.2: Distribution of the difference between the effective Born radii computed with the different generalised Born Models and the 'perfect' Born radii given by Poisson Boltzmann calculations.

kcal mol^{-1} and 1.26 kcal mol^{-1} respectively. Thus, these models show roughly equal differences with respect to the PB calculations. As an illustrative example, figure 2.1 plots the electrostatic component of solvation of each compound in the dataset obtained with PDAnum1 against the results obtained with PBrankin.

To determine how much of this difference is caused by the Pairwise Descreening Approximation, the same calculation was performed with the GB model that employs the numerical integration scheme to compute the effective Born radii (referred in the text as FDGB). The average unsigned difference is 1.01 kcal mol^{-1}. Thus, the PDA is only responsible for a moderate amount of the discrepancy between the PB and GB results.

Figure 2.2 is a plot of the deviations between the 'perfect' Born radius derived with PBrankin and the effective Born radius computed with the GB models for the entire dataset. FDGB comes closest to the perfect Born radii because it uses a more accurate scheme to compute the effective Born radii. PDAnum1 and PDAnum2, essentially indistinguishable on this plot, are more noisy and show larger deviations. PDAexp2 shows the largest deviations and a tendency to overestimate the effective Born radii. On a few occasions the deviations are very large for PDAexp1 and PDAexp2 (greater than 1 Å). Thus a seemingly uniform level of agreement with the PB calculations hides a different behavior in the computation of the effective Born radii for the various GB models.

Table 2.5: Polarisation energy (ΔG_{pol}) and hydration free energy (ΔG_{hyd}) of octafluoropropane in kcal mol^{-1}

Model	ΔG_{pol}	ΔG_{hyd}	Error[a]
PDAnum-1	-2.08	0.06	4.24
PDAexp-1	+1.94	4.33	0.03
PDAexp-2	+2.27	4.67	0.37
FDGB	-2.13	0.01	4.29
PBrankin	-2.18	-0.04	4.34

a against the experimental hydration free energy of octafluoropropane (+4.30 kcal mol^{-1}, taken from ref[118])

Table 2.6: Effective Born radii of every atom in octafluoropropane calculated by the different models (in Å)

Atom	PDAnum1	PDAexp1	PDAexp2	FDGB	PBrankin
CT	2.54	3.61	4.04	2.49	2.49
CT	2.84	4.15	4.73	2.61	2.62
CT	2.54	3.61	4.04	2.49	2.51
F	2.09	2.69	2.88	1.95	1.96
F	2.11	2.71	2.91	1.96	1.97
F	2.03	2.53	2.67	1.90	1.90
F	2.14	2.81	3.03	1.96	1.97
F	2.14	2.81	3.03	1.96	1.96
F	2.10	2.70	2.89	1.95	1.97
F	2.10	2.70	2.89	1.96	1.97
F	2.03	2.53	2.67	1.90	1.95

It is somewhat surprising that the models that give the best agreement with experimental observable are the most different from the PB calculations. The presence of some very large deviations is also a source of concern. Their origin can be explained by the protocol employed during the parameterisation. A good example is given by the compound octafluoropropane. Table 2.5 lists the electrostatic component of solvation (ΔG_{pol}), the predicted hydration free energy (ΔG_{hyd}), and the agreement with experiment. The experimental hydration free energy of octafluoropropane is difficult to reproduce because its large positive value (+4.30 kcal mol^{-1}, taken from ref[118]) cannot be recovered by the values of surface

tension coefficients that are required for other groups of small molecules. Because this free energy is the target value during the optimization process and PDAexp1, PDAexp2 are free to adopt any set of parameters, they produce a hydration free energy in good agreement with experiment. This is achieved, however, at the cost of a positive polarisation energy which is clearly unphysical. By contrast, PDAnum1 (PDAnum2 is identical to PDAnum1 for this particular compound) yields polarisation energies in good agreement with FDGB or PBrankin (and a large error against experiment). The source of these differences is related to the very different effective Born radii that the GB models have computed. Table 2.6 clearly shows that with PDAnum1, the effective Born radii are in good agreement with those computed by FDGB which themselves are in very good agreement with the 'perfect' Born radii of PBrankin. PDAexp1 and PDAexp2, yield however very different effective Born radii.

2.3.5 Behaviour of the parameterised models in potential of mean force calculations

Potentials of mean force for the association of various species were also computed for the best models. Given that GBSA models are often used to perform simulations of molecular complexes, the ability of these models to reproduce accurate potentials of mean force is more important than their ability to predict solvation free energies in good agreement with experiment.

Systems for which PMFs have been previously derived in the literature[124–127] were selected so that comparison could be made. Quantitative agreement is not expected because issues such as long range treatment of electrostatics, solvent model and partial charges can lead to different results for the same system (see ref[128,129] for good examples). However, if the physics is adequately modelled, one would expect broad agreement between the various solvation models. The selected systems encompass a broad range of interactions (hydrophobic, aromatic, polar, ionic).

To compute the PMFs, Metropolis Monte Carlo sampling[39] was used and free energy differences were computed using the Free Energy Perturbation Method

(FEP)[130] as implemented in the program MCPRO.[131] The simulations were performed at 298 K. Windows were positioned approximately every 0.2 Å along the coordinate of interest. The number of moves used to equilibrate the system and collect statistics depended on the system considered. For very simple systems such as methane, each window was equilibrated for 1000 (1K) moves and data collected for 10000 (10K) moves. For more complex systems where more internal degrees of freedom have to be considered, each window was equilibrated for 10K moves and data collected for 100K moves. For constrained PMFs, each window was equilibrated for 5K moves and data collected for 50K moves. The number of moves needed to converge free energy differences is much smaller than would be needed in an explicit solvent simulation, since there are far fewer degrees of freedom.

Solvent moderated packing effects are usually not modelled by continuum solvation models although it is possible to parameterise empirical models to obtain a solvent separated minimum by means of a penalty related to the solvent volume excluded by two approaching species.[132] In this study, the continuum solvation models only yield a contact minimum (CM).

Because our models PDAnum-1 and PDAnum-2 are very similar (they can only differ on systems that involve nitrogen atoms of AMBER type N2, N3 or oxygen atoms of AMBER type O2), PMFs involving PDAnum-2 were only computed when the results would differ from those obtained with PDAnum-1.

Unconstrained PMFs, where all the internal degrees of the freedom but the reaction coordinate are freely sampled, have been computed for a methane pair, a benzene pair and a N-methylacetamide (NMA) pair. The results can be compared with previous work from the Jorgensen group.[124–126] The remaining PMFs have been computed by constraining the pair of molecules in a defined orientation. The systems selected are amino acids side chains. PMFs for these systems have been computed by Masunov[127] with the CHARMM[36] force field using a generalised Born model,[133] the EEF1 solvation model[134] and a primitive electrolyte model (similar to a vacuum simulation with ε set to 80). The results were compared to those obtained with the Spherical Solvent Boundary Potential (SSBP) hybrid solvation model developed by Beglov and Roux.[135]

Methane pair

This PMF is a test case for hydrophobic interactions. Because the polarisation term is negligible for this compound, the solvent effects are only due to the surface area dependent term. The (all atom) molecules are kept rigid. A very small amount of sampling over each window is required to converge the free energy differences. We report here PMFs generated with 1K moves for equilibration and 10K moves for collection. By contrast, Jorgensen et al.[124] have carried out a simulation on this system (united atom) with explicit solvent and used 500K moves for equilibration, followed by 2M moves for collection. The PMF computed in vacuum and with a GB/SA term are reported below on figure 2.3.

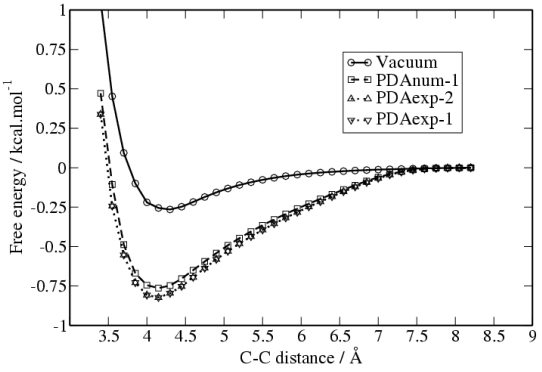


Figure 2.3: Potential of Mean Force Methane-Methane

A CM (Contact-Minimum) is present at a separation of 4.1 Angstroms. In vacuum, the well-depth is about -0.3 kcal mol^{-1} and arises from attractive Lennard-Jones interaction. In solution, our various GB/SA models perform similarly with a net attraction of around -0.80 kcal mol^{-1}; this is because the surface tension term does not vary much between the different models. Since the GB term is negligible for methane, the energy change in solution is solely related to the change in SASA and the association of the two species is favoured to reduce the total SASA of the system. Jorgensen has computed a binding free energy of -0.42 +/-0.34 kcal mol^{-1} for that system. Thus our model is slightly more attractive. We note however, that the hydration free energy of a single methane molecule is under-estimated with our solvation models (between 1.1-1.2 kcal mol^{-1} instead of 2.0 kcal mol^{-1}). If our models predicted a more accurate solvation energy, the dimer would become more stabilised.

Benzene pair

This PMF has been computed along the distance of the centre of mass of each benzene molecule. Our parameters for benzene are quite similar to the ones employed in a study by Jorgensen et al.[125] A Monte Carlo optimisation in vacuum reveals that the global minimum is a roughly T-shaped dimer with a separation of 4.8 Å. The interaction energy is -2.40 kcal mol^{-1}. Jorgensen has found a slightly more perpendicular configuration at a separation of 5.0 Å, with an interaction energy of -2.31 kcal mol^{-1}.

The PMF in vacuum and with our GBSA models for the association of two rigid benzene molecules are reported below. For each window 11K moves were performed and averaging was usually done on the last 10K.

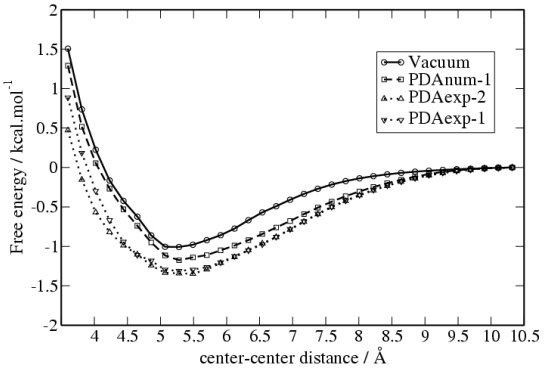


Figure 2.4: Potential of Mean Force Benzene-Benzene

According to these results, the strength of the interactions between two benzenes is stronger in water than in vacuum. This is expected because of the hydrophobic effect. However figure 2.4 shows that our solvation models exhibit more marked differences than with methane. A CM is observed around 5.1-5.3 Å with a well depth ranging from 1.2 to 1.6 kcal mol^{-1}. In the study of Jorgensen a CM at 5.5 Å with a well-depth of -1.5 kcal mol^{-1} was found. Jorgensen also noted that integration of his computed PMF overestimated the association of benzene in water according to experimental measurements.[136]

If the association was purely driven by hydrophobic forces, we would expect the PMF in solution to be more stabilised by about 0.7 kcal mol^{-1} than in vacuum, because of the difference of SASA for the two separated species, and the

dimer at the CM (difference of 100 $Å^{-2}$ with a surface tension of around 0.007 kcal $mol^{-1}.Å^{-2}$). This is almost the case with PDAexp1 and PDAexp2, but in PDAnum1, as the two species get closer, the GB term becomes more positive, reducing the amount of stabilisation due to the reduction of SASA. This is expected as the dimer exhibits an attractive Coulombic attraction and part of the GB equation is anti-correlated with the Coulombic term. Benzene is expected to be able to form weak-hydrogen bonds with water and desolvation would reduce its ability to form these hydrogen bonds.[137] The different behaviour of the solvation models is caused by the different scaling factors for H and C. In PDAexp1 or PDAexp2, as the two molecules get closer, the Born radii do not increase much, because the descreening influence of one atom is strongly reduced by the scaling factors. If there are little or no variations in Born radii, then the GB term does not vary much, and thus does not oppose the Coulombic attraction. With PDAnum1, larger variations are seen because the scaling factors reduce less the descreening, causing ultimately the PMF to be weaker.

Jorgensen predominantly observed around the CM a range of distorted T-shaped pairs, including some roughly parallel stacked and displaced pairs. In the vicinity of the CM with our GBSA models we tend to see more perpendicular T-shaped pairs and rarely parallel stacked and displaced pairs. As noted before, at a short distance of around 4 Å, where only face to face stacking is possible, the net interaction is repulsive, even though the gas phase interaction energy is almost -2.0 kcal mol^{-1}. This is because this conformation, configurationally restricted, is entropically disfavoured.

N-methylacetamide N-methylacetamide

The association of two N-methylacetamide molecules (NMA) in solution is representative of a polar interaction driven by the formation of a hydrogen bond at short separation. This PMF was constructed by constraining the distance between the centre of geometry of each molecule.. PMFs in vacuum and for the different models are reported in figure 2.5. In vacuum the formation of a hydrogen bond is favoured by approximately -4.5 kcal mol^{-1}. In water this interaction is

considerably reduced because formation of the hydrogen bond between the two amides involve breaking hydrogen bonds with water. With TIP4P water,[138] Jorgensen found no attraction between two NMA molecules, which is consistent with the experimental association constant.[126] PDAnum-1 and PDAexp-1 show similar variations and exhibit a CM at 4.7 Å with a well-depth of -0.8 kcal mol^{-1} and -1.0 kcal mol^{-1} respectively. PDAexp-2 varies similarly until about 5.2 Å where the free energy drops until it finally exhibits a contact minimum (CM) at 4.7 Å with a well-depth of -1.8 kcal mol^{-1}.

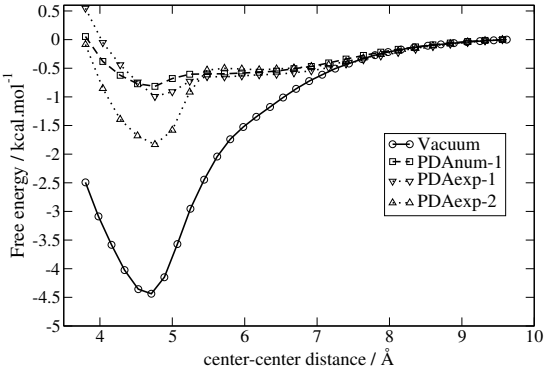


Figure 2.5: Potential of mean force NMA-NMA

From the average value of the total GBSA energy of the system at large separation, one can obtain the free energy of hydration of NMA which is reported in table 2.7. Note that this value can differ slightly from that predicted for the rigid model of NMA used to parameterise the solvation models, because the NMA molecules used in this simulation are flexible and the hydration energy is averaged over all the configurations available to each molecule.

Table 2.7: Hydration free energy of N-methylacetamide in kcal mol^{-1}

Model	Hydration free energy
PDAnum-1	-6.4
PDAexp-1	-7.0
PDAexp-2	-6.5
exp	-10.1[a]

a from ref[98]

Table 2.7 shows that the three solvation models predict approximately similar

hydration free energies. Note that error against experiment is large. Yet PDAnum-1 and PDAexp-1 yield energies at the CM that are within 1 kcal mol^{-1} of the results obtained by Jorgensen. Furthermore the error on the hydration free energy cannot explain why PDAexp-2 is twice as attractive as the other solvation models. As noted in tables 2.2 and 2.3, the only difference between PDAexp-2 and PDAexp-1 for this system are the scaling factors of N and O. Figure 2.6 shows the average value of the Born radius of the polar hydrogen involved in hydrogen bonding at different separation distances. With PDAexp-2 when the hydrogen atom comes into close contact with the oxygen atom of the other NMA molecule, the PDA approximation significantly overestimates the value of the Born radius for the hydrogen. This stabilises the hydrogen bonded configurations and causes the attraction between the two NMA molecules to increase. Since models that have the same accuracy for the prediction of the free energy of hydration of NMA can behave quite differently for the computation of free energies of interaction, this suggest that parameterisation of continuum models solely against free energies of hydration does not ensure a proper treatment of intermolecular interactions in solution.

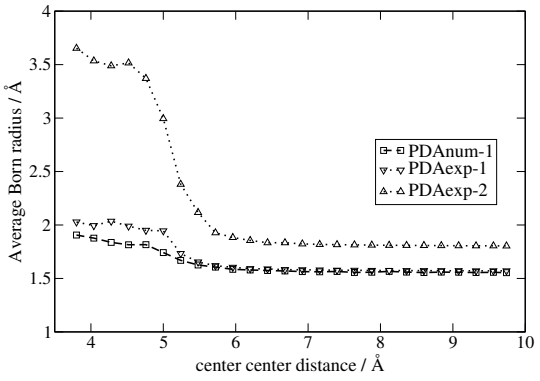


Figure 2.6: Average Born radius of hydrogen bonded atom H

To determine the extent of the error on the computation of the Born radius for the polar hydrogen involved in the hydrogen bond of the NMA dimer, a snapshot representative of a hydrogen bonded configuration was extracted from one simulation with a separation of 4.5 Å between the centres of mass of each molecule. The Born radius of the polar hydrogen is computed using the previous solvation models, and compared to the value of the Born radius found by the numerical integration method (eq 1.54) using the set of intrinsic Born radii appropriate for each

solvation model.

Table 2.8: Born radius of the hydrogen bonded atom H in the N-methylacetamide dimer in Å

Model	PDA	Numerical
PDAnum-1	1.8	1.9
PDAexp-1	2.0	2.0
PDAexp-2	3.2	2.1

The Born radii reported in table 2.8 shows that for models PDAnum-1 and PDAexp-1, the PDA approximation reproduces fairly well the Born radius of the polar hydrogen obtained with the numerical integration method (even though for PDAexp-1 the scaling factors were not specifically derived to reproduce accurately the Born radii), but PDAexp-2 shows a large error.

Glu- Arg+

In this PMF, the carboxylate group of the ionized glutamic acid side chain (Glu-) and the guanidinium group of the arginine side chain (Arg+) are constrained to lie in the same plane. The reaction coordinate is the distance between the carbon atom in the carboxylate group of the glutamic acid and the carbon atom in the guanidinium group of the arginine.

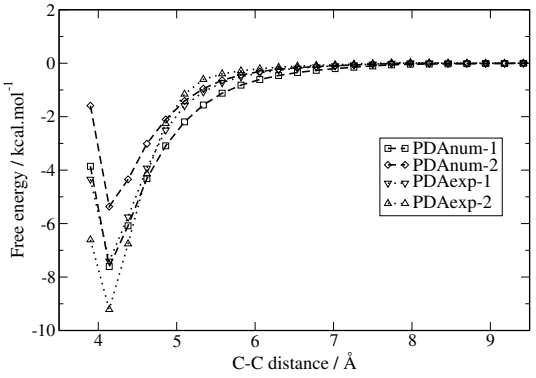


Figure 2.7: Potential of mean force Glu- Arg+

Masunov reports a well-depth of -4.5 kcal mol^{-1} with the SSBP simulation. The GB model from CHARMM is attractive by about -4.0 kcal mol^{-1}. Figure 2.7 shows that all the solvation models exhibit a strong attraction. A CM is observed

at 4.1 Å and the well-depth is -9.2 kcal mol^{-1} for PDAexp-2, -7.4 kcal mol^{-1} for PDAexp-1 and -7.6 kcal mol^{-1} for PDAnum-1. Interestingly, with PDAnum-1 the well-depth can be reduced significantly by just changing the offset for the parameter O2 from 0.88 to 0.85 and increasing the offset for the parameter N23 from 0.87 to 0.95. The resulting model, called PDAnum-2 yields an attraction of -5.3 kcal mol^{-1} which is in closer agreement with Masunov results. It was therefore decided to also compute PMFs involving AMBER atom types O2, N2 or N3 with PDAnum-2. This illustrates that potentials of mean force can be used to adjust parameters of continuum solvation models.

Glu0 Glu-

In this PMF, the functional groups of the neutral glutamic acid side chain (Glu0) and the charged glutamic acid side chain (Glu-) are constrained to stay in the same plane (head to head approach) and the reaction coordinate is the distance between the two carbon atoms in each functional group.

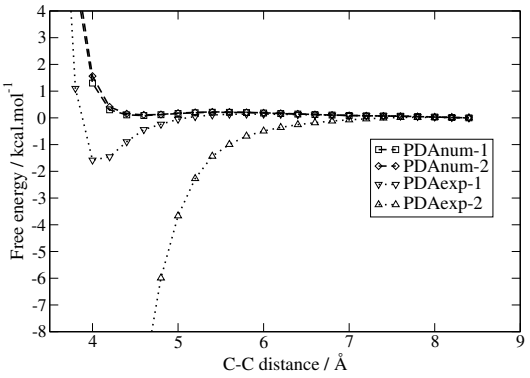
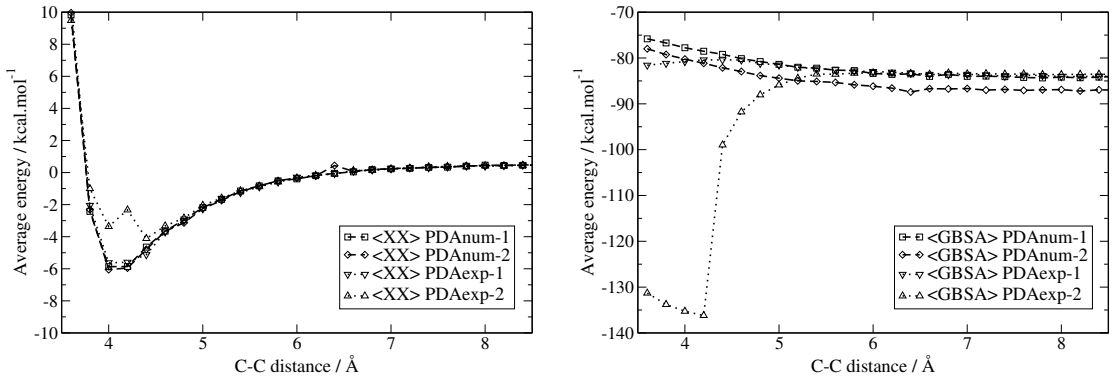


Figure 2.8: Potential of mean force Glu0 Glu-

Masunov reports a well-depth of -2.0 kcal mol^{-1} with the SSBP simulation. The GB model from CHARMM is repulsive by almost 2.0 kcal mol^{-1} at the CM. Figure 2.8 shows that our simulations exhibit different behavior. PDAnum-1 and PDAnum-2 show no attraction, PDAexp-1 is attractive by about -1.6 kcal mol^{-1} at 4.0 Å. Results with PDAexp-2 appear unphysical. As the two molecules becomes closer, the attraction increases very quickly. At a separation smaller than 4.5 Å the free energy differences between neighboring windows becomes too large for the results to be reliable.

Figure 2.9 plots the average GBSA energy and the vacuum intermolecular energy between the two molecules. It shows that the increased attraction observed in the PMF using PDAexp-2 is due to a sudden drop of the GBSA energy at short separation (figure2.9b). Interestingly, this causes the simulation with PDAexp-2 to sample different configurations than those observed with the other simulations at around 4-4.5 Å, as shown by the different values of the average vacuum intermolecular energy (figure 2.9a). The difference in behavior between PDAnum-2 and PDAexp-1 in figure 2.8 also seems to be due to the slightly decreasing GBSA energy of PDAexp-1 at short separation (figure 2.9b).



(a) (b)

Figure 2.9: Glu0 Glu-: Average vacuum intermolecular and GBSA energies (a) Average vacuum intermolecular energy (b) Average GBSA energy

The different behavior of the solvation models can be linked to the Born radius of the polar hydrogen HO. Figure 2.10 shows that with PDAexp-2 the Born radius of atom HO increases very quickly and tends toward infinity at short separation. This large, unrealistic, increase in the Born radius is due to the large value of the scaling factors of the nearby oxygen atoms, which increase their overlap with atom HO. In this case, the Pairwise Descreening Approximation is a particularly poor approximation which leads to these large errors. PDAexp-1 also sees a lesser increase of the Born radius because of the value of the scaling factor for oxygens it uses. This causes an increased attraction and fortuitously a better agreement with the SSBP simulation from Masunov. PDAnum-1 and PDAnum-2, with a smaller Born radius for HO, yield no attraction and are the most similar to the GBSA model from CHARMM.

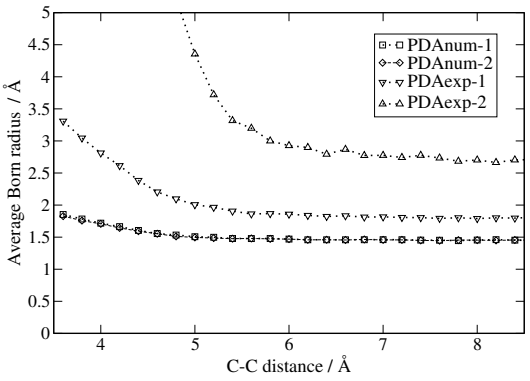


Figure 2.10: Average Born radius of atom HO

As before, a snapshot was extracted from the simulations at a separation of 4.0 Å between the carbon atoms of the carboxylic/carboxylate groups. In table 2.9, the Born radius for HO computed with the PDA approximation was compared with the Born radius obtained by numerical integration (eq 1.54).

Table 2.9: Born radius of the hydrogen bonded atom HO in the Glu0 Glu- dimer in Å

Model	Numerical	PDA
PDAnum-1	1.7	1.7
PDAnum-2	1.7	1.7
PDAexp-1	1.9	2.8
PDAexp-2	1.8	14.3

PDAnum-1 and PDAnum-2 find values that match closely the Born radius obtained by numerical integration because the scaling factors have been optimised to correct for the systematic errors in the PDA approximation. PDAexp-2 and PDAexp-1 show deviations because the scaling factors were not used to correct the systematic errors of the PDA approximation, but were instead used to reduce the error against the experimental hydration free energies.

Glu0-Glu0 pair

A PMF for the interaction of two neutral Glutamic acid side chains has been derived for a coplanar approach. The C-C distance corresponds to the distance between the carbon atoms from each carboxy group.

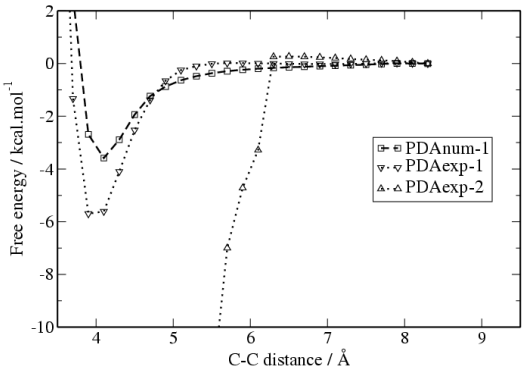


Figure 2.11: Potential of Mean Force Glu0 Glu0

Figure 2.11 shows that, apart from PDAexp-2 which exhibits a complete failure, all models exhibits a CM around 4.0 Å. The well depth is -3.6 kcal mol^{-1} for PDAnum-1. and -5.7 kcal mol^{-1} for PDAexp-1. In the study by Masunov, a well-depth of approximately -1.8 kcal mol^{-1} has been found. This suggest that our model are too attractive. Quite interestingly, the PMF computed with the GBSA model from CHARMM was found to be slightly repulsive. The origin of the failure of PDAexp-2 is similar to the one observed in the potential of mean force between Glu- and Glu0.

Arg+ Arg+ pair

Two possible orientations have been considered. In the first, the two guanidinium moieties are constrained to stay in the same plane and the reaction coordinate is the distance between the carbon atom present in the functional group.

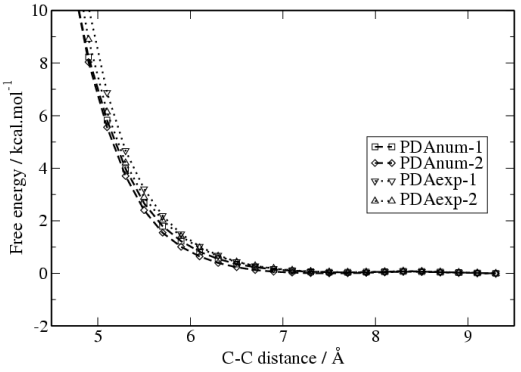


Figure 2.12: Potential of Mean Force Arg+ Arg+ coplanar approach

All the PMFs on figure 2.12 are repulsive until about 6.5 angstroms. No CM is really observed. The SSBP simulation of Masunov finds a CM at 4.6 angstroms

with a well-depth of 0.0 $kcal.mol^{-1}$. The GBSA model from CHARMM is similar to ours but falls off with distance more quickly . It is repulsive by 2 $kcal.mol^{-1}$ at 4.6 angstroms. The difference most likely arise from the different parameters for the non-bonded interaction of the polar hydrogens of the guanidinium moiety in AMBER (r* = 0.60 Angstroms) and CHARMM (r* = 0.22).

The two side chains can also approach each other in a stacked fashion which has also been considered.

By analysing the PDB database, Soetens et al. have observed a stacked conformation as the preferred mode of interaction between two hydrated arginine side-chains when the C-C distance of the guanidinium groups is small (below 4 Å).[128] The amount of stabilisation due to this interaction is controversial and Soetens reports values ranging between -10.0 $kcal.mol^{-1}$ and -2.7 $kcal.mol^{-1}$, depending on the water model used . Masunov reports a well depth of approximately -1.0 $kcal.mol^{-1}$ for the SSBP simulation, but the solvent separated minimum (SSM) that lies at about 6.5 angstroms is slightly deeper. Soetens noticed only a very shallow SSM but argued that it could have been due to insufficient sampling. The GBSA model of CHARMM yields a well-depth of -1.0 $kcal.mol^{-1}$ at about 3.6 angstroms.

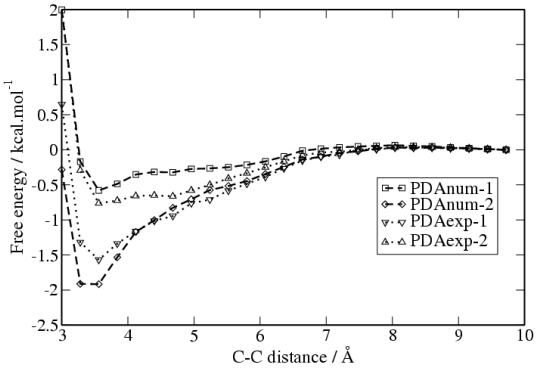


Figure 2.13: Potential of Mean Force Arg+ Arg+ stacked approach

The PMFs from figure 2.13 are quite sensitive to the solvation model. Stabilisation free energies ranges between -1.9 and -0.6 $kcal.mol^{-1}$ approximately. We note that, although the side chains were kept parallel, they were allowed to rotate and the two hydrocarbons moieties could assume a parallel or anti-parallel conformation, while Masunov kept the orientation fixed (anti-parallel).

As noted above, quantitative comparison is difficult. We see however, that our results agree with the analysis of Soetens: the stacked approach is clearly preferred over the coplanar at short distance. Trends between stacked, planar and a third T-shaped geometry are more complex at longer distance and could not presumably be reproduced by a continuum model.

HisP Glu- pair

The system simulated is a protonated histidine (HisP) and a glutamic acid (Glu-) side chain. The molecules are constrained in a coplanar approach and the reaction coordinate is the distance between the δ nitrogen of the imidazole group and one oxygen atom of the carboxylate group.

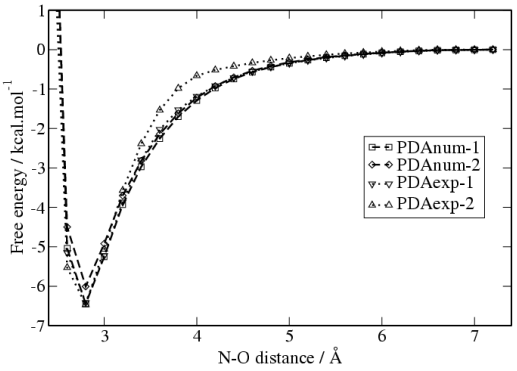


Figure 2.14: Potential of Mean Force HisP Glu- coplanar approach

Masunov reports a well-depth of about -1.0 kcal mol^{-1} at a CM of 2.8 Å with the SSBP potential. The GBSA model from CHARMM is attractive by about -2.5 kcal mol^{-1}. PMFs on figure 2.14 are more attractive, with well-depths between -6.0 and -6.5 kcal mol^{-1}. Since other PMFs including Glu- appears correct, the stronger attraction is probably caused by the protonated Histidine.

HisD Glu- pair

The system simulated is a delta tautomer histidine (HisD) and a glutamic acid (Glu-) side chain. The molecules are constrained in a coplanar approach and the reaction coordinate is the distance between the δ nitrogen of the imidazole group and one oxygen atom of the carboxylate group.

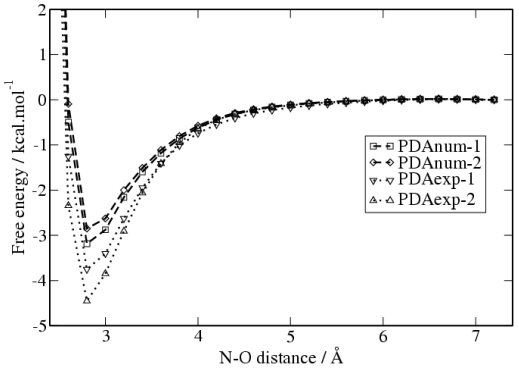


Figure 2.15: Potential of Mean Force HisD Glu- coplanar approach

Masunov reports a well-depth of about -2.5 kcal mol^{-1} at a CM of 3.0 Å with the SSBP potential. The GBSA model from CHARMM is attractive by about -1.0 kcal mol^{-1}. Figure 2.15 shows that PDAnum1 and PDAnum2 comes closest to the SSBP results with well depths of -3.2 and -2.8 kcal mol^{-1} respectively.

HisD HisE pair

The system simulated is a delta tautomer histidine (HisD) and a epsilon tautomer histidine (HisE) side chain. The molecules are constrained in a coplanar approach and the reaction coordinate is the distance between the δ nitrogen of the δ tautomer of the imidazole group and the δ nitrogen of the ε tautomer of the other imidazole group.

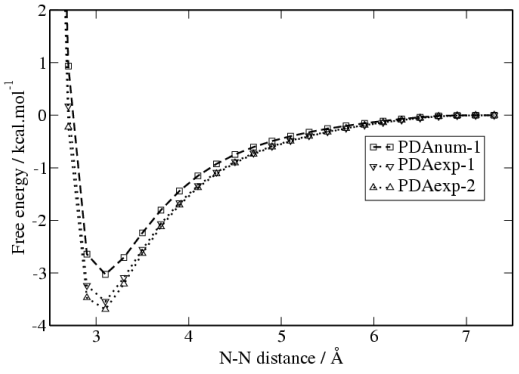


Figure 2.16: Potential of Mean Force HisD HisE coplanar approach

Masunov reports a well-depth of about -1.75 kcal mol^{-1} at a CM of 3.0 angstroms with the SSBP potential. The GBSA model from CHARMM is attractive by about -2.0 kcal mol^{-1}. With our models, PDAnum1 comes closest to the

SSBP results with a well-depth of -3.0 kcal mol^{-1}.

HisP HisE pair

The system simulated is a protonated histidine (HisP) and a ε tautomer histidine (HisE) side chain. The molecules are constrained in a coplanar approach and the reaction coordinate is the distance between the δ nitrogen of the δ tautomer of the imidazole group and the δ nitrogen of the ε tautomer of the other imidazole group.

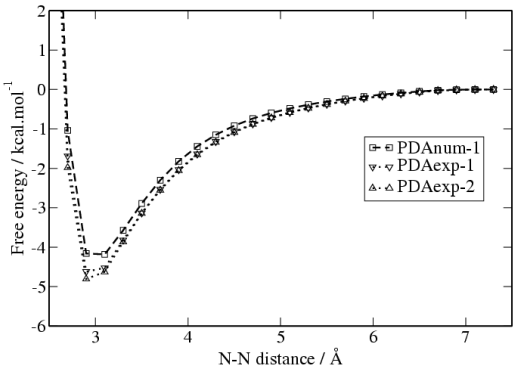


Figure 2.17: Potential of Mean Force HisP HisE coplanar approach

Masunov reports a well-depth of about -2.0 kcal mol^{-1} at a CM of 3.0 Å with the SSBP potential. The GBSA model from CHARMM is attractive by about -1.0 kcal mol^{-1}. Figure 2.17 shows that PDAnum1 comes closest to the SSBP results with a well-depth of -4.2 kcal mol^{-1}.

HisP HisP pair

The system simulated is two protonated histidines. Here two possible approaches have been considered. In the first one, the two imidazoles rings are constrained to stay in the same plane. The reaction coordinate is the distance between the δ nitrogen of the δ tautomer of the imidazole group and the δ nitrogen of the ε tautomer of the other imidazole group.

As can be seen on figure 2.18 all the GBSA models exhibits litlle or no attraction. They are quite similar to the result Masunov reports for the GBSA model from CHARMM. However the SSBP simulation exhibits a well-depth of -2.0 kcal mol^{-1} around 4.7 angstroms.

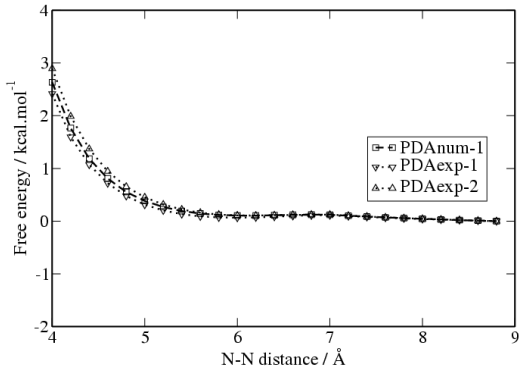


Figure 2.18: Potential of Mean Force HisP HisP planar approach

The second considered approach is a stacking between the two rings. The reaction coordinate is the distance between the centre of each imidazole ring.

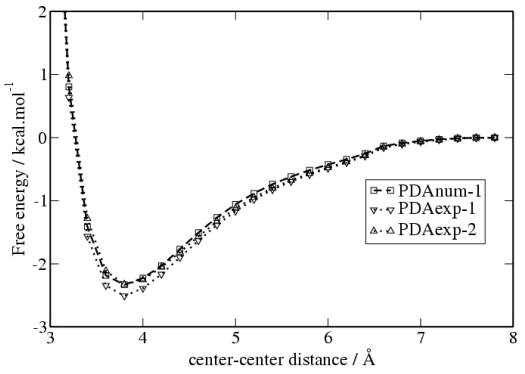


Figure 2.19: Potential of Mean Force HisP HisP stacked approach

Masunov reports a SSM deeper (+0.0 kcal mol^{-1}) than the CM (+1.0 kcal mol^{-1} at 3.8 Å).The GBSA model from CHARMM is entirely repulsive. Present models behave quite differently, and yield well-depths of about -2.5 kcal mol^{-1} at about 3.8 Å (see figure 2.19).

Lys+ Glu- pair-coplanar

The molecules are constrained in a coplanar approach and the reaction coordinate is the distance between the nitrogen of the amino group and the carbon of the carboxylate group.

Masunov reports a CM at 3.2 angstroms with a well-depth of about -2.2 kcal mol^{-1}. The GBSA model from CHARMM exhibits a CM at 3.5 Å with a well depth of -1.75 kcal mol^{-1}. Our model PDAnum1is strongly attractive with a well-depth

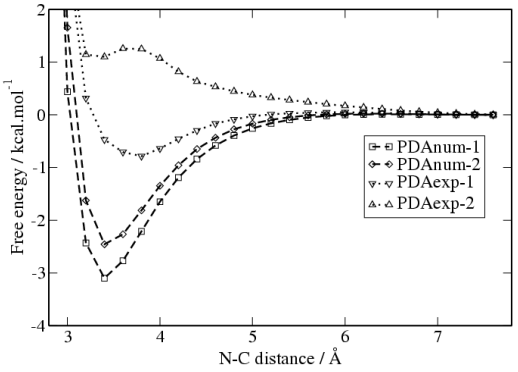


Figure 2.20: PMF Lys+ Glu- coplanar approach

of -3.1 kcal mol^{-1} and a CM at 3.2 Å. PDAnum2 yields a weaker well depth of -2.4 kcal mol^{-1} in good agreement with the SSBP simulations. PDAexp1 is only weakly attractive with a free energy at the CM of -0.8 kcal mol^{-1} while PDAexp2 exhibits a repulsive profile. The behavior of PDAexp1 and PDAexp2 is once again linked to the large error on the Born radii of the polar hydrogen atoms of the lysine moiety that are overlapped by the oxygen atoms of the glutamic acid moiety.

2.3.6 Cause of errors in the PMFs

Inspection of equation 1.55 reveals that the large errors in the computation of Born radii for PDAexp-2 are not due to values greater than one for some scaling factors. This is best illustrated using the simple scheme in figure 2.21.

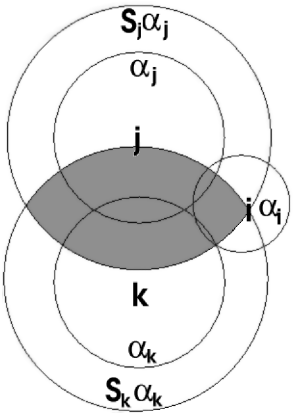


Figure 2.21: Typical case of failure of the PDA

As explained before, when computing the effective Born radius of an atom i of intrinsic Born radius α_i, spheres of adjacent atoms j, k of radius $S_j\alpha_j$ and $S_k\alpha_k$ are considered, where S_j and S_k are scaling factor for the atom j and k of intrinsic

Born radius α_j and α_k. If these products are such that the radius of spheres $S_j\alpha_j$ and $S_k\alpha_k$ is significant compared to the distance between atoms j, k and i, the spheres may have a significant overlap, here represented by the shaded area. Note that this could also happen with an atom that has a large intrinsic Born radius and a small associated scaling factor.

While this is well known, of particular interest in this case is that a fraction of the overlap occurs very close to the centre of sphere α_i. Because the amount of descreening caused by atom j on atom i is related to the ratio of the volume of atom j to the distance between atoms i and j raised to the fourth power,[139] the error due to the Pairwise Descreening Approximation will be large here.

In a recent study, Feig et al. have reported a comparison of the performance of various GB models against PB models.[116] The authors noted that a number of GB models implemented in AMBER that use the Pairwise Descreening Approximation failed to provide solvation energies for some proteins because of the computation of negative Born radii. It is quite possible that these errors were caused for the reasons described here. Interestingly, an implementation of the method of Qiu[98] in CHARMM,[133] also computed negative Born radii on a few occasions, even though that method makes use of a 'close contact function' that reduce the radii of nonbonded atoms that are very close to atom i.

In a molecular simulation, because polar hydrogens usually have a small collision diameter, they are the most likely to come into close contact with polar atoms in a fashion similar to figure 2.21, resulting in an overestimation of the Born radius of the polar hydrogen. Inspection of equation 1.55 easily shows that if the contribution of neighboring atoms j to the effective Born radii of atom i is overestimated, the effective Born radii can become negative. Our implementation of the Pairwise Descreening Approximation resets to α_i any negative value of B_i, explaining why PMFs involving PDAexp-2 could be computed entirely, even when some Born radii became negative.

A simple solution to reduce the risk of this occurring is to ensure that for one atom, the product of its intrinsic Born radius by its associated scaling factor is not larger than its van der Waals radius. In this fashion, overlaps such as those seen on figure 2.21 will become unlikely because they will be associated with a repulsive

non bonded energy. In practice, for the cases we have examined, we have found that optimizing the PDA scaling factors against the polarisation energy obtained by the numerical integration of the Born radii, along with offsets to the van der Waals radii smaller or equal to one, does not lead to cases similar to figure 2.21. Whether such situations could still occur in other systems, especially larger ones involving deeply buried atoms, is unknown.

2.4 Conclusion

Molecular mechanics studies are often faced with the problem that most force fields have been designed to study a limited class of molecules. As a result, parameters are often missing when one tries to study a new class of molecules. This situation is undesirable if a large number of drug-like molecules are to be studied by free energy simulations. By adopting a framework proposed in the literature, the risk of such situation occurring has been greatly reduced. The General Amber Force Field ensures that parameters will be available in the vast majority of the cases, while the use of the AM1/BCC method to derive atomic partial charges means that parameter generation will be fast and reliable.

To simplify the systems to be studied, the influence of the solvent is represented by a generalised Born surface area theory. A number of generalised Born surface area solvent models using the Pairwise Descreening Approximation have been derived by two different methods. In the first the scaling factors that are used to compensate for systematic errors in the Pairwise Descreening Approximation have been optimised along with all other parameters against experimental free energies of hydration of organic molecules. It is observed that the scaling factors sometimes adopt values that appear unphysical but that compensate for the other approximations in the generalised Born theory. With the second method, the scaling factors are optimised so that they compensate only for systematic errors of the Pairwise Descreening Approximation. The first method yields models that predict the hydration free energy of organic molecules more accurately than with the second method, explaining why this approach has been adopted in a number of previous studies.[117,118] However, the effective Born radii computed by these models

show larger deviations to the 'perfect' Born radii computed by Poisson-Boltzmann calculations. Subsequently, when applied to the computation of potentials of mean force, models derived with the first method sometimes fail to provide meaningful results because of large errors in the computation of Born radii. Results also show that good agreement with other PMF simulations reported in the literature does not necessarily require a good prediction of experimental hydration free energies. Instead, variations of the intermolecular interaction energy terms must be correctly balanced by variations of the GBSA solvation energy to yield a reliable potential of mean force. Because variations of the GBSA energy are controlled by the Born radii, it is essential that they are computed accurately. Therefore, efforts in the parameterisation of GBSA models using an approximate method to obtain Born radii should be directed toward an accurate computation of this property and not solely the prediction of experimental solvation free energies. While the computation of Potentials of Mean Force provide a stringent test of the quality of a generalised Born solvation model, comparisons against effective Born radii computed by Poisson-Boltzmann calculations may also be helpful in quickly identifying problematic sets of parameters, as illustrated in this study. Because some solvation models have been previously developed without considering these issues,[117,118] they should be further tested before being used to model the solvent screening of intermolecular interactions. Finally, it appears that the errors due to the PDA observed here can be reduced by ensuring that the product of the intrinsic Born radius of one atom by its associated scaling factor is smaller than its van der Waals radius. As a last note, while the set of parameters in the PDAnum series are clearly better than those in the PDAexp series over the range of PMFs tested, further modifications to the models may be needed to use them effectively in protein simulations because of the tendency of the PDA of Hawkins to under-estimate the effective Born radii of buried atoms.[123] Improvements over the PDA that deal with this problem have been suggested in the literature and their integration with the parameterised models reported here should be simple.[140,141]

Chapter 3

Efficient generalised Born models for Monte Carlo simulations

> "It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could safely be regulated to anyone else if machines were used"
>
> Gottfried Wilhelm von Leibniz

3.1 Introduction

One of the goals of the previous chapter was to parameterise a generalised Born model to use it in Monte Carlo simulations of protein-ligand complexes. While the generalised Born method is very efficient when used to represent solvent effects on small molecules, it quickly loses efficiency in the context of Monte Carlo simulations. Methods that address this issue are discussed in this chapter.

3.2 Generalised Born in a Monte Carlo Simulation

GBSA is widely used in the context of molecular dynamics simulations. For example, some interesting studies of GBSA molecular dynamics simulations of RNAs are discussed by Sorin et al.[142,143] while Felts et al. used GBSA molecular dynamics to study the potential of mean force of small peptides.[105] To date, few Monte Carlo GBSA simulations have been reported in the literature. The flexible

docking algorithm of Taylor et al. uses Monte Carlo moves and a GBSA model of water.[17] The Concerted Rotation with Angles (CRA) algorithm of Ulmschneider et al. uses novel Monte Carlo moves to fold peptides in a GBSA force field.[144,145] While the treatment of small peptides with a Monte Carlo GBSA method is still efficient compared to the explicit solvent alternative, or desirable in the case of Monte Carlo protein backbone moves, the method quickly loses its appeal as the system size increases. A Monte Carlo simulation of a biomolecule requires many more moves than molecular dynamics time steps because only portions of the system under study are updated at every move. Because most of the system does not change coordinates, and the force field terms are usually separable, it is generally sufficient to calculate only the change in energy of the part that has moved, which is very efficient. However, inspection of equation 1.55 shows that the Born radius of atom i depends on the position of every other atom j in the system. In turn, this means that the pairwise energies from equation 1.52 have the same dependency. As a result, the energy between atoms that did not move must be recomputed after every Monte Carlo move and a full GB energy calculation must be performed after every Monte Carlo move. For even a mid-sized protein the computational cost can be very high. This is not a problem in a molecular dynamics simulation because the total energy of the system is calculated after every step in any case.

The aim of this article is to introduce methods that can overcome the limitations of a standard GBSA implementation within the framework of a Monte Carlo simulation. This work is motivated by the availability of powerful Monte Carlo methods such as concerted rotations[144] or configurational bias for sampling protein systems,[146] and the efficiency that can be attained by combining them with a GBSA model.

3.3 Implementing a generalised Born force field suitable for free energy calculations

In typical relative binding free energy calculations, one ligand is converted into a closely related ligand. However, it is often the case that the two molecules do not have the same number of atoms. In this case, it is necessary to introduce dummy

atoms in one or two ligands such that one molecule can be converted completely into the other. At one end state ($\lambda = 0.0$ or $\lambda = 1.0$), these atoms do not interact at all with other particles, but they can be gradually transformed into interacting particles, such that at the other end of the perturbation they are fully interacting with the surrounding environment.

While dummy atoms at one end of the perturbation have zero charge, if their Born radii Bi are also zero, then very near to the end of the perturbation, dummies have a small but still substantial charge in a very small dielectric cavity. Such a charge has a huge electrostatic energy and the presence of this would lead to energetic instabilities. Therefore, the minimum intrinsic Born radius of an atom was set to 1.10 Å. This would mean that at the end states of a perturbation, dummy atoms would be (incorrectly) displacing a portion of dielectric. To avoid this artefact, dummy atoms are assigned a scaling factor of zero. As it can be verified by inspecting equation 1.55, this means that when the Born radii of non dummy atoms is calculated, the displacement of dielectric by a dummy atom is zeroed. When the effective Born radii of the dummy atoms are calculated, the algorithm returns a non zero value but since the charge on the dummy atom has been reduced to zero, the dummy atom does not contribute at all to the generalised Born energy. In order for the implementation to handle generic perturbations, the scaling factor and offsets to the intrinsic Born radii are treated as force fields parameters and are thus linearly interpolated between the end states of the perturbed system.

3.4 Selecting a test system and setup

In this work, the GBSA method was implemented in a modified version of the Monte Carlo package ProtoMS2.1.[147] Polarisation energies were computed using equation 1.52 and the Born radii were calculated with equation 1.55. The Surface Area calculations were implemented using the method of Shrake and Rupley[148] and a probe of 1.4 Å radius was used. The parameter set used for this GBSA model comes from a previous study ("PDAnum2" in this reference).[100]

To test the approximations introduced below we selected as a test case a set of protein-ligand relative binding free energy calculations which are shown in figure

3.1. The perturbations are typical of the mutations performed in a protein ligand binding free energy calculation and cover apolar to apolar (**1to2**), polar to apolar (**1to3** and **4to5**) and polar to polar (**4to6**) perturbations. The two different proteins considered exhibit a very different binding site. Neuraminidase has a polar, solvent exposed, binding site while cyclo-oxygenase-2 has a buried, fairly hydrophobic binding site. The test case should therefore represent a broad class of protein-ligand interactions that are studied by free energy perturbation methodologies. The binding mode of the inhibitors was inferred on the basis of a similar ligand complexed to a monomer of the N9 strain of influenza A (PDB code 1BJI)[149] or the PDB structure of murine COX-2 complexed to SC-558 (PDB code 1CX2).[150] When necessary, hydrogens were added to the crystallographic structure using the program reduce.[151] Sugars, co-factors, crystallographic waters and ions were removed. The protein was setup with the AMBER99 force field, inhibitors were setup with the GAFF force field and the atomic partial charges were derived using the AM1/BCC method[109] as implemented in the package AMBER8.[112] The system was energy minimised using the Sander module of AMBER8 and a generalised Born force field (the igb keyword was set to 1).[112] The backbone of the energy minimised protein was kept rigid for subsequent Monte Carlo simulation which were conducted with a modified version of the ProtoMS2.1 package.[147] To reduce the computational cost, only the protein residues that have one heavy atom within 15 Å of any heavy atom of the ligands were retained. The bond angles and torsions of the protein side chains within 10 Å of any heavy atom of the ligand and all the bond angles and torsions of the ligand were sampled during the simulation, with the exception of rings. The bond lengths of the protein and ligand were kept rigid. The total charge of the system was brought to zero by neutralizing lysine residues lying in the outer (frozen) part of the scoop (residues number 511 and 532 for COX-2, 432 and 273 for neuraminidase). The protonation state of the histidines was decided by visual inspection of the crystallographic structures. The resulting model of COX-2 had 155 residues and neuraminidase 145 residues. A 10 Å switched residue based cutoff was employed in all simulations. In the generalised Born simulations, a cutoff of 20.0 Å for the calculation of the Born radii

was applied.

Replica exchange thermodynamic integration[59,60] (RETI) was applied to these systems and the necessary ensemble of states were formed using Metropolis Monte Carlo sampling[39] at a temperature of 25 °C. In the RETI protocol, standard finite difference thermodynamic integration (FDTI)[38] is performed at each value of the coupling parameter λ (Δλ =0.001). Occasionally, moves that exchange system coordinates between replica i at $\lambda = A$ of energy E_A(i), and replica j at $\lambda = B$ of energy E_B(j) are attempted, subject to the RETI acceptance test described in chapter 1.

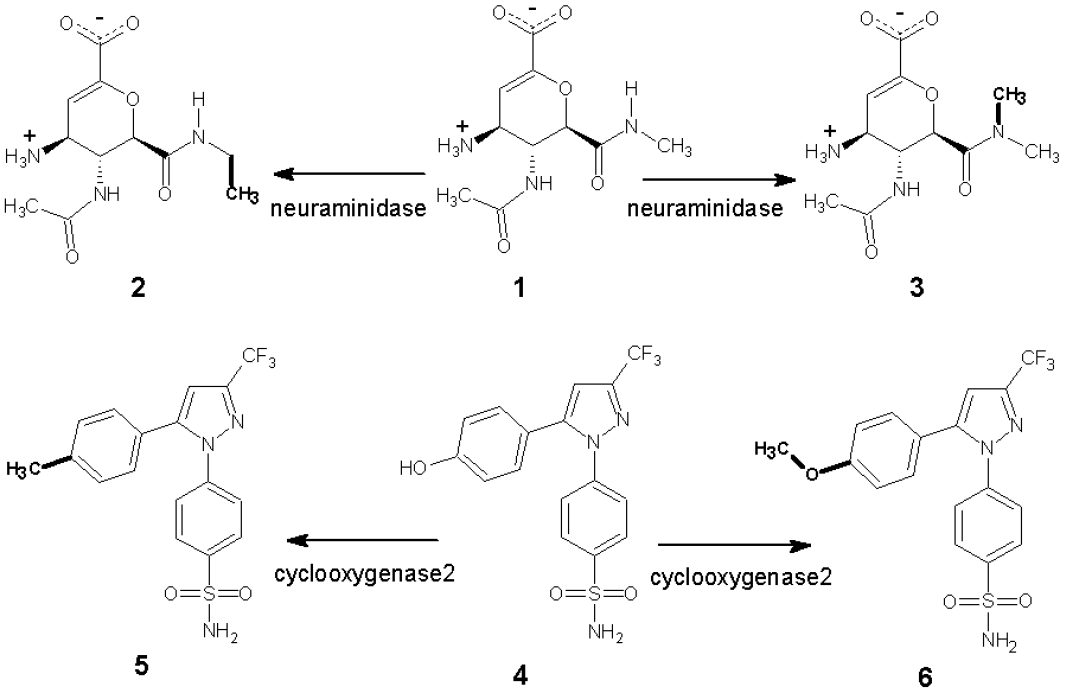


Figure 3.1: Representation of the ligands considered in this free energy study. For visual emphasis, the parts of the ligands being perturbed are highlighted by bold, straight lines.

Solute moves were attempted 10% of the time, with the remainder being protein side chain moves. In the unbound state, two thousand (K) moves of equilibration were performed before 200 K moves of data collection. In the bound state, the system was pre-equilibrated at one value of λ for 600 K moves. The resulting configuration was distributed over 12 values of the coupling parameter λ (0.00, 0.10, ..., 0.90, 0.95, 1.00) and further equilibration was performed for 100 K moves. Data were collected over the remaining 900 K moves. Replica exchange moves were attempted every 5 K configurations.

The error on the free energy gradients was calculated by taking the standard error of batch averages (size 1 K). The standard error of these averages was then integrated over the λ coordinate to yield the maximum error.

The speed up reported in the next sections are calculated as ratios of the time taken to complete a 1000 MC moves on the test systems between two particular simulation protocols.

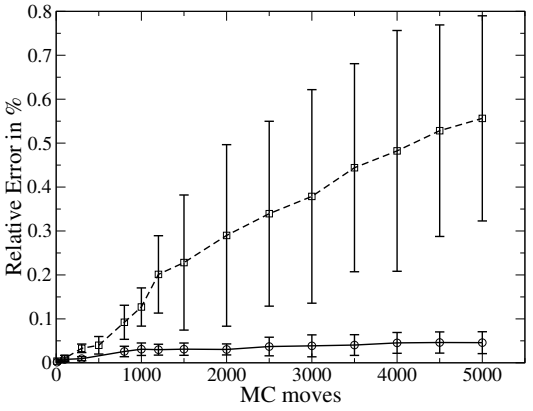
3.5 Approximated generalised Born Potential

A rigorous GB calculation means that the GB energy between all pairs of atoms must be recalculated after every move. However, the impact of a moving atom on the Born radius of a distant atom is small. We have therefore structured the implementation of the GB calculation such that the energy of a pair of atoms is recalculated only if the Born radius of either atom has changed by more than a specified threshold value after a MC move. A large number of pair interactions can be skipped in this fashion, resulting in a significant speed-up. In this implementation, only the necessary old and new GB energy pair terms are recalculated to update the total GB energy. This keeps additional memory requirements low and makes the method easily applicable to larger system.

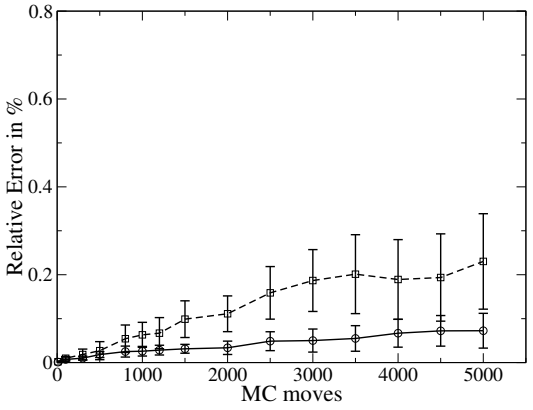
This approximation may have unwanted effects. For example, the total energy would not be completely conserved in a hypothetical Monte Carlo simulation in the NVE ensemble. However, the fact that useful results can be obtained from Molecular Dynamics simulations where fluctuations in the total energy are introduced because of errors in the integrator suggests that as long as the impact of the approximation is small, the resulting ensemble will closely mirror the correct one.

To assess the impact of this approximation, we have run series of short free energy simulations with a GBSA force field at different values of the threshold parameter. The plots reported in figure 3.2 are constructed by running a simulation for N steps with a specified value of the threshold parameter. The total energy of the last generated configuration is then recorded and compared to the value that is obtained by calculating the total energy with no approximation. This procedure is repeated ten times for a number of values of N. An arguably acceptable error

on the total energy would be about 0.1 percent since this is in the accepted range of MD integrator errors.[37] The systems run in neuraminidase are more sensitive to the approximation and at a high threshold value, the deviations become quickly large. The systems run on COX-2 appear much less sensitive. In both systems, at a threshold of 0.005 Å and up to 5000 MC moves the error is below 0.1 percent.



(a) Perturbation **1to3**, neuraminidase



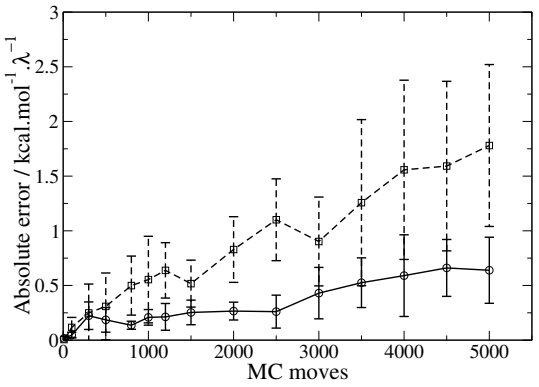
(b) Perturbation **4to5**, COX-2

Figure 3.2: Relative error in percent of the total energy as a function of the number of Monte Carlo moves for a threshold of 0.005 Å (black line, circles) and 0.05 Å (dashed line, squares). Each point is the average of ten different simulations and the error bar represents the associated standard error. Similar plots for the perturbations **1to2** and **4to6** are observed (data not shown).

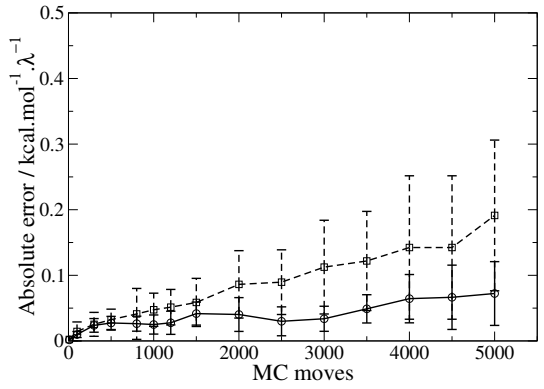
An added requirement for a free energy calculation is that the free energy gradients are not too sensitive to this approximation. In this application, we use a finite difference scheme and the gradients are formed from the difference in total energy at a value of λ - dλ and λ + dλ. In figure 3.3 the protocol described previ-

ously is applied to report the free energy gradients accumulated at $\lambda = 0.50$. The gradients are formed from the difference of two large numbers and are therefore more sensitive to small errors in the total energies. The free energy gradients of the perturbations run on neuraminidase are seen to be much more sensitive to the threshold than those run on COX-2. Because the binding site of neuraminidase is much more solvent exposed and comprises several polar amino acids, a rigorous treatment of the GB energy appear more important than for COX-2, where the buried, hydrophobic binding site is less sensitive to solvent effects. After 1000 MC moves, at a threshold of 0.001 Å, the average error on the free energy gradients is still 0.20 ± 0.07 kcal $\text{mol}^{-1}.\lambda^{-1}$ for **1to3**, while for **4to5** it is essentially negligible at high and low threshold values. From a computational perspective, the cost of a full GBSA calculation after 1000 approximate GBSA calculations is small. By updating completely the GB energy every 1K MC moves and with a suitably small threshold parameter, the errors on the total energy and the free energy gradients can be kept sufficiently low such that they have a small or negligible influence on the computed free energy.

To verify more rigorously the sensitivity of the systems to the threshold, a series of GBSA free energy simulations are run for each system with a varying threshold parameter. The impact of the threshold parameter on the calculated free energy is shown in figure 3.4. For the perturbations in COX-2, the calculated free energies are within the statistical error of the exact simulation over the range of thresholds studied. For the perturbations in neuraminidase, the free energy is more sensitive to the value of the threshold parameter and a high value of the threshold yields results that deviate significantly from the rigorous calculations; this is more marked for **1to3** than **1to2**, with results agreeing to within statistical sampling error from a threshold value of 0.002 Å or less. These results are consistent with the increased sensitivity of the free energy gradients to the threshold for the neuraminidase systems observed in figure 3.3. Figure 3.5 shows the speed-up relative to a full GBSA calculation. Because the computational expense is similar for the systems run on the same protein, speeds up are shown for **1to3** and **4to5** only. Even with a threshold as low as 0.001 Å, a considerable speed up is achieved because the Born radii of several protein atoms are insensitive to the displacement



(a) Perturbation **1to3**, neuraminidase



(b) Perturbation **4to5**, COX-2

Figure 3.3: Absolute error in the free energy gradients as a function of the number of Monte Carlo moves for a threshold of 0.001 Å (black line, circles) and 0.005 Å (dashed line, squares). Each point is the average of ten different simulations and the error bar represents the associated standard error. Similar plots for the perturbations **1to2** and **4to6** are observed (data not shown).

of a distant residue. On these systems and over the range of thresholds studied, the simulations run 2.4 to 3.8 times faster.

On the range of systems studies here, the influence of the threshold parameter on the calculated binding free energies has been shown to be negligible (COX-2) or minor (neuraminidase) and can be minimised by reducing sufficiently the threshold parameter, at the cost of additional simulation time. That a balance can be struck between speed-up and accuracy can prove useful. In applications where accuracy is important, almost rigorous calculations can be made with a sufficiently low threshold. On the other hand, less accurate calculations that could be useful in the context of fast free energy calculations could be run with a higher threshold.

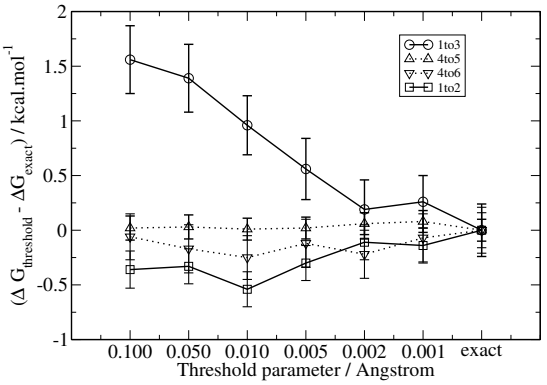


Figure 3.4: Influence of the threshold parameter on the calculated free energy for the selected perturbations in the bound state. The error bars represents the associated statistical error.

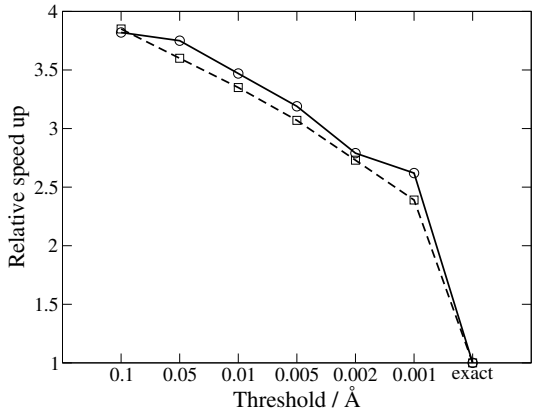


Figure 3.5: Relative speed up that can be achieved in the simulation of the perturbations **1to3** (circles, black line) and **4to6** (squares, dashed line) in the bound state with a varying threshold parameter. The speed up are based on the time taken to perform 1K Monte Carlo moves.

The optimum value of the threshold varies according to the system, but can be estimated rapidly by plotting the drifts in the free energy gradients for a series of short simulations. From the systems studied here, it appears that a good compromise would be achieved with a threshold of 0.002 Å or 0.005 Å.

3.6 Simplified Sampling Potential

3.6.1 Theory

A novel methodology to perform Monte Carlo simulations has recently been proposed by Gelb.[152] He shows that it is possible to perform a Monte Carlo simulation in which the potential energy is evaluated using an approximate, less expensive potential E_ζ than a more realistic potential E_π, and still form an ensemble of states

that are distributed according to the rigorous potential. The method is very powerful as in principle any kind of simplified potential/expensive potential combination can be devised.

The method can be briefly summarised as follow:

1. Start a simulation in state i
2. Performs N steps of standard Metropolis sampling with a simple potential E_ζ of limiting distribution ζ until a state j is reached
3. Set state i = state j with probability $\chi = (\pi_j \zeta_i)/(\pi_i \zeta_j)$. In this equation π_i and ζ_i are the probability of state i in the two different distributions π and ζ.
4. Accumulate any property of interest that is a function of the coordinates of state i
5. Return to 1 or terminate after a number of iterations

In essence, a standard Monte Carlo simulation is conducted for N steps with a potential chosen for its convenience (usually computational efficiency). However, because the probabilities of state i and j in the two distributions π and ζ generally differ it is necessary in step three to correct for any bias introduced by the potential E_ζ. This acceptance test makes sure that the ensemble formed during the simulation converges towards the distribution π instead of ζ. In the NVT ensemble, step three amounts to accepting state j according to

$$\exp\left[\beta\left([E_\pi(j) - E_\pi(i)] - [E_\zeta(j) - E_\zeta(i)]\right)\right] \geq rand(0,1). \tag{3.1}$$

The acceptance test is therefore based on the difference of the difference of energies of state i and j between the two potentials E_π and E_ζ. With this method, no statistics for the target ensemble π can be collected during step two and the number of data points accumulated is reduced compared to a traditional Monte Carlo simulation. This does not necessarily affect convergence because subsequent configurations in a Markov Chain are typically highly correlated and do not contribute

new information to the running average. That is to say, it is equally good to sample the distribution of interest less often if the samples are less correlated. The methodology is described in figure 3.6.

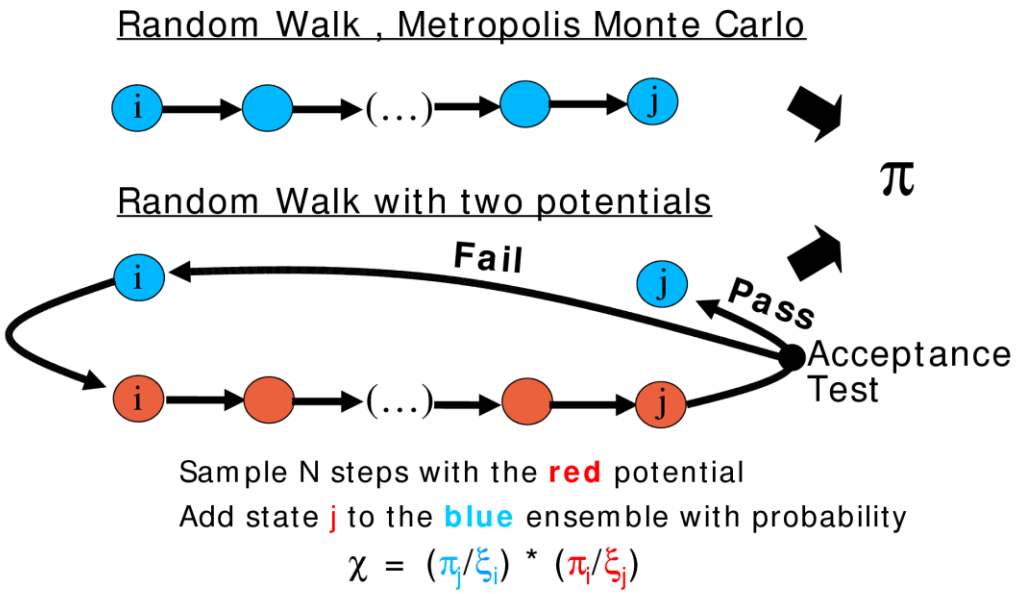


Figure 3.6: The simplified sampling methodology. In a standard random walk conducted with the Metropolis Monte Carlo algorithm, the successive generation of trial moves converges the density of states towards the equilibrium distribution π corresponding to the blue potential. The same results can be achieved by performing a random walk with the red potential and periodically considering the generated configurations, subject to the acceptance test χ, to calculate the thermodynamic properties of the blue potential.

To date, applications of this methodology have been reported by Hetenyi et al.[153,154] (who seem to have developed a similar method independently). Hetenyi reported a 3.0 to 4.7 speed up in the simulation of a Lennard-Jones fluid using for E_ζ a potential similar to E_π but with a shorter cutoff. Gelb reported similar results on a similar system.[152] In a second application by Hetenyi, a MC Ewald sum simulation of water was running 4.5 to 7.5 times faster using this methodology. This method has been employed by Iftimie et al. to perform ab-initio simulations using a classical potential.[155]

If the change of energy in going from state i to j is similar with the two Hamiltonians, the probability of accepting the configuration j will be close to unity. On the other hand, if the two potentials differ too much, then the acceptance rate will drop and the method will lose efficiency since all the steps performed with E_ζ have

been wasted. Therefore a good approximate potential E_ζ must be faster than, and yet very similar to E_π. This is of course difficult to achieve.

3.6.2 Application to a GBSA model

The complete application of the GBSA theory requires the calculation of a Surface Area (SA) dependent term to yield a solvation free energy. The inclusion of this term can be expensive and becomes significant once the GB calculations have been accelerated with the use of a threshold. For example, the simulation of **4to5** with a GB threshold of 0.005 Å is about 1.8 times slower once the SA calculations are enabled.

The fluctuations in the SA term are known to be small compared to the other energy components of the force field. This observation has led other workers to devise schemes where the SA term is only periodically updated.[17,145,156] While reasonable, this approximation is not completely rigorous. Other workers have developed faster, approximate SASA calculations schemes, but these algorithms do not calculate reliably the small changes in SASA associated with the small conformational changes observed between MC moves.[157]

However, the effect of the SA term can be rigorously included in the GBSA simulation by adopting a particular simplified sampling potential methodology. The simple potential E_ζ correspond to a GB simulation run without a SA term while the correct potential E_π includes SA calculations.

In addition, we consider other means to further speed up the calculations by adopting a less rigorous solvation model for the simplified potential. Here two different simplified solvation models are investigated, a distance dependent dielectric (DDD) force field and a simplified GB force field (fastGB). In the DDD force field the GB equations are replaced by a $\varepsilon(r) = 4r$ distance dependent dielectric. In the fastGB force field smaller cutoffs are applied: a residue based cutoff, Born radii cutoff and threshold of 6.0 Å, 12 Å and 0.05 Å respectively. In addition, since no statistics are collected with fastGB, it is not necessary to compute free energy gradients, which avoids the expensive GB energy calculations for the perturbed states. The rigorous potential is taken as a GB simulation with a threshold of 0.005 Å and

a rigorous SA calculation.

Table 3.1: Acceptance rate at the correction step and relative speed up for different combinations of potentials and number of moves M with the approximate potential[a]

M	DDD Rate [b]	DDD Speed up[c]	fastGB Rate [b]	fastGB Speed up[c]
1to3				
5	56.7 %	2.2	90.2 %	2.1
10	36.7 %	2.4	83.4 %	2.3
20	18.2 %	2.9	76.3 %	2.7
25	13.6 %	3.1	71.5 %	2.8
4to5				
5	65.4 %	2.2	91.2 %	2.0
10	47.7 %	2.5	87.1 %	2.3
20	27.9 %	2.9	80.8 %	2.5
25	22.5 %	3.1	78.2 %	2.7

a The results for **1to2** and **4to6** are similar to **1to3** and **4to5**.

b Average across all values of λ

c Relative to a GBSA simulation with a threshold of 0.005 Å

Table 3.1 lists the average acceptance rate of the correction step for the two different potentials as a function of the number of moves performed. The speed up compared to the rigorous GBSA simulation is also reported. The parameter M is the number of moves performed with the quick potential before attempting to add the generated configuration to the ensemble. As this quantity increases, the acceptance rate diminishes. As has been pointed out, a tradeoff must be made between computational efficiency and sampling efficiency.[152]

With the DDD model, the acceptance rate decreases faster than the speed up increases, and a short value of M is favored. Even after only 5 steps, the acceptance rate is only 55-65 %. For the systems in neuraminidase, the acceptance rate of the correction step is actually similar to using vacuum conditions (data not shown). This illustrate that the configurations favored by a GBSA force field are rather different from those preferred by a DDD force field.

With the fastGB model the decrease in the acceptance rate is more or less counterbalanced by the increase in speed up and no value of M is clearly favored. In addition, the acceptance rates are much higher and around 90 % for M equal to

5.

If we make the assumption that simulations run with the DDD and fastGB force field explore the configurational space at the same rate, then these results suggest that the combination of the two potentials fastGB/GBSA yields more efficient sampling than the DDD/GBSA combination.

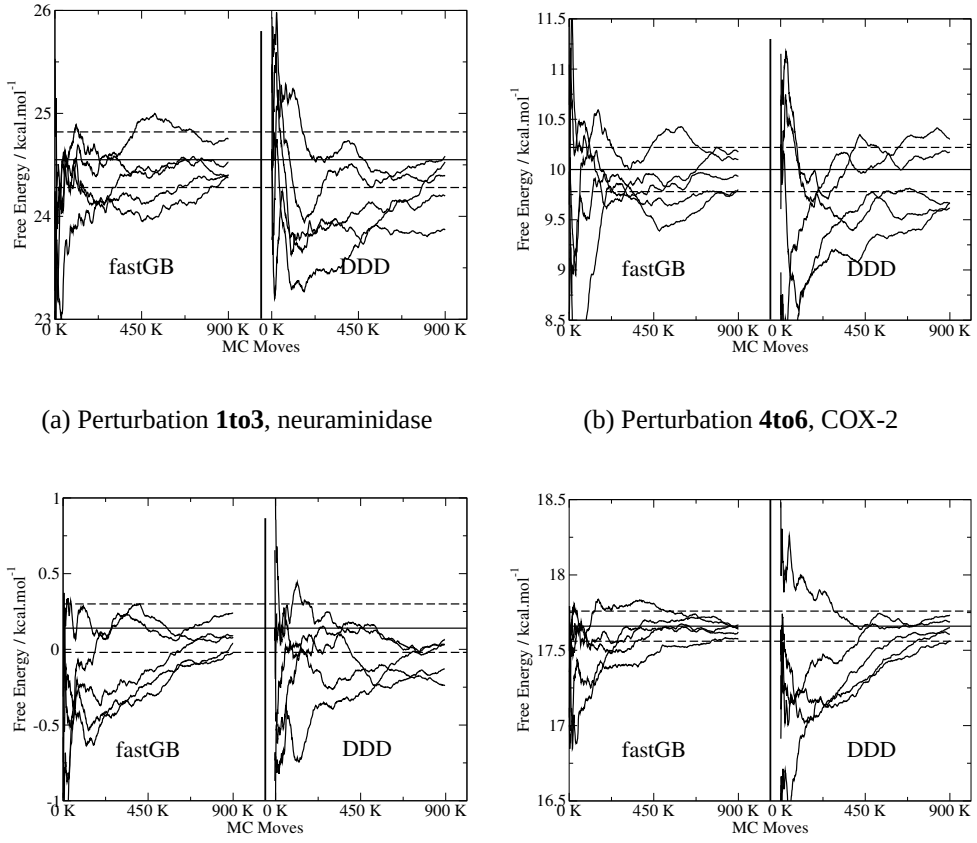
To demonstrate this more decisively, in figure 3.7 we investigate the convergence of the calculated free energies in the bound state for 5 independent simulations performed with the different protocols and a value of M set to 10. After 900K moves of data collection, almost all the fastGB simulations have converged to within the error bounds of the results obtained with GBSA 0.005 Å. With the DDD protocol, the results are more spread and several simulations are outside the error bounds. It is apparent that the fastGB protocol converges better the free energies than the DDD protocol for the same number of iterations.

Taken together, the results in table 3.1 and figure 3.7 suggest that a low acceptance rate for the correction step hinders convergence. The DDD simulations are slightly faster than the fastGB simulations. However, since the simulation results are much better converged with the fastGB protocol, it should be preferred over a DDD model. By combining the fastGB potential with the value of M set to 10 and with a GBSA 0.005 Å potential described in the previous section, an approximately 2.3 fold speed up over a standard MC simulation run with GBSA 0.005 Å can be achieved. The present results demonstrate that the simplified sampling potential methodology, applied here to increase the efficiency of generalised Born calculations for the first time, allows significant computational savings without additional approximations.

3.7 Conclusion

A novel methodology by which free energy calculations in a generalised Born framework can be made more efficient within Monte Carlo simulations has been proposed. It can be summarised as:

1. An approximate generalised Born potential in which the energy of the system is only partially updated after a MC move. The impact of this approx-



(a) Perturbation **1to3**, neuraminidase

(b) Perturbation **4to6**, COX-2

(c) Perturbation **1to2**, neuraminidase

(d) Perturbation **4to5**, COX-2

Figure 3.7: The convergence of the calculated free energies in the bound state using different sampling potentials. The estimated free energy is plotted as a function of the number of Monte Carlo moves performed. In each figure, on the left hand side, 5 independent simulations run with fastGB are shown. On the right hand side, 5 independent simulations run with DDD are shown. The horizontal line is the estimate of the free energy obtained after 900K moves with the potential GBSA 0.005 Å. The dashed lines represents the statistical error associated with this number.

imation on the calculated free energies can be made arbitrarily small at the expense of computational time by adjusting a single parameter.

2. Sampling driven by an inexpensive potential with a special Monte Carlo acceptance test that removes any bias in the distribution introduced by the cheap potential. This allows in addition the rigorous incorporation of surface area calculations at a minimum computational cost.

In table 3.2 timings for various combinations of these approximations on two of the four systems are reported. Protocol 4, which combines the two approxima-

tions is seven to eight times faster compared to protocol 1 which would correspond to a standard implementation of a GBSA force field. Protocol 4 is also only about 4.1-4.3 times slower than a simulation run in vacuum (protocol 5) which compares favorably with the typical efficiency of molecular dynamics GBSA simulations.[112] While the increased efficiency has been demonstrated on a free energy calculation, improvements in standard MC simulations could be sought with the same method.

Table 3.2: Time required to complete a block of 1K moves for selected approximations[a]

Protocol	Solvation model	Simplified Potential	Time **1to3**	Time **4to5**
1	GBSA exact	No	633.4 s	746.3 s
2	GBSA threshold 0.001 Å	No	241.8 s	312.0 s
3	GBSA threshold 0.005 Å	No	198.6 s	243.0 s
4	GBSA threshold 0.005 Å	fastGB M=10	79.8 s	104.2 s
5	Vacuum	No	19.5 s	24.1 s

a ProtoMS2.1 on a Pentium IV 2 GHz compiled with g77

It is important to recall that the loss of efficiency of the GBSA method when employed to simulate large systems is not observed in molecular dynamics simulations. However, there are several reasons which render Monte Carlo simulations in a GBSA force field desirable. First, the method allow the use of complex Monte Carlo moves such as RETI,[59] configurational biased,[146] or concerted rotations moves[144] that enhance significantly configurational sampling. These features are not available in a molecular dynamics simulation. Second, Monte Carlo simulations often yield converged free energies more efficiently. This is because the contribution of several unimportant degrees of freedom to the ensemble averages can be trivially removed by not sampling them. By contrast, molecular dynamics would require methods such as SHAKE[48] or positional restraints which add overhead to the potential energy function evaluation.

We may ask whether or not the partial rigidity of the system and the simplified treatment of solvation affects the accuracy of the calculated binding free energy. The simulation results can be compared with experimental figures by constructing a thermodynamic cycle, which requires the ligand perturbations in the unbound state to be performed. We stress that free energy calculations in the unbound state

are extremely rapid (on the order of a few minutes) and there is no need to introduce the methods developed to speed up the simulations in the bound state. The calculated binding free energies are listed in table 3.3. The implicit solvent protocol reproduces well the relative binding free energy of **1to3** but underestimates somewhat the binding free energy of the three other systems, although the trends are respected.

Table 3.3: Calculated and Experimental Binding Free Energies of the tested systems with the ApproxGB+SA protocol[a]

Perturbation	$\Delta\Delta G_{exp}$	$\Delta\Delta G_{bind}$	ΔG_{prot}	ΔG_{wat}
1to2	-1.6	-0.3 ± 0.3	0.1 ± 0.2	0.4 ± 0.1
1to3	-2.7	-2.7 ± 0.3	24.5 ± 0.3	27.2 ± 0.2
4to5	$<$ -4.6	-2.4 ± 0.1	17.7 ± 0.1	20.1 ± 0.1
4to6	$<$ -5.6	-3.1 ± 0.2	10.0 ± 0.2	13.1 ± 0.1

a The threshold was set to 0.005 Å. The figures are in kcal mol^{-1}. The experimental figures were taken from ref[78] for the neuraminidase inhibitors and from ref[58] for the COX-2 inhibitors.

The main emphasis of this chapter was to introduce a novel methodology to perform generalised Born Monte Carlo free energy calculations efficiently. A thorough investigation of the influence of the solvation model on the relative binding free energies will require the comparison of explicit and implicit solvent simulations on a larger set of systems. Such studies will be reported in the following chapters.

Chapter 4

Application to a protein-ligand system : cyclooxygenase-2

> "Certainly no subject or field is making more progress on so many fronts at the present moment, than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life it is that all things are made of atoms, and that everything that living things do can be understood in terms of jigglings and wigglings of atoms."
>
> Richard P. Feynman

4.1 Introduction

The application of generalised Born free energy techniques to the prediction of the relative binding free energies of a series of cyclooxygenase-2 (COX-2) selective nonsteroidal anti-inflammatory drugs is explored. The enzyme COX-2 was selected as a test case because its relatively hydrophobic, buried binding site provides a mean to assess the ability of a generalised Born model to treat properly the desolvation of the ligand and binding site. Simulation results are systematically compared to explicit solvent simulations performed with the same system setup and force field, typical empirical scoring functions and results published in the literature.

4.2 Presentation of the system

The enzyme cyclooxygenase is known to be responsible for the cyclooxygenation of arachidonic acid to prostaglandin PGG_2. PGG_2 is involved in the biosynthesis of numerous prostaglandins that possess analgesis, antipyretic and anti-inflammatory activity. At least two isoforms of this enzyme, COX-1 and COX-2 exist.[158] COX-1 is expressed in most cells and is thought to be responsible for the the production of prostaglandins that provide gastrointestinal tolerability. COX-2 only exist in inflammatory states. Traditional nonsteroidal anti-inflammatory drugs (NSAIDs) (aspirin, ibuprofen) inhibit both enzymes and as a result show ulcerogenic side effects. A second generation of NSAIDs has been shown to inhibit COX-2 selectively over COX-1.[159] One of the drugs in that series, celecoxib (compound **1**), available commercially under the name of Celebrex, shows a 375-fold selectivity of COX-2 over COX-1.[160] Celebrex is prescribed for acute pain, menstrual cramps, and the pain and inflammation of osteoarthritis and rheumatoid arthritis. Recently, the Food and Drug Administration has announced that based on preliminary studies from the National institute of Health, risks of cardiovascular events may be increased in patients receiving Celebrex. Shortly after other NSAIDs COX-2 selective inhibitors were removed from the market (Pfizer, valdecoxib, marketed as Bextra and Merck, rofecoxib, marketed as Vioxx).

The binding site of COX-2 is a long hydrophobic channel extending from the membrane region of the protein. A depiction of the interactions between important amino acids in the binding site and the brominated analogue of celecoxib compound **2** is presented in figure 4.1.

At the entrance of the channel the residues Arg120, Glu524, Tyr355 and Arg513 (bottom of figure 4.1) form a network of hydrogen bonds that acts as a gate to the binding site. The sulfonamide group of the ligand extends into a relatively polar pocket and makes interactions with Val523, Arg513, Gln192 and His90 (right corner of figure 4.1). The orientation of the sulfonamide group is ambiguous. Another X-ray structure from reference[150] shows the N-S-C-C torsion of the sulfonamide group flipped by 180 degrees. This is not surprising as it is not possible to distinguish between oxygen and nitrogen atom in the electron density plots obtained by

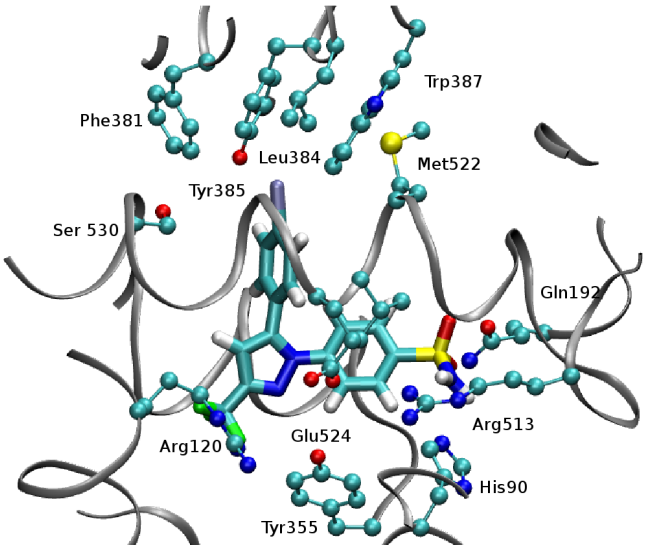


Figure 4.1: The binding site in the crystal structure of the brominated analogue of celecoxib SC-558 complexed with mouse cyclooxygenase-2.[150] The hydrogen atoms on the amino acid sidechains and some residues mentioned in the text are not shown for clarity.

X-ray crystallography. In the structure of figure 4.1 one of the oxygen atom of the sulfonamide group makes unfavourable electrostatic interactions with the amide carbonyl oxygen of Gln192 and the backbone carbonyl oxygen of Ser353 and one hydrogen bond with an amide hydrogen atom is not formed. In the alternative conformation where the sulfonamide group is rotated, water molecules can donate an hydrogen bond to the amide hydrogen atom and NH-O hydrogen bonds are formed with Gln192 and Ser353. Free energy perturbation studies from Price et al. predict this alternative conformation to be favoured over the conformation present in the X-ray structure by over 4.5 kcal mol^{-1}.[58]

The 5-aryl ring of SC-558 extends into a hydrophobic pocket on the top of figure 4.1 and makes hydrophobic contacts with Phe381, Leu384, Tyr385, Trp387, Met522. The range of binding affinity of the series of inhibitors presented in reference[159] is modulated by subsitution of the bromine atom in SC-558 by various hydrophobic and polar substituents. The common scaffold of these inhibitors is presented in figure 4.2. Their activity against COX-2 is shown in table 4.1.

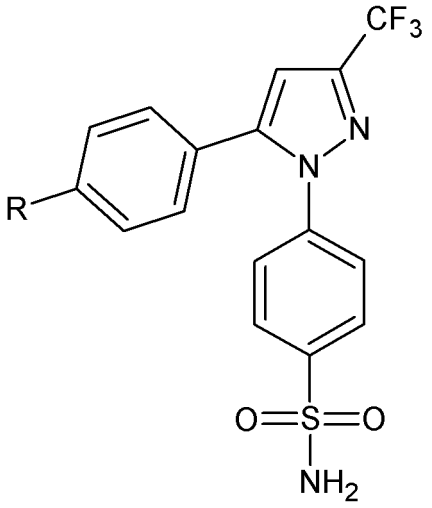


Figure 4.2: Structure of the series of celecoxib analogues.

Table 4.1: Experimental activity of a series of celecoxib analogues against COX-2

Compound	R	IC_{50} (μM)
1	CH_3	0.040
3	CH_2CH_3	0.86
4	CH_2OH	93.3
5	SCH_3	0.009
6	OCH_3	0.008
7	CF_3	8.23
8	OH	> 100
9	Cl	0.01
10	F	0.041
11	H	0.032

The range of measured binding affinities spans 5 orders of magnitude, from low nanomolar to hundreds of micromolar. Hydrogen bonding (**4**, **8**) and electron withdrawing groups are shown to reduce binding affinity (**7**) for COX-2 while electron-donating groups increase affinity (**5**, **6**). Increasing steric bulk also decreases affinity (**3**).

From a computational perspective, the high structural similarity of the compounds in this series and the large span of inhibition constants make this system ideal for study by free energy simulation methodologies. Furthermore the features of the binding site of COX-2 (buried, hydrophobic) make it an interesting test case

for the generalised Born methodology.

4.3 System setup and simulation protocols

The PDB structure of murine COX-2 complexed to SC-558 was selected as a starting point for this study (PDB code 1CX2). Hydrogen atoms had already been assigned by the crystallographers.[150] Ligand **3** was modelled in the complex on the basis of the binding mode of **2**. As noted in the previous section, previous theoretical studies and crystallographic evidences have pointed out that the sulfonamide moiety of SC-558 in the crystallographic structure of 1CX2 has been misplaced. The N-S-C-C torsion around this functional group was rotated to position it to interact favourably with neigbouring residues. A nearby heme was removed as it is not involved in any direct interactions with the binding site and its inclusion would have required the derivation of AMBER parameters for the heme group. The protonation state of histidines was selected by visual inspection and resulted in the assignment of δ-tautomers for His 90, 95, 133, 204, 207, 214, 226, 232, 242, 278, 309, 320, 351, 356, 386, 388 and 417. The protein was setup with the AMBER99 force field, inhibitors were setup with the GAFF force field and the atomic partial charges were derived using the AM1/BCC method[109] as implemented in the package AMBER8.[112] The system was energy minimised using the Sander module of AMBER8 and a generalised Born force field (the igb keyword was set to 1, which corresponds to the default generalised Born force field in AMBER8).[112] The backbone of the energy minimised protein was kept rigid for subsequent Monte Carlo simulations which were conducted with a modified version of the ProtoMS2.1 package.[147] To limit the computational cost, only the protein residues that have one heavy atom within 15 Å of any heavy atom of compound **3** were retained. The resulting protein scoop consisted of 155 residues. For the explicit solvent simulations, the complex was hydrated by a sphere of TIP4P water molecules[138] of 22 Å radius and centred near the geometric centre of the ligand. To prevent evaporation, a half-harmonic potential with a 1.5 kcal $mol^{-1}Å^{-1}$ constant was applied to water molecules whose oxygen atom distance to the ligand centre of geometry was

greater than 22 Å. A similar sphere of water was employed to solvate the ligands in the unbound state. The bond angles and torsions of the protein side chains within 10 Å of any heavy atom of the ligand and all the bond angles and torsions of the ligand were sampled during the simulation with the exception of rings. The bond lengths of the protein and ligand were kept rigid. The total charge of the system was brought to zero by neutralizing lysine residues lying in the outer (frozen) part of the scoop (residues 511 and 532). A 10 Å residue based cutoff was employed in all simulations. In the generalised Born simulations, a cutoff of 20.0 Å for the calculation of the Born radii was applied. To increase the efficiency of the generalised Born simulations, the techniques developed in chapter 3 were employed. In particular, a threshold of 0.005 Å to the recalculation of the Born energy was applied. A simplified sampling potential with a residue based cutoff, Born radii cutoff and threshold of 6.0 Å, 12 Å and 0.05 Å respectively was adopted and generated configurations were considered for the calculation of thermodynamic properties every 10 steps of this potential. A series of perturbations that transform one ligand into another in the series of celecoxib analogues was devised. In some instances, there is uncertainty as to how the substituent R should be positioned in the binding site. In this situation, observations from Price et al.[58] were used to model the ligand in the correct conformation.

Replica exchange thermodynamic integration[59] was applied to these systems and the necessary ensemble of states were formed using Metropolis Monte Carlo sampling[39] at a temperature of 25 °C. For the explicit solvent simulations in the bound state, solvent moves were attempted with a probability of 85.7%, protein side chain move with a probability of 12.8% and solute move with a probability of 1.4%. In the unbound state, solvent moves were attempted 98.4% of the time. Preferential sampling was used to increase the convergence of the calculated free energy.[40] The solvent was equilibrated for 20 million (M) configurations to remove any repulsive contact with the solute(s). The system was then equilibrated in one end state (typically corresponding to the largest ligand) for 20M further moves where solute, protein and solvent moves were attempted. The resulting configuration was distributed over 12 values of the coupling parameter

λ (0.00,0.10,...,0.90,0.95,1.00) and equilibrated for 10M moves before collecting statistics for 30M moves. In addition to the standard set of moves, replica exchange moves were attempted every 200 K configurations.

In the implicit solvent simulations, solute moves were attempted 10% of the time, with the remainder being protein side chain moves. In the unbound state, two thousand (K) moves of equilibration were performed before 200 K moves of data collection. In the bound state, the system was pre-equilibrated at one value of λ for 600 K moves. The resulting configuration was distributed over the 12 values of λ and further equilibration was performed for 100 K moves. Data was collected over the remaining 900 K moves. Replica exchange moves were attempted every 6 K configurations.

The error on the free energy gradients was calculated by taking the standard error of batch averages (size 3 K for the implicit solvent simulations in the bound state, 2 K in the unbound state and 200 K for the explicit solvent simulations in the bound and unbound state). The standard error of these averages was then integrated over the λ coordinate to yield the maximum error. This method will overestimate the statistical error, but the statistical error typically underestimate the precision of a free energy calculation.

The celecoxib analogues were also scored using common empirical scoring functions available in the literature. The program GOLD[12] was used to obtain scores with the scoring function Goldscore,[21] Chemscore[19,20] and Astex Statistical Potential (ASP).[22] The compounds were scored on the basis of the modelled binding mode in the energy minimised crystallographic structure of COX-2 used in the free energy study. To avoid artifacts due to use of a different force field, the ligands were locally minimised according to each scoring function, prior to scoring.

4.4 Explicit solvent simulations results

The calculated relative binding free energies with the explicit solvent protocol for the series of celecoxib derivatives are shown in table 4.2. The results are in good agreement with the experimental data.

Table 4.2: Comparison between experimental and calculated relative binding free energies and relative solvation free energies with the explicit solvent protocol[a]

Pert	Exp[b]	$\Delta\Delta G_{bind}$	$\Delta\Delta G_{solv}$	ΔG_{prot}	ΔG_{wat}	ΔG_{vac}
1t3	1.82	2.24 ± 0.36	0.58 ± 0.26	3.56 ± 0.27	1.32 ± 0.24	0.74 ± 0.09
4t3	-2.78	-2.18 ± 0.45	6.93 ± 0.38	0.8 ± 0.26	2.98 ± 0.37	-3.95 ± 0.09
1t7	3.16	3.9 ± 0.25	0.39 ± 0.24	20.71 ± 0.11	16.81 ± 0.23	16.42 ± 0.07
8t1	$<$ -4.64	-2.9 ± 0.37	4.39 ± 0.35	15.03 ± 0.15	17.93 ± 0.34	13.54 ± 0.08
8t9	$<$ -5.46	-3.49 ± 0.31	5.20 ± 0.29	17.12 ± 0.11	20.61 ± 0.29	15.41 ± 0.02
10t9	-0.84	-1.33 ± 0.18	-0.08 ± 0.16	-0.19 ± 0.08	1.14 ± 0.16	1.22 ± 0.01
11t10	0.15	0.01 ± 0.17	0.95 ± 0.16	-3.47 ± 0.05	-3.48 ± 0.18	-4.43 ± 0.02
11t8	$>$ +4.77	1.66 ± 0.29	-4.49 ± 0.27	-21.47 ± 0.10	-23.13 ± 0.27	-18.64 ± 0.02
3t5	-2.7	-1.7 ± 0.44	-0.46 ± 0.34	-3.19 ± 0.36	-1.49 ± 0.26	-1.03 ± 0.22
5t6	-0.07	-1.75 ± 0.52	-1.18 ± 0.64	-6.56 ± 0.38	-4.81 ± 0.36	-3.63 ± 0.53
8t6	$<$ -5.59	-3.68 ± 0.75	4.03 ± 0.73	9.35 ± 0.39	13.03 ± 0.64	9.00 ± 0.35

a Figures in kcal mol^{-1}. XtY means that compound X was perturbated into compound Y. $\Delta\Delta G_{bind}$ is the relative binding free energy. $\Delta\Delta G_{solv}$ is the relative solvation free energy. ΔG_{prot} is the free energy difference in the protein environment. ΔG_{wat} is the free energy difference in the aqueous environment. ΔG_{vac} is the free energy difference in vacuum.

b Relative free energies are calculated using the formula $\Delta\Delta G = \Delta G_2 - \Delta G_1 = RTln(K_1/K_2)$ with the approximation that the ratio of the IC_{50} is equal to the ratio of the dissociation constants.[161]

Two different cycles were closed for the binding free energies and solvation free energies and the resulting hysteresis is shown in figure 4.3. The hysteresis is low in both cases, particularly if we consider that the cycles involve 4 or 5 steps. This indicates that the simulation results should be well converged.

Table 4.3 shows the free energy difference of the ligands with respect to compound **1** (celecoxib). The mean unsigned error (MUE) is found to be 0.76 kcal mol^{-1}.

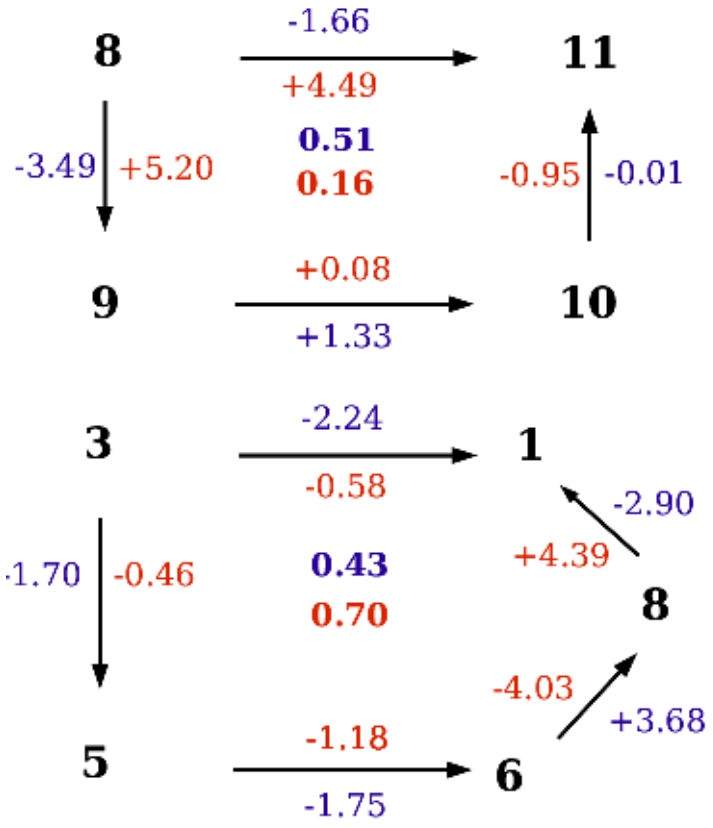


Figure 4.3: Explicit solvent protocol: closure of two thermodynamic cycles for the calculated relative binding (blue) and solvation (red) free energies of the celecoxib analogues. All the figures are in kilocalories per mole.

Table 4.3: Explicit solvent protocol: the experimental and calculated binding free energies with respect to celecoxib **1**.[a]

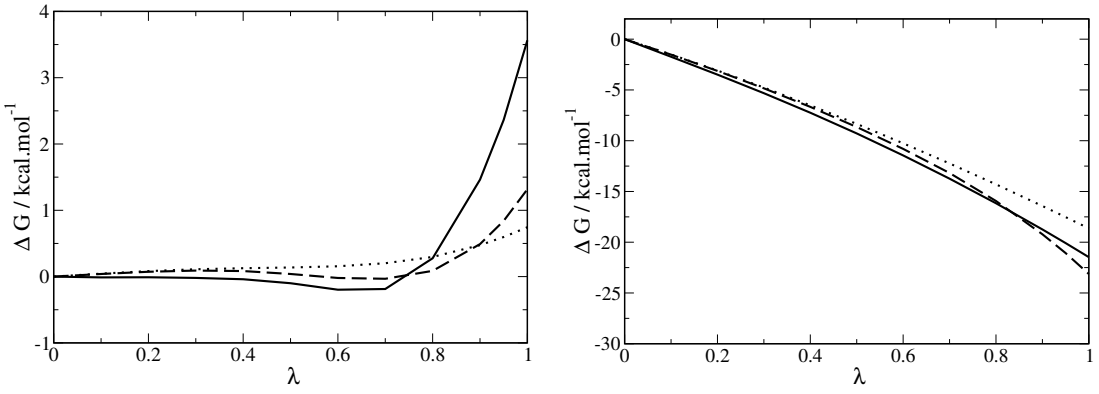
Compound	Perturbation pathway[b]	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
6	[**1t8**+**8t6**] ; [**1t3**+**3t5**+**5t6**]	-0.99 $\pm$ 0.81	-0.95
5	[**1t3**+**3t5**]	0.54 $\pm$ 0.57	-0.88
9	[**1t8**+**8t9**]	-0.58 $\pm$ 0.48	-0.82
11	[**1t8**+**8t9**+**9t10**+**10t11**];[**1t8**+**8t11**]	0.99 $\pm$ 0.51	-0.13
1		0	0
10	[**1t8**+**8t9**+**9t10**];[**1t8**+**8t11**+**11t10**]	1.00 $\pm$ 0.51	0.01
3	**1t3**	2.25 $\pm$ 0.36	1.82
7	**1t7**	3.90 $\pm$ 0.25	3.15
4	[**1t3**+**3t4**]	4.42 $\pm$ 0.58	4.59
8	**1t8**	2.90 $\pm$ 0.37	4.63

a Figures in kcal mol^{-1}.

b Figures obtained by summing free energy changes over different perturbations, and in some cases, averaging over two different pathways.

The free energy profiles for the transformation of **1** to **3** and **11** to **8** are shown in figure 4.4. The free energy changes over the coupling parameter are seen to be smooth. The free energy profile for the perturbation **1** to **3** indicates that growing an extra methyl group is initially slightly favourable in the binding site of COX-2. However, at around a λ value of 0.7 the free energy in the bound state increases more rapidly than in the unbound state, and ultimately the ethyl analogue **3** is less favoured than celecoxib **1**. This reflects the steric restriction that the larger ethyl group experiences in the binding site.

In the perturbation of the unsubstituted derivative **11** into a hydroxy group compound **8**, the free energy profiles in the protein, water and vacuum are very similar until about λ 0.5 after which the free energy decreases rapidly in water but somewhat less in the binding site, resulting in the hydroxy group being less stable than the unsubstituted derivative. This difference is due to the inability of the hydroxy group to donate its hydrogen bond in the binding site of COX-2. In water this can of course be accomplished easily.



(a) (b)

Figure 4.4: Explicit solvent protocol: the change in free energy along the coupling parameter. The solid line represents the change in COX-2, the dashed line represents the change in water and the dotted line represent the change in vacuum.
(a) Perturbation of **1** to **3** (b) Perturbation of **11** to **8**

A useful way to analyse the predictive power of the explicit solvent simulations is to calculate predictive indices for this series of compounds. The method of the predictive indices has been proposed by Pearlman et al.[28] to measure the ability of a binding free energy prediction method to rank a series of inhibitors in their order

of affinity. The predictive indices are calculated as follows:

$$PI = \frac{\sum_{j>i}\sum_i w_{ij}C_{ij}}{\sum_{j>i}\sum_i w_{ij}} \tag{4.1}$$

with

$$w_{ij} = |E(j) - E(i)| \tag{4.2}$$

and

$$\begin{aligned} C_{ij} &= -1 \quad if \quad \frac{E(j)-E(i)}{P(j)-P(i)} < 0 \\ &= +1 \quad if \quad \frac{E(j)-E(i)}{P(j)-P(i)} > 0 \\ &= 0 \quad if \quad P(j)-P(i) = 0 \end{aligned} \tag{4.3}$$

Where E(i) is the experimental binding free energy of compound i and P(i) is the predicted binding free energy (or some score) of compound i. This index ranges from -1 to +1 depending on how well the predicted ranking matches the experimental ordering. A value of +1 indicates perfect prediction, a value of -1 indicates that predictions are always wrong and a value of 0 arises from predictions that are completely random. The predictive index method essentially considers each pair of compounds i and j in turn. Large differences in binding free energies will have a large weight w_{ij} and successfully predicting which of the two compounds is the most potent will provide a large positive contribution to the final PI. If i and j have a small difference in binding free energy, an incorrect prediction of the most potent binder will have a minor impact on the predictive index.

By applying equations 4.1 to 4.3 a predictive index of 0.96 is found for the explicit solvent simulations reported above. This is an excellent result, demonstrating impressive predictions on this set of ligands. Figure 4.5 summarises the results of the explicit solvent simulations. The coefficient of determination (calculated from figure 4.5 and ignoring **1**) has a value of 0.85 which suggests a respectable agreement between experiment and theory.

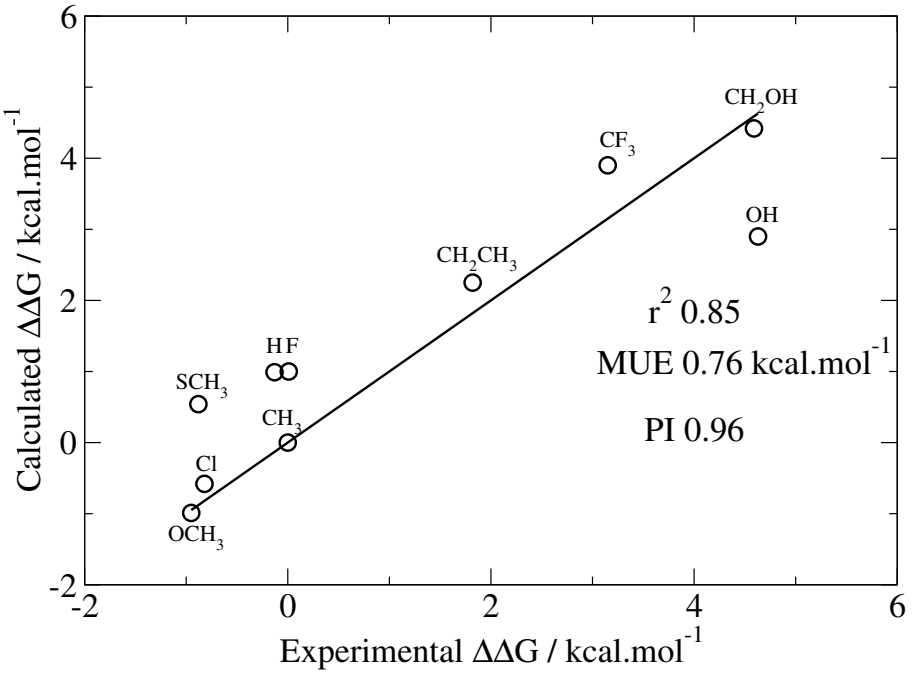


Figure 4.5: Summary of the explicit solvent protocol results. The binding free energies are relative to compound **1**.

The same set of compounds has been studied by Price et al.[58] and comparison of their results with ours is of interest to validate the current protocol. Price et al.[58] reported results in somewhat better agreement with experimental data, with a mean unsigned error of 0.4 kcal mol^{-1} and a coefficient of determination of 0.96. The origins of the differences between the two studies most likely arise from two factors. We have employed the AMBER99[34] force field for the protein and the GAFF[107] force field with AM1/BCC atomic partial charges for the ligands.[108] Price et al. used the OPLS/AA[33] force field for the protein and the OPLS/AA force field and CM1A atomic partial charges for the ligand.[162] Another difference lies in the system setup. In our simulation, no water molecules were present in the active site of COX-2 while, depending on the perturbation studied, one or two water molecules were present in the simulations of Price et al. There is no structural evidence supporting the presence of water molecules in this buried, hydrophobic binding site and Price et al. could not rule out the possibility that the water molecule was an artefact of the procedure used to build the water cap in their simulations. The experimental binding free energy difference between **10** and **11** is -0.15 kcal mol^{-1}. In the mutation of **10** to **11** Price et al. reported a binding

free energy of -1.24 kcal mol^{-1} in the presence of two water molecules in the binding site. When a water molecule bridging interactions between the para substituent on the 5-aryl ring of the celecoxib derivatives and Met522 was manually removed from the binding site and the calculation run again, a binding free energy of +1.52 kcal mol^{-1} was found. The presence of water molecules in the binding site can therefore affect significantly the relative binding free energies. Our simulation yield a binding free energy difference of -0.01 kcal mol^{-1}, in good agreement with experimental results.

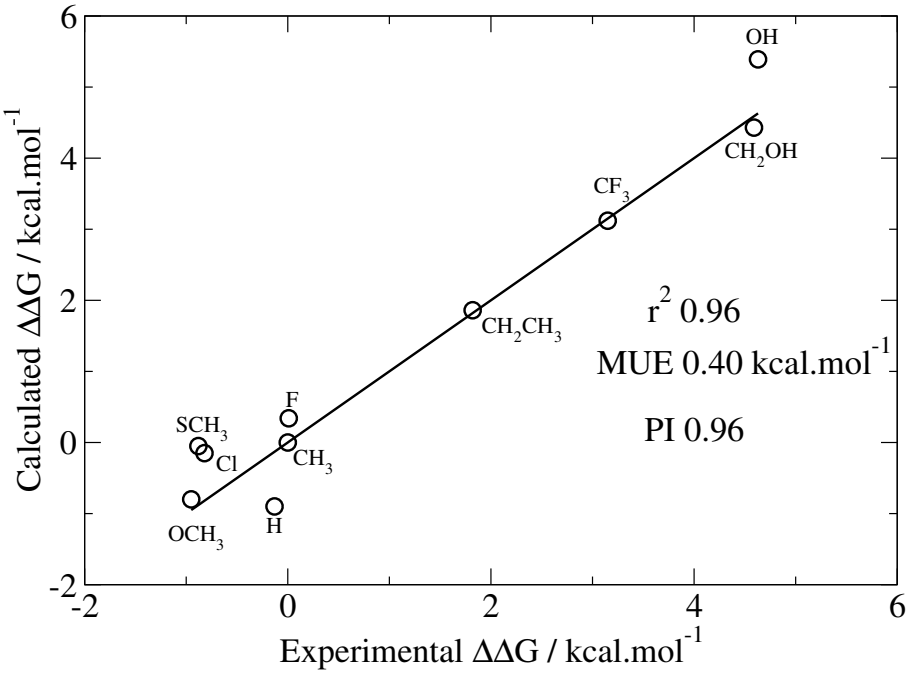


Figure 4.6: Results reported by Price et al.[58]

The results of Price et al. are reproduced in figure 4.6 for comparison. Their results are clearly in more quantitative agreement with experiment than ours. It is worth mentioning that the calculated predictive index is identical to ours, meaning that our simulations predict the ordering of the ligands as well.

4.5 Generalised Born simulations results

The calculated relative binding free energies with the implicit solvent protocol for the same series of perturbations is presented in table 4.4. The results are in reasonable agreement with experimental data and match closely the trends observed with the explicit solvent simulations. The free energy change in vacuum, necessary to

calculate the relative solvation free energies, was taken from the figures reported in table 4.2.

Table 4.4: Comparison between experimental and calculated relative binding free energies and relative solvation free energies[a] with the implicit solvent protocol

Pert	Exp[b]	$\Delta\Delta G_{bind}$	$\Delta\Delta G_{solv}$	ΔG_{prot}	ΔG_{wat}
1t3	1.82	1.92 ± 0.31	0.36 ± 0.13	3.02 ± 0.30	1.10 ± 0.09
4t3	-2.78	-1.03 ± 0.23	7.87 ± 0.11	2.89 ± 0.22	3.92 ± 0.07
1t7	3.16	2.24 ± 0.18	-1.30 ± 0.11	17.36 ± 0.16	15.12 ± 0.08
8t1	$<$ -4.64	-2.48 ± 0.19	6.55 ± 0.11	17.61 ± 0.17	20.09 ± 0.08
8t9	$<$ -5.46	-3.07 ± 0.07	5.86 ± 0.03	18.20 ± 0.07	21.27 ± 0.02
10t9	-0.84	-1.15 ± 0.07	-0.38 ± 0.01	-0.31 ± 0.07	0.84 ± 0.01
11t10	0.15	-1.08 ± 0.05	0.16 ± 0.01	-5.35 ± 0.05	-4.27 ± 0.01
11t8	$>$ +4.77	0.42 ± 0.07	-6.05 ± 0.03	-24.27 ± 0.07	-24.69 ± 0.02
3t5	-2.7	-0.94 ± 0.45	-1.18 ± 0.31	-3.15 ± 0.39	-2.21 ± 0.22
5t6	-0.07	-1.98 ± 0.68	-1.60 ± 0.76	-7.21 ± 0.41	-5.23 ± 0.54
8t6	$<$ -5.59	-3.23 ± 0.54	4.05 ± 0.50	9.82 ± 0.40	13.05 ± 0.36

a Figures in kcal mol^{-1}. XtY means that compound X was perturbated into compound Y. $\Delta\Delta G_{bind}$ is the relative binding free energy. $\Delta\Delta G_{solv}$ is the relative solvation free energy. ΔG_{prot} is the free energy difference in the protein environment. ΔG_{wat} is the free energy difference in the aqueous environment. ΔG_{vac} is the free energy difference in vacuum.

b Relative free energies are calculated using the formula $\Delta\Delta G = \Delta G_2 - \Delta G_1 = RTln(K_1/K_2)$ with the approximation that the ratio of the IC_{50} is equal to the ratio of the dissociation constants.[161]

The closure of the same two cycles shown in figure 4.7 shows that the hysteresis is even lower than in the explicit solvent simulations. This is not entirely surprising as the statistical error associated with individual generalised Born simulations is lower than for their explicit solvent counterpart. Most of this difference arise from the unbound state as simulations in pure implicit water can sample thoroughly all the degrees of freedom of the system more easily than in an explicit solvent simulation.

Table 4.7 shows the free energy difference of the ligands with respect to compound **1**.

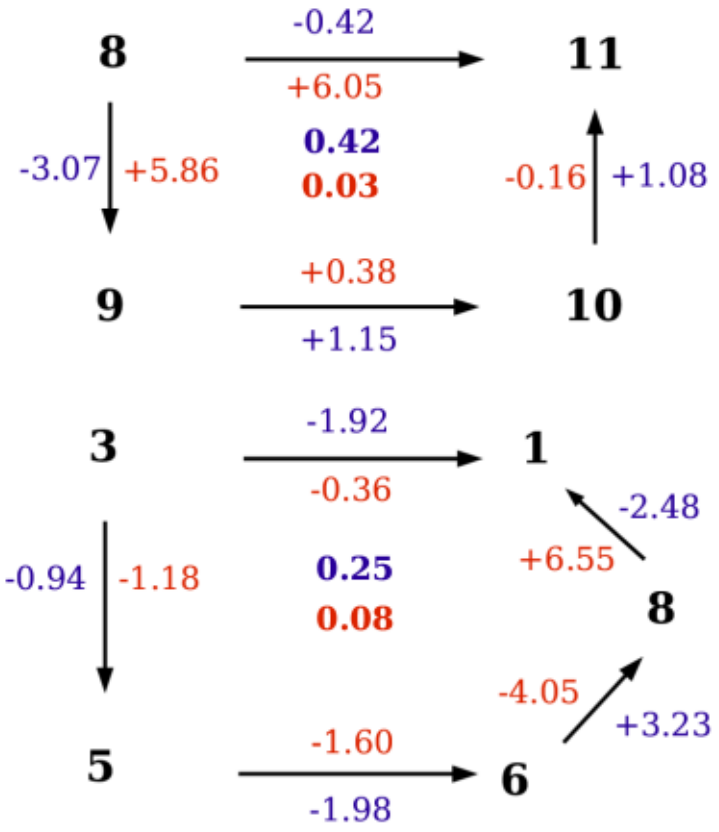


Figure 4.7: Implicit solvent protocol: the closure of two thermodynamic cycles for the calculated relative binding (blue) and solvation (red) free energies of the celecoxib analogues. All the figures are in kilocalories per mole.

Table 4.5: Implicit solvent protocol: the experimental and calculated binding free energies with respect to celecoxib **1**.[a]

Compound	Perturbation pathway[b]	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
6	[**1t8**+**8t6**] ; [**1t3**+**3t5**+**5t6**]	-0.88 $\pm$ 0.72	-0.95
5	[**1t3**+**3t5**]	0.98 $\pm$ 0.55	-0.88
9	[**1t8**+**8t9**]	-0.59 $\pm$ 0.20	-0.82
11	[**1t8**+**8t9**+**9t10**+**10t11**];[**1t8**+**8t11**]	1.84 $\pm$ 0.21	-0.13
1		0	0
10	[**1t8**+**8t9**+**9t10**];[**1t8**+**8t11**+**11t10**]	0.76 $\pm$ 0.21	0.01
3	**1t3**	1.92 $\pm$ 0.31	1.82
7	**1t7**	2.24 $\pm$ 0.18	3.15
4	[**1t3**+**3t4**]	2.95 $\pm$ 0.39	4.59
8	**1t8**	2.48 $\pm$ 0.19	4.63

a Figures in kcal mol^{-1}

b Figures obtained by summing free energy changes over different perturbations, and in some cases, averaging over two different pathways

The results for the generalised Born simulations on this set of ligands are summarised in figure 4.8. The MUE at 1.08 kcal mol^{-1} is higher than the one obtained for the explicit solvent simulations and accordingly the coefficient of determination has dropped to 0.70. However the calculated predictive index stands at 0.96 and is identical to that obtained with the other protocols. Thus, while the predicted binding free energies deviate more from their experimental figure, the ordering of the compounds is as good as with the previous explicit water protocol.

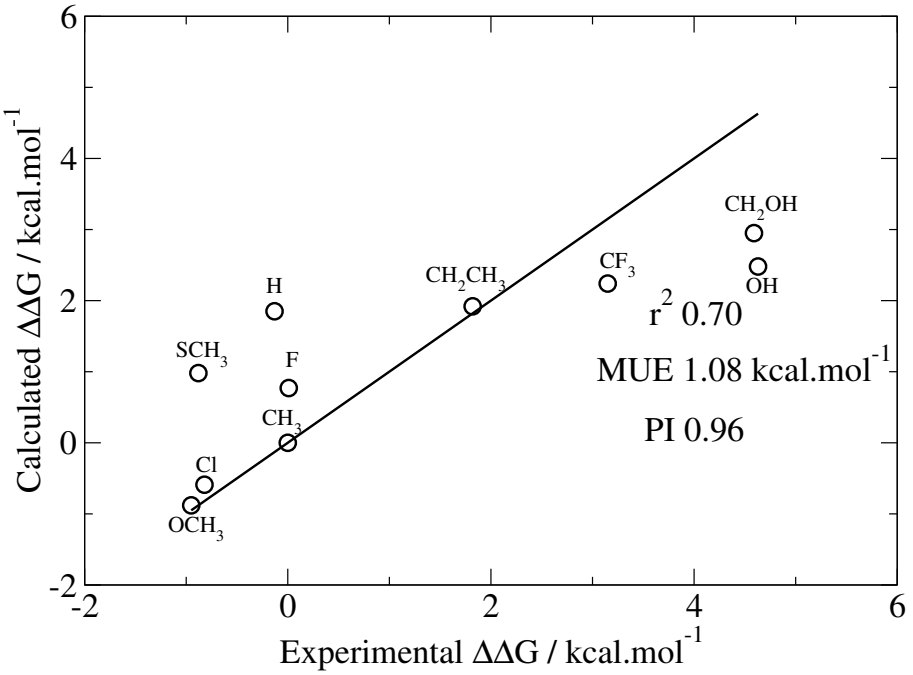
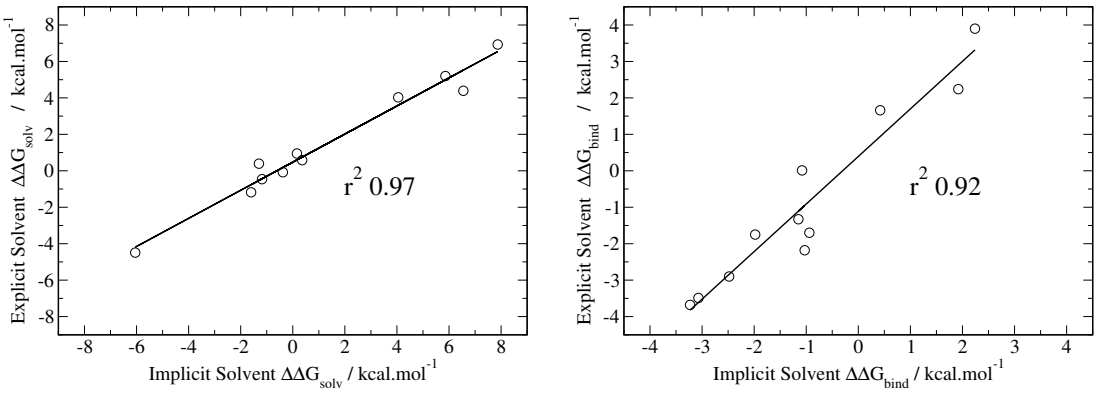


Figure 4.8: Summary of the implicit solvent protocol results. The binding free energies are relative to compound **1**.

It is interesting to compare the predicted solvation and binding free energies for the generalised Born and TIP4P simulations. As can be seen in figure 4.9, there is a strong correlation between the two protocols. In the case of the relative solvation free energies, a coefficient of determination of 0.97 is obtained. It is known that solvation free energies obtained by a generalised Born approach correlate very well with the solvation free energies of small molecules calculated by explicit solvent simulations.[163] Here we demonstrate that this relationship still holds true in the case of more complex, flexible molecules.

The correlation between the calculated binding free energies is lower and the coefficient of determination is 0.92. This suggest that some aspects of solvation in the protein complex are not captured similarly by the two simulation methods.



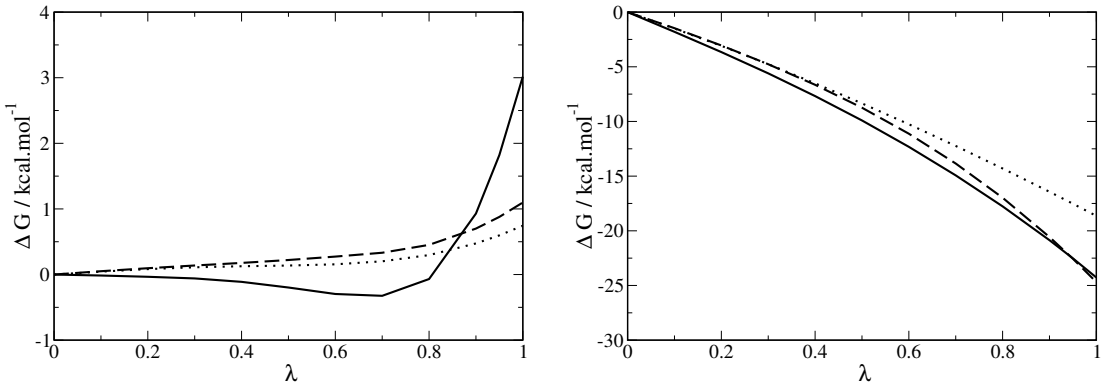
(a) (b)

Figure 4.9: The correlation between predicted solvation and binding free energies by the explicit and implicit solvent simulation protocols.
(a) Relative solvation free energies (b) Relative binding free energies

The free energy profiles for the transformation of **1** to **3** and **11** to **8** are shown in figure 4.10. The free energy changes over the coupling parameter follow similar trends to those observed with the explicit solvent protocol (figure 4.4). The free energy profile for the perturbation of **1** to **3** in implicit water is smoother and somewhat different from the one observed in TIP4P water and the free energy changes at intermediate values of the coupling parameter are different as well. However, the double free energy difference at the end of the perturbation is almost identical to the one obtained with the explicit solvent protocol.

In the perturbation of the unsubstituted derivative **11** into an hydroxy group **8** the main difference observed is that after a λ value of 0.50 the free energy in aqueous implicit water decreases more rapidly than in the complex of COX-2, resulting in ligand **8** being only modestly less stable than ligand **11**. This trend is also observed in the perturbation of **8** to **6**, **8** to **1** and **8** to **9**. This suggest that compound **8** complexed with COX-2 is systematically more stable by about 0.50 to 1.00 kcal mol^{-1} in a generalised Born simulation than in an explicit solvent simulation.

This observation is particularly interesting because in water, **8** is better solvated in the implicit solvent simulations than in the explicit solvent simulations, as evidenced by the calculated relative solvation free energies for the perturbations **8** to **1**, **8** to **9** or **11** to **8** reported in table 4.4. Since the hydroxy substituent on **8** is



(a) (b)

Figure 4.10: Implicit solvent protocol: the change in free energy along the coupling parameter. The solid line represents the change in COX-2, the dashed line represents the change in water and the dotted line represents the change in vacuum. (a) Perturbation of **1** to **3** (b) Perturbation of **11** to **8**

transferred to an essentially hydrophobic pocket in the binding site of COX-2, one would expect the generalised Born simulation to yield binding free energies that disfavour more the binding of this compound than with the explicit solvent protocol. That this behaviour is not observed is due to the treatment of desolvation by the algorithms employed to calculate the Born radii. In the binding pocket, small regions of void exist between the hydroxy group of **8** and the protein side chains. These regions of space are not occupied by water in the explicit solvent simulations. In the generalised Born protocol however, these small regions are treated by the Pairwise Descreening Approximation algorithm as regions of high dielectric ($\varepsilon = 78.5$). As a result, the hydroxy group is still partially solvated even in the binding site. This leads to the relative stabilisation of the polar hydroxy group with respect to the other, less polar groups. This hypothesis is supported by the good agreement in the relative binding free energy between **1** (methyl) and **3** (ethyl) between the two protocols, where the contribution of the generalised Born energy to the solvation free energy is negligible compared to the influence of the non polar term.

Artifacts in solvation due to the presence of small pockets of high dielectric constant in the interior of proteins have been noted by other workers.[140,141,164]

Here, we investigate a simple method that attempts to compensate for the improper treatment of desolvation by the generalised Born approach. By visualising the binding site of COX-2, we locate three small pockets of void that surround the 5-aryl group of the ligand and we position 3 spheres of radius 2 Å in each pocket. (See Figure 4.11 for clarity) The spheres are assigned generalised Born parameters suitable for a carbon atom (more precisely, a scaling factor of 0.77 and an offset to the van der Waals radius of 0.68). Other force field parameters (charge and Lennard Jones well-depth) are set to 0. As a result the only impact of these spheres on the simulation is that they displace a volume of dielectric. Because these spheres make close contact with the parts of the ligands that are subject to a perturbation they affect their Born radii which in turn changes the generalised Born energy of the ligands. This protocol bears some resemblance to the method proposed by Liu et al.[165] to take into account the presence of small voids between the ligand and receptor atoms.

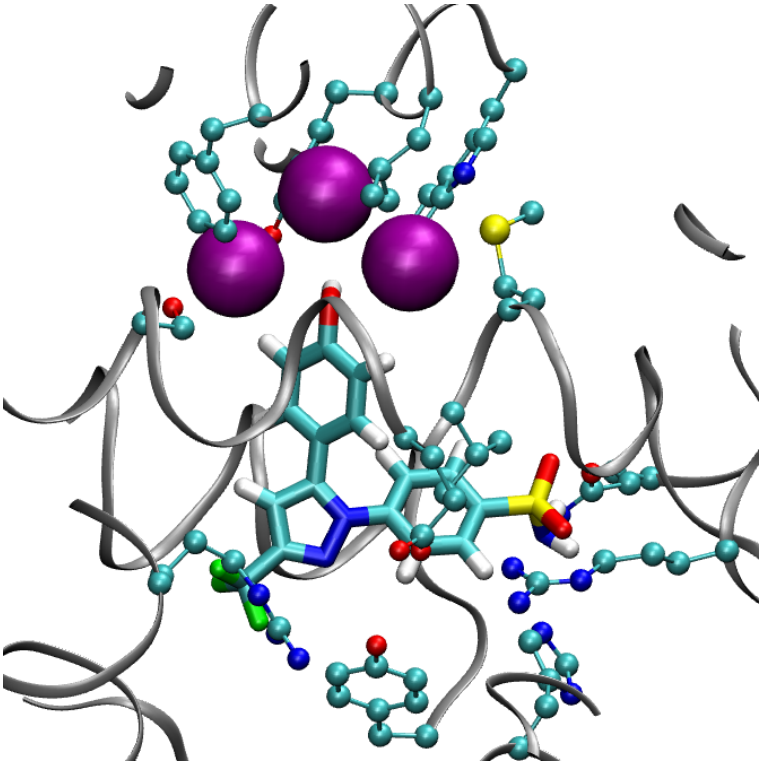


Figure 4.11: A model of compound **8** in the binding site of COX-2 with the addition of three 2 Å radii spheres that cover approximately the small regions of void left between the ligand and the pocket where the 5-aryl ring extends. Hydrogen atoms on the amino acid side chains are not shown for clarity.

Table 4.6 shows the results of the generalised Born simulations conducted in the presence of these extra particles.

Table 4.6: Comparison between experimental and calculated relative binding free energies with the modified implicit solvent protocol[a]

Pert	Exp	$\Delta\Delta G_{bind}$ [b]	$\Delta\Delta G_{bind}$ [c]
1t3	1.82	2.00 ± 0.31	1.92 ± 0.31
4t3	-2.78	-2.96 ± 0.26	-1.03 ± 0.23
1t7	3.16	2.34 ± 0.17	2.24 ± 0.18
8t1	$<$ -4.64	-3.07 ± 0.18	-2.48 ± 0.19
8t9	$<$ -5.46	-3.74 ± 0.08	-3.07 ± 0.07
10t9	-0.84	-1.22 ± 0.06	-1.15 ± 0.07
11t10	0.15	-1.15 ± 0.05	-1.08 ± 0.05
11t8	$>$ +4.77	0.99 ± 0.09	0.42 ± 0.07
3t5	-2.7	-1.27 ± 0.39	-0.94 ± 0.45
5t6	-0.07	-1.94 ± 0.41	-1.98 ± 0.68
8t6	$<$ -5.59	-3.85 ± 0.36	-3.23 ± 0.54

a Figures in kcal mol^{-1}

b Modified implicit solvent protocol

c Standard implicit solvent protocol

The closure of the thermodynamic cycle **8** to **9**, **9** to **10**, **10** to **11** and **11** to **8** is 0.38 kcal mol^{-1}. The closure of the thermodynamic cycle **3** to **5**, **5** to **6**, **6** to **8**, **8** to **1** and **1** to **3** is 0.43 kcal mol^{-1}. These low figures suggest that the simulation results can be interpreted with confidence.

Comparison with the results listed in table 4.4 shows that the perturbations involving an hydroxy group are now less favourable for the hydroxy substituent by 1.93 kcal mol^{-1} (**4** to **3**), 0.59 kcal mol^{-1} (**8** to **1**), 0.67 kcal mol^{-1} (**8** to **9**) and 0.57 kcal mol^{-1} (**8** to **11**). These changes are much larger than the standard error associated with each figure and hence significant. The addition of the three dielectric displacing particles has resulted in the destabilisation of the more polar groups. This is because their effect is to increase the Born radii of the polar hydrogen and oxygen atom of compound **4** and **8**. This results in a lowered solvation energy for these compounds, effectively making their introduction into the binding site less favourable. Note that the difference in binding free energy for the remain-

ing perturbations between the two protocols is much lower and within the error bars of each simulation. This is slightly unexpected as some perturbations exhibit large differences in solvation free energy between the watercap and generalised Born protocol. For example in the perturbation of **1** to **7** the solvation free energy obtained with an explicit solvent protocol is 0.39 $\pm$ 0.23 kcal mol^{-1} and -1.30 $\pm$ 0.09 kcal mol^{-1} with the implicit solvent protocol. In the binding site of COX-2, the relative binding free energy is 3.90 $\pm$ 0.25 kcal mol^{-1} with the explicit solvent protocol and 2.24 $\pm$ 0.17 kcal mol^{-1} with the modified implicit solvent protocol. Since it is easier to desolvate compound **7** from explicit water than for implicit water, one would expect to find a lower binding free energy for this compound with the explicit solvent protocol. However, the opposite is the case. Furthermore the binding free energy of **1** to **7** is similar between the two modified generalised Born protocols (the difference of 0.1 kcal mol^{-1} is within the statistical error). These observations suggest that the comparison of the difference in solvation free energies between implicit and explicit solvent protocols does not always rationalise the difference of binding free energies predicted by the two simulation methods. It is worth mentioning that compound **7** contains three fluorine atoms and that this class of compounds was identified as significant outliers during the parameterisation of a generalised Born force field reported in chapter 2.

In figure 4.12 the correlation of the calculated binding free energies between the explicit and modified implicit protocols is plotted. The correlation has increased and the coefficient of determination is now 0.96 which is similar to the degree of correlation observed between the solvation free energies calculated with the two protocols.

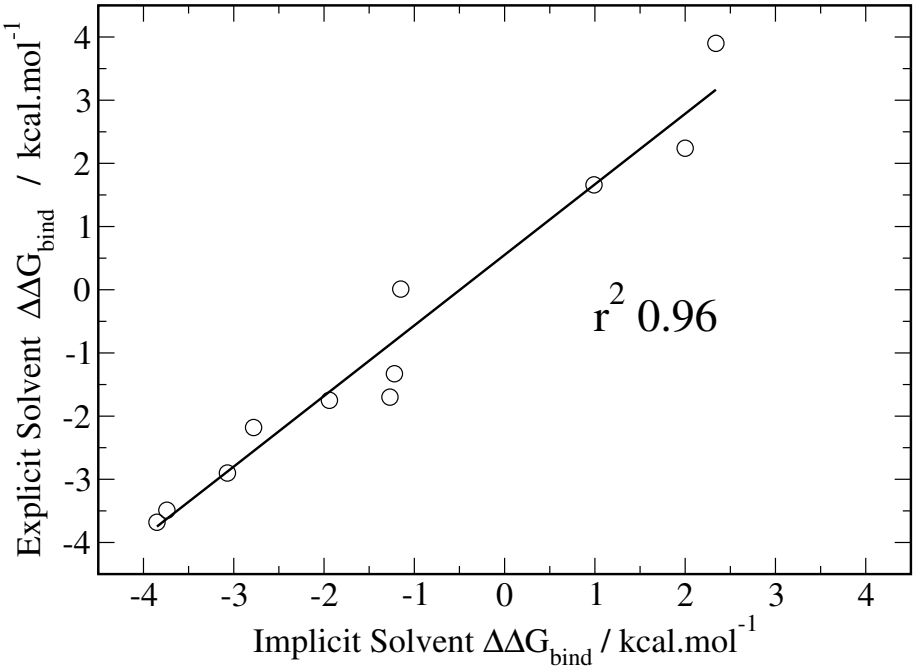


Figure 4.12: Correlation between the calculated binding free energies of the explicit solvent and modified implicit solvent protocol

Table 4.7: Modified implicit solvent protocol: the experimental and calculated binding free energies with respect to celecoxib **1**.[a]

Compound	Perturbation pathway[b]	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
6	[**1t8**+**8t6**] ; [**1t3**+**3t5**+**5t6**]	-1.00 $\pm$ 0.71	-0.95
5	[**1t3**+**3t5**]	0.73 $\pm$ 0.55	-0.88
9	[**1t8**+**8t9**]	-0.67 $\pm$ 0.21	-0.82
11	[**1t8**+**8t9**+**9t10**+**10t11**];[**1t8**+**8t11**]	1.89 $\pm$ 0.22	-0.13
1		0	0
10	[**1t8**+**8t9**+**9t10**];[**1t8**+**8t11**+**11t10**]	0.74 $\pm$ 0.22	0.01
3	**1t3**	2.00 $\pm$ 0.32	1.82
7	**1t7**	2.34 $\pm$ 0.19	3.15
4	[**1t3**+**3t4**]	4.96 $\pm$ 0.43	4.59
8	**1t8**	3.07 $\pm$ 0.20	4.63

a Figures in kcal mol^{-1}

b Figures obtained by summing free energy changes over different perturbations, and in some cases, averaging over two different pathways

Next we consider the agreement of the simulations with the experimental measurements (table 4.7 and figure 4.13). The mean unsigned error is now 0.83 kcal mol^{-1}, the predictive index 0.96 and the coefficient of determination 0.79. It appears there-

fore that a better treatment of desolvation has increased the quantitative accuracy of the implicit solvent calculations, even though they remain, overall, slightly inferior to the explicit solvent calculations. It appears that the PI method, while good at assessing the ability to rank a set of inhibitors in their order of affinity, does not discriminate between different protocols once they have reached a sufficiently high accuracy. It is therefore important to characterise quantitative agreement with other methods.

The modified generalised Born protocol is very simple as it only involves filling pockets of void with spheres. Improvements in relative free energy calculations are observed because the spheres have been positioned close to the substituents that are being perturbed. In principle, there are several pockets in the protein that would need to be filled which could render the protocol cumbersome. Furthermore, in other binding sites, there might be a partial occupancy of a pocket by a water molecule. The proposed protocol suffers thus from some limitations.

The results are summarised in figure 4.13.

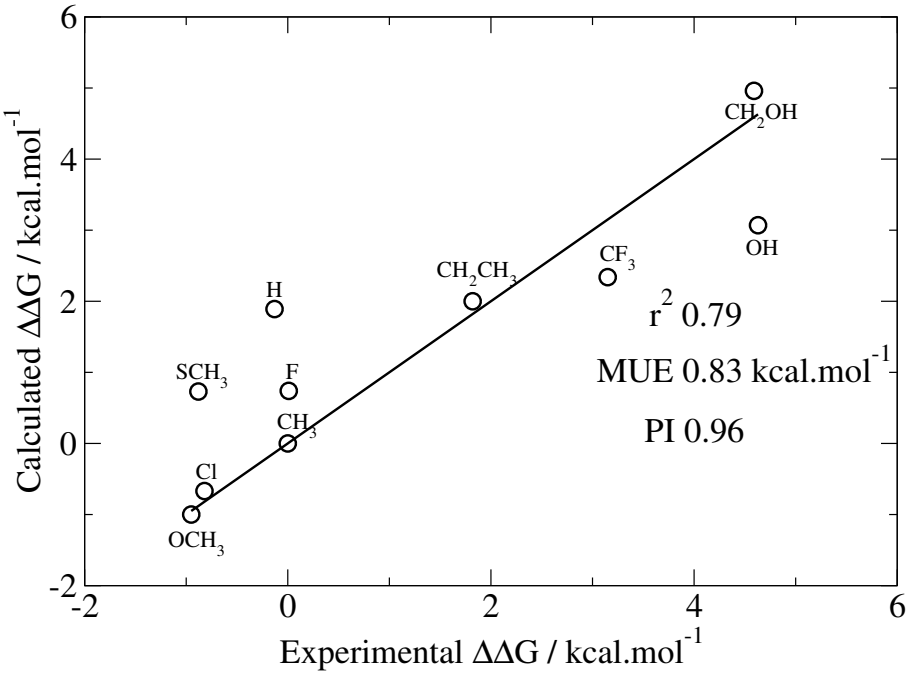


Figure 4.13: Summary of the modified implicit solvent protocol results. The binding free energies are relative to compound **1**.

4.6 Influence of protein flexibility

A significant difficulty in the calculation of protein ligand relative binding free energy calculations arises from the sampling of the several protein and solvent degrees of freedom in addition to the ligand degrees of freedom. An implicit solvent framework reduces such complexity, but the degrees of freedom of several protein side chain must still be sampled. Here we consider the impact of such protein side chains flexibility on the calculated binding free energies. A rigid model of protein solvated in explicit water would not be feasible. This is because the simulation results would be heavily affected by the position and orientation of the surrounding water molecules. In addition, the explicit solvent would restrict severely the configurations the ligand can sample, particularly for the simulation in the unbound state. The simulation results would be dependent on the initial placement of the water molecules, which is arbitrary. By contrast, in the crystal structure of a protein, the position of the protein side chain is known with some precision. By adopting an implicit model of water, no uncertainties associated with the orientation of water molecules is introduced in the model.

In the following section, protein-ligand binding free energies were calculated with the protein rigid and a generalised Born model of water. The simulation properties were averaged for only 300 K moves because the dimensionality of the energy landscape has been reduced to only the degrees of freedom of the ligand, and faster convergence of the free energies is expected. No pre-equilibration was necessary, and each window was equilibrated for 30 K moves before data collection. Because solute moves are more expensive on average than protein side chain moves, the simulations were only about 1.5 times faster than the generalised Born simulations with protein flexibility even though one third of Monte Carlo moves were performed in total. The simulations were run in presence of the three solvent displacing particles.

Table 4.8: Comparison between experimental and calculated relative binding free energies with a rigid protein and the modified implicit solvent protocol[a]

Pert	Exp	$\Delta\Delta G_{bind}$[b]	$\Delta\Delta G_{bind}$[c]
1t3	1.82	1.04 ± 0.22	2.00 ± 0.31
4t3	-2.78	-2.65 ± 0.21	-2.96 ± 0.26
1t7	3.16	3.27 ± 0.14	2.34 ± 0.17
8t1	$<$ -4.64	-3.62 ± 0.15	-3.07 ± 0.18
8t9	$<$ -5.46	-3.65 ± 0.15	-3.74 ± 0.08
10t9	-0.84	-1.57 ± 0.03	-1.22 ± 0.06
11t10	0.15	-1.48 ± 0.02	-1.15 ± 0.05
11t8	$>$ +4.77	0.77 ± 0.07	0.99 ± 0.09
3t5	-2.7	-1.27 ± 0.36	-1.27 ± 0.39
5t6	-0.07	-1.58 ± 0.62	-1.94 ± 0.41
8t6	$<$ -5.59	-4.74 ± 0.63	-3.85 ± 0.36

a Figures in kcal mol^{-1}

b Modified implicit solvent protocol, no protein flexibility

c Modified implicit solvent protocol, protein flexibility

The closure of the thermodynamic cycle **8** to **9**, **9** to **10**, **10** to **11** and **11** to **8** is 0.17 kcal mol^{-1}. The closure of the thermodynamic cycle **3** to **5**, **5** to **6**, **6** to **8**, **8** to **1** and **1** to **3** is 0.69 kcal mol^{-1}.

The figures are remarkably similar to the results obtained with protein flexibility. The perturbations exhibiting the biggest differences are **5** to **6**, **1** to **3**, **1** to **7** and **8** to **1**.

The binding free energies with respect to celecoxib are shown in table 4.9.

Table 4.9: Modified implicit solvent protocol and rigid protein: the experimental and calculated binding free energies with respect to celecoxib **1**.[a]

Compound	Perturbation pathway[b]	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
6	[**1t8**+**8t6**] ; [**1t3**+**3t5**+**5t6**]	-1.47 $\pm$ 0.72	-0.95
5	[**1t3**+**3t5**]	-0.23 $\pm$ 0.43	-0.88
9	[**1t8**+**8t9**]	-0.03 $\pm$ 0.17	-0.82
11	[**1t8**+**8t9**+**9t10**+**10t11**];[**1t8**+**8t11**]	2.94 $\pm$ 0.17	-0.13
1		0	0
10	[**1t8**+**8t9**+**9t10**];[**1t8**+**8t11**+**11t10**]	1.46 $\pm$ 0.17	0.01
3	**1t3**	1.04 $\pm$ 0.22	1.82
7	**1t7**	3.27 $\pm$ 0.14	3.15
4	[**1t3**+**3t4**]	3.69 $\pm$ 0.30	4.59
8	**1t8**	3.62 $\pm$ 0.15	4.63

a Figures in kcal mol^{-1}.

b Figures obtained by summing free energy changes over different perturbations, and in some cases, averaging over two different pathways.

When the perturbations are carried out with a rigid protein, the binding free energy of the bigger substituent like **3** or **5** is lowered by approximately 1 kcal mol^{-1} in both cases and 1.3 kcal mol^{-1} for **4**. On the other hand the free energy of the smallest substituents has increased, for **11** by 0.90 kcal mol^{-1}, for **10** by 0.55 kcal mol^{-1}, for **9** by 0.50 kcal mol^{-1}. While this is not conclusive evidence, it is tempting to argue that since the protein is rigid in these simulations, the motion of the protein side chains in the binding site is not affected by the presence of bulkier substituent. Thus there would be no penalty for growing a larger substituent into the binding site, and conversely, it would not be as favorable to put a smaller substituent into the binding site. One could also have reasoned that in the absence of protein flexibility, larger substituents are more likely to make bad contacts with the protein side chains and hence be less stabilised. However, the protein ligand system was prepared by energy minimisation of the the largest ligand, compound **3**, complexed in the binding site. Thus there is ample room to fit all the ligands from this series without incurring steric clashes with the protein. This could also explain the trends in the increased/decreased binding affinity of the larger/smaller compounds in this series. The mean unsigned error is 1.03 kcal mol^{-1} and the predictive index is 0.93. These figures are a bit lower than those obtained when

protein flexibility is enabled. This is not entirely surprising, as the present model should be less realistic. However, the predictions are still of a very good quality, and the associated reduction in computational expense could be worth the introduction of such an approximation. The present results challenge the assumption that receptor flexibility is necessary to obtain meaningful protein ligand binding free energies. To emphasise this last point, for this series of compounds, a more elaborate treatment of solvation has more impact on the binding free energies than the inclusion of protein flexibility.

The results obtained with this simulation protocol are summarised in figure 4.14.

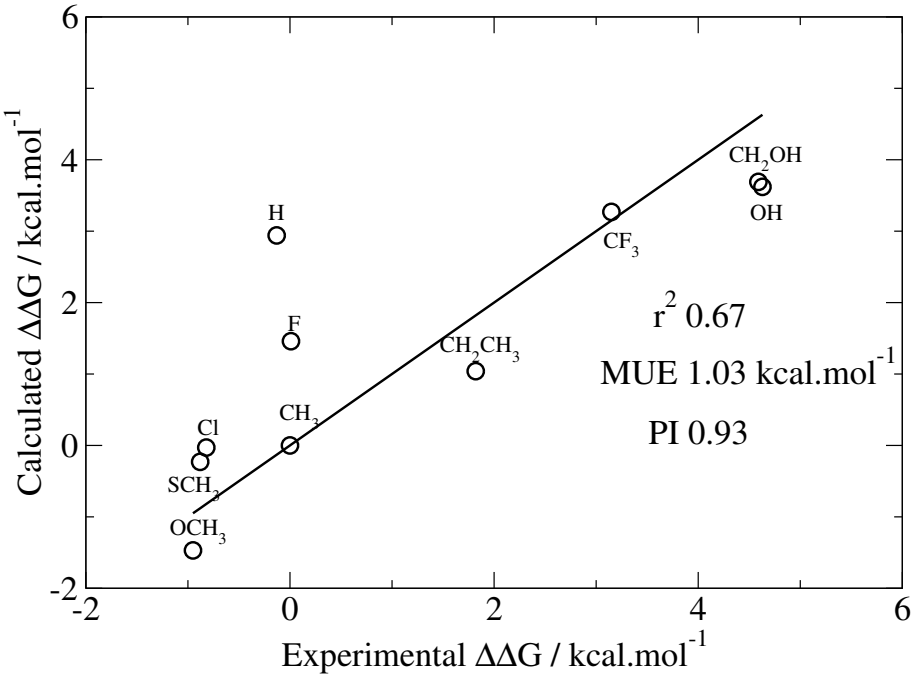


Figure 4.14: Summary of the results with the modified implicit solvent protocol and no protein flexibility. The binding free energies are relative to compound **1**.

4.7 Computational cost and convergence

No specific rule dictated the choice of the number of Monte Carlo moves employed to calculate the free energy changes reported in the previous sections. It is interesting to evaluate, *a posteriori*, the quality of the predictions as a function of the amount of computational resources invested. This would also provide a fair comparison of the implicit and explicit solvent protocol.

In figure 4.15, the mean unsigned error between experimental and predicted binding free energies relative to compound **1**, is plotted as a function of the average CPU time taken by one simulation performed at one value of λ. This represents the shortest amount of time that one would have to wait to obtain a predictivity plot, assuming enough CPUs are available to run all the perturbations simultaneously (in this case, 11 perturbations with 12 windows each, meaning 132 CPUs). In addition, while simulations in the unbound state are very fast in the implicit solvent simulations (about 20 minutes), they do take longer in the explicit solvent simulations and their cost has to be considered. All the timings are based on the time taken to complete a simulation on a 2.2 GHz AMD Opteron dual processor.

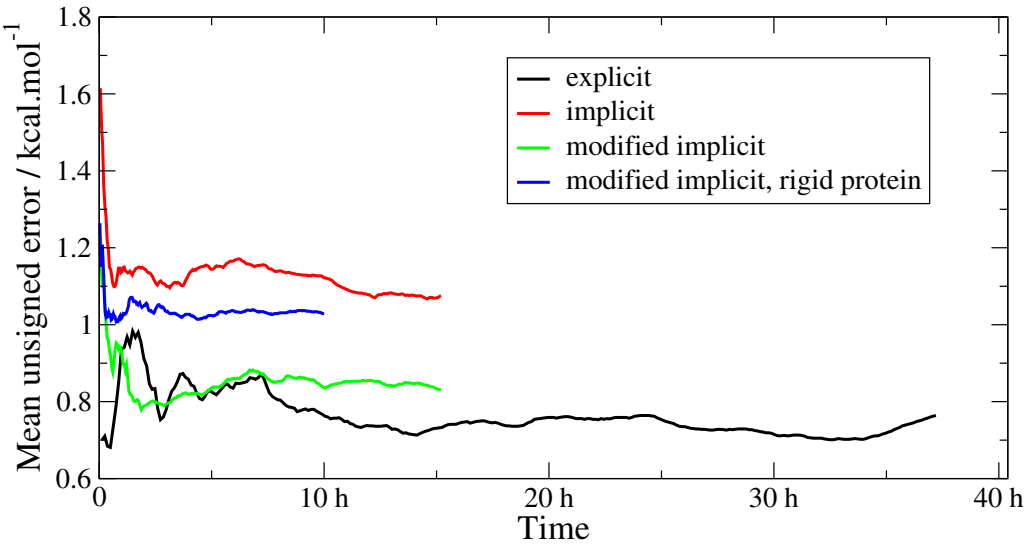


Figure 4.15: The convergence of the mean unsigned error as a function of the time taken to complete a single simulation at one value of λ.

Figure 4.15 shows that the mean unsigned error converges very quickly on this system. For the explicit solvent simulations, stable results require about 10 hours of simulation. The MUE of the implicit and modified implicit simulation protocols does not evolve much after 5-7 hours of simulation. For the simulations conducted with a rigid protein the MUE is stable after 2-3 hours.

The speed at which a good qualitative ranking can be obtained is measured in figure 4.16. The set of values the predictive index can adopt for this set of compounds is discrete, which explains the jumps in PI observed in this figure, which occur when the relative binding free energy of one compound has changed sufficiently such that it has become more or less stable than the two other closest

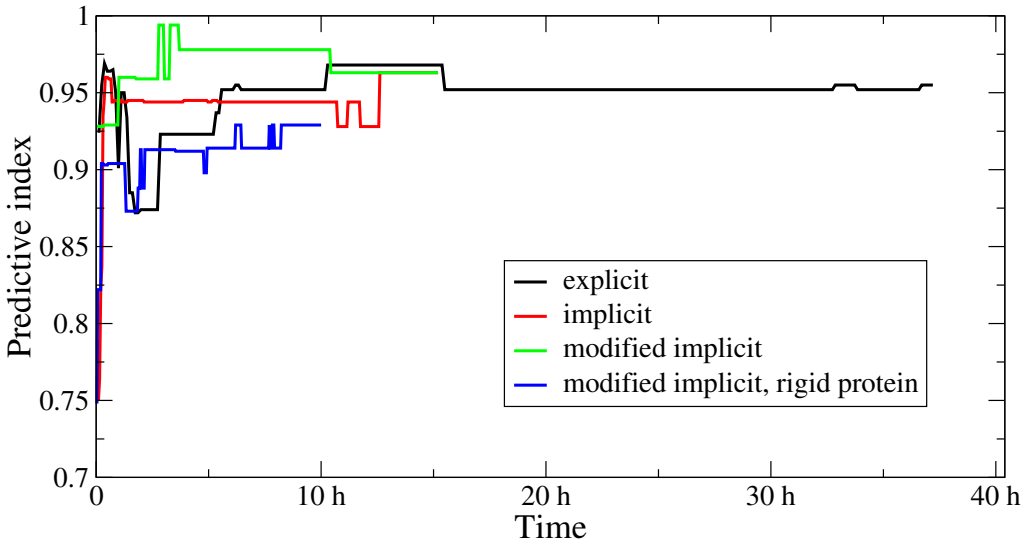


Figure 4.16: The convergence of the predictive index as a function of the time taken to complete a single simulation at one value of λ.

compounds in order of affinity. We observe that very little sampling is required to obtain a high predictive index. All methods yield a PI greater than 0.90 after only 3 hours of simulation. The PI for the explicit solvent simulations is stable after about 5 hours of simulation. This varies between 1 to 4 hours for the implicit and modified implicit solvent simulations. The PI for the simulations conducted with a rigid protein is essentially stable after 2 hours.

The following observations suggest that the protocols employed to calculate the relative binding free energies have dramatically overestimated the number of moves required to obtain stable predictions. It is likely that errors due to lack of convergence of individual perturbations cancels out to some extent when the results of several perturbations are used to build a predictivity plot, meaning that reliable predictions can be obtained before all individual perturbations are fully converged.

In a real-world situation, one would not know the experimental binding free energies or ranking of the simulated compounds and the previous plots could not be constructed to decide when sufficient sampling has been performed. It could be possible however to plot the convergence of the closure of a thermodynamic cycle as a function of time. In the limit of infinite sampling, and assuming all the models used in the different perturbations are fully consistent, this quantity should converge to zero. In practice, this quantity is likely to fluctuate around this figure

once the results are reasonably converged.

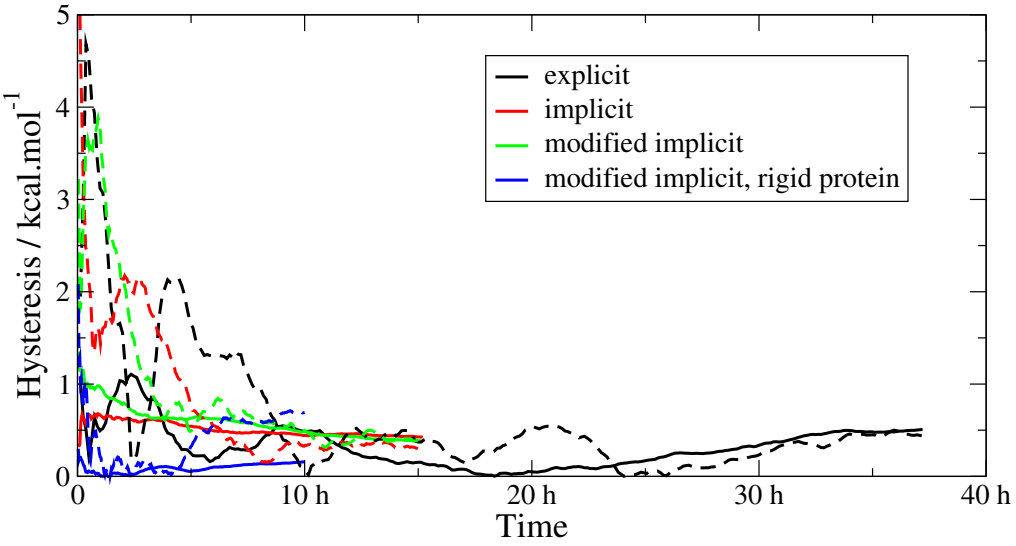


Figure 4.17: The convergence of the closure of the thermodynamic cycles as a function of the time taken to complete a single simulation at one value of λ, in the bound and unbound state. The straight lines are for the cycle involving compounds **8**, **9**, **10** and **11**. The dashed lines for the cycle involving compounds **1**, **3**, **5**, **6** and **8**.

In figure 4.17, the convergence of the two closed cycles in this series is plotted as a function of time. If we assume that the results are considered converged if the closure for both cycles is under 1 kcal mol^{-1}, then the explicit solvent simulations would be considered converged after 8 hours, the implicit solvent simulations after 5 hours, the modified implicit solvent simulations 3 hours, and the implicit solvent simulations with a rigid protein in under 1 hour. This metric is unlikely to be perfect. For example, while the hysteresis is very quickly essentially 0 kcal mol^{-1} for both cycles with the rigid protein simulations, it increases to 0.7 kcal mol^{-1} after about 5 hours for one cycle. In the explicit solvent simulations, the hysteresis for one cycle slowly decreases to 0 kcal mol^{-1} in 18 hours, but increases then to finish at 0.5 kcal mol^{-1} at the end of the simulation.

4.8 Comparison with empirical scoring functions

Predictive indices for the series of celecoxib analogues have been computed using the Chemscore,[19,20] GoldScore[12] and ASP energy function.[22] The results are presented graphically in figures 4.18, 4.19 and 4.20. Since it is unclear how the scores

predicted by these methods can be related to binding free energies, quantitative descriptors such as the mean unsigned error or the coefficient of determination were not calculated.

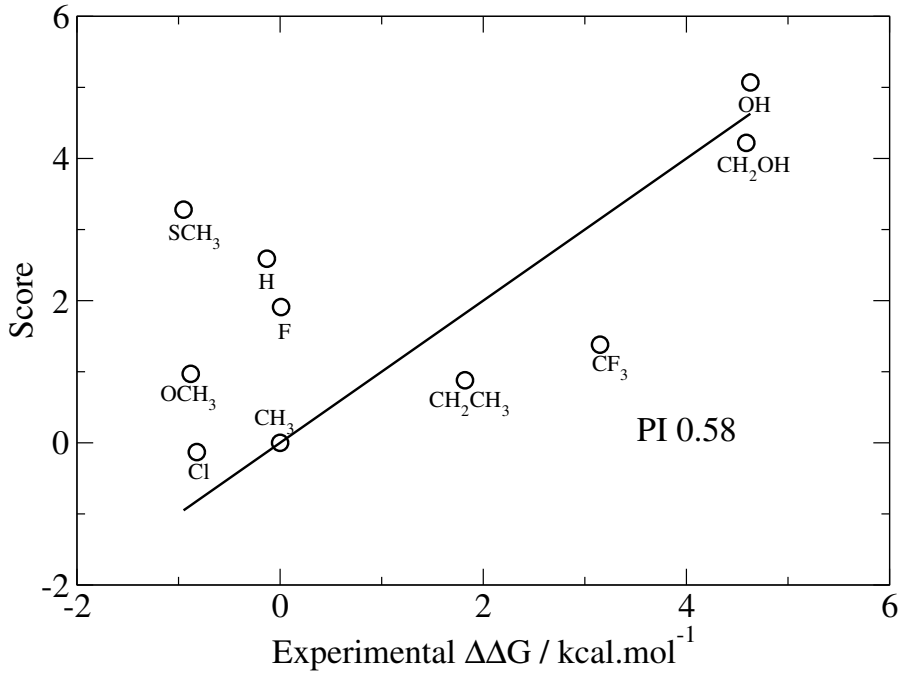


Figure 4.18: Chemscore: Calculated score and experimental binding free energy of a series of celecoxib analogues. All the data is relative to compound **1**

Chemscore performs reasonably and the predictive index is 0.58. The two high micromolar inhibitors **4** and **8** have been discriminated from the other compounds and the chlorinated substituent **9** is indeed found to be a more potent binder than celecoxib **1**. There is however no discrimination between the remaining compounds and the thiol ether derivative, the second most potent binder in that series, scores as the third worst, just below the hydroxy groups.

Goldscore performs very poorly on this set. While the ether substituent **6** has been identified as the best binder, the thio ether **5** which has an almost identical affinity is the predicted worst binder and lies off the scale of the plot. The PI is actually negative which means that Goldscore rank these compounds worse than a random ranking. This is essentially because the hydroxy groups score as well as the best binders.

ASP shows no discrimination either and the trifluoro group is predicted to be significantly better than any other group while it is actually a poor binder. The PI of -0.44 indicates that the predicted rankings tend to be anti-correlated with the

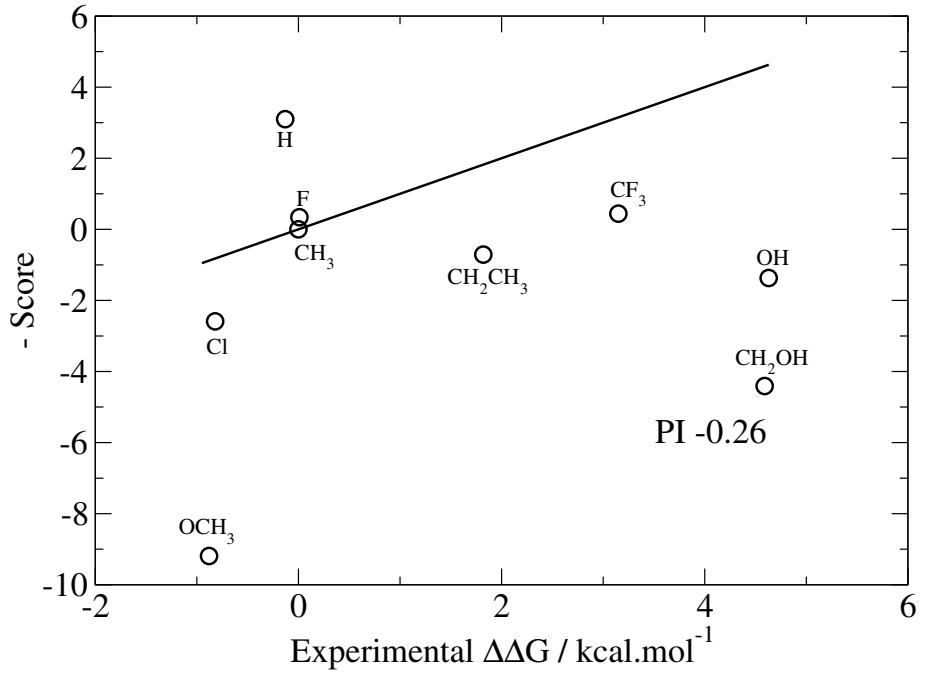


Figure 4.19: Goldscore: Calculated score and experimental binding free energy of a series of celecoxib analogues. All the data is relative to compound **1**. The negative of the Goldscore is plotted such that if the method can explain the variation of the relative binding free energies, a positive correlation would be observed.

experimental ordering.

4.9 Conclusion

The relative binding free energies of a series of NSAIDs COX-2 specific analogues of celecoxib have been calculated by means of explicit solvent (TIP4P) and implicit solvent (generalised Born) free energy simulations. The results are in good agreement with experimental measurements. The explicit solvent simulations yield a mean unsigned error of 0.76 kcal mol^{-1} and a coefficient of determination of 0.85. The implicit solvent simulations yield a mean unsigned error of 1.08 kcal mol^{-1} and a coefficient of determination of 0.70. Predictive indices which measure the ability of the predictions to rank the inhibitors according to their relative affinities are calculated for both methods. The very high values obtained (0.96 in both cases) validate the application of these methods to this system. Systematic differences between the implicit and explicit simulation results are investigated and it is shown that the origin of some of the differences lies in the incorrect treatment of desolvation by the generalised Born algorithms employed in this study.

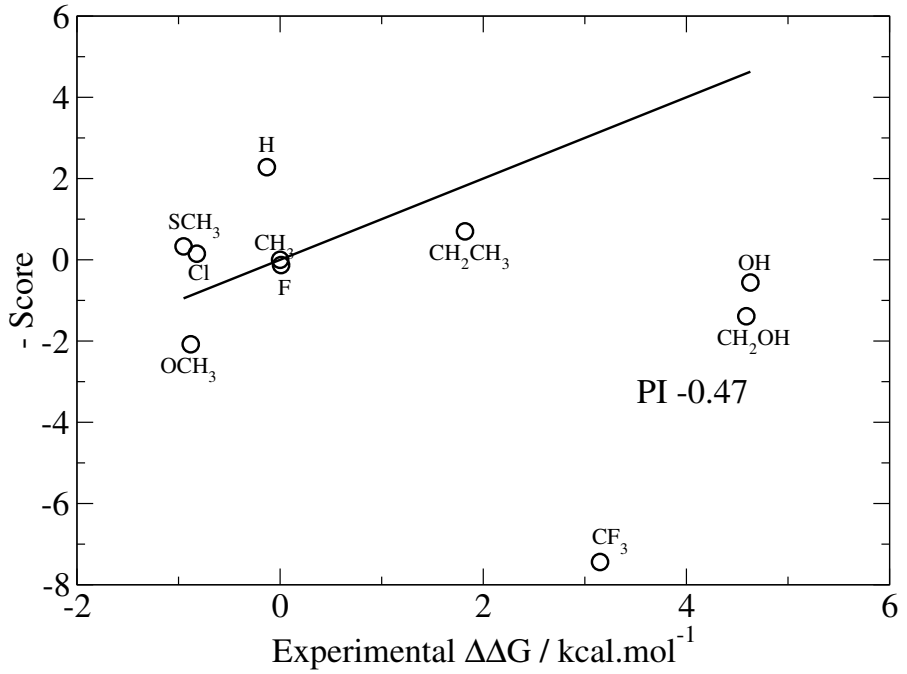


Figure 4.20: ASP: Calculated score and experimental binding free energy of a series of celecoxib analogues. All the data is relative to compound **1**, and the negative of the score is plotted for reasons similar to the Goldscore method.

The elimination of small pockets of high-dielectric that surround the perturbed part of the ligands yields a generalised Born model in better agreement with experimental results (PI 0.96, MUE 0.83 kcal mol^{-1}, r^2 0.79). While a method that would treat desolvation correctly in a more general fashion has yet to be devised, we demonstrate here that there is ample room for further optimisation of implicit solvation methodologies on this system. The effect of keeping rigid the protein environment while perturbating the ligands is then investigated. It is found that the quality of the predictions is affected, but only to a small extent (PI 0.93, MUE 1.03 kcal mol^{-1}, r^2 0.67). The time necessary to obtain converged predictions is then assessed *a posteriori* and it is found that high PIs can be obtained after only 1 to 5 hours of simulation, even though the individual free energy differences take longer to converge. Finally, the ability of commonly used empirical scoring functions to rank these compounds correctly has been assessed by calculating predictive indexes for Chemscore, Goldscore and ASP. The obtained PI of 0.58, -0.26 and -0.47 are significantly lower and demonstrate the superiority of the free energy method for the ranking of the inhibitors in this series against COX-2.

Chapter 5

Application to a protein-ligand system : neuraminidase

> “If it’s green or wriggles, it’s biology.
> If it stinks, it’s chemistry.
> If it doesn’t work, it’s physics.”
>
> Handy Guide to Science

5.1 Introduction

The application of generalised Born free energy techniques to the prediction of the relative binding free energies of a series of inhibitors of the influenza enzyme neuraminidase is explored. Neuraminidase has a polar, solvent exposed binding site and ligand binding involves several identified crystallographic bound waters. It is therefore considered a challenging test case of implicit solvent methodologies. In this chapter, simulation results are systematically compared to explicit solvent simulations performed with the same system setup and force field, and typical empirical scoring functions.

5.2 Presentation of the system

The common flu is a contagious respiratory illness caused by influenza viruses. It can cause mild to severe illness, and at times can lead to death. There are three

types of influenza viruses, designated influenza A, B and C and distinguished by their different genetic sequences. The most severe flu in humans is usually caused by type A influenza viruses. Each type of flu is divided into several subtypes, depending on the type of the surface proteins hemaglutinin (HA) and neuraminidase (NA) that are sticking through the viral envelope of the virus. There are 16 types of HA, designated H1-H16 and 9 NA sub types, designated N1-N9. All of these possible combinations of surface proteins are able to infect a variety of animals, but so far, only those containing the H1, H2, H3, H5, H7 and H9 and the N1, N2 and N7 surface proteins infect humans, and of these, so far, only H1, H2, H3 and N1 and N2 do so to any extent. Finer distinctions between subtypes are necessary because influenza viruses mutate readily and within subtypes there may be many genetic variants, called strains.[166]

Usually, 'avian influenza virus' refers to influenza A viruses found chiefly in birds, but infections caused by these viruses can occur in humans. The risk from avian influenza is generally low to most people, because the viruses do not usually infect humans. However, health autorithies worldwide are particularly worried by the existence of a virulent H5N1 subtype of avian influenza virus that is being spread over the world by diseased migratory birds. While the H5N1 avian virus does not readily infect people, repeated exposure to infected birds or poultry can cause infection, with a high mortality rate. Virologists fear that through repeated exposures to humans, the H5N1 avian influenza will be able to mutate into a strain that can be passed between humans. Because these viruses do not commonly infect humans, there is little or no immune protection against them in the human population and if the virus were to gain capacity to spread easily between humans, a deadly influenza pandemic could occur, putting the lives of billions of people at risk.[166]

Besides vaccination, a number of available anti-viral treatments can help to slow the spread of the flu virus in an infected body and reduce the mortality rate. Most neuraminidase inhibitors prevent this surface protein of the flu virus from cleaving sialic acid residues from the carbohydrate sidechains present in the membranes of cells. As a result, the newly produced viruses find themselves unable to leave the infected cell and propagate the infection. It is therefore important to re-

ceive anti-viral treatment within the first 48 hours of infection, otherwise it might be too late to slow down the spread of the disease and avoid the most severe symptoms.

The binding site of neuraminidase is a polar pocket, extending onto the surface of the protein. A depiction of the interactions between important amino acids in the binding site and an analogue of sialic acid, compound **10** is presented in figure 5.1.

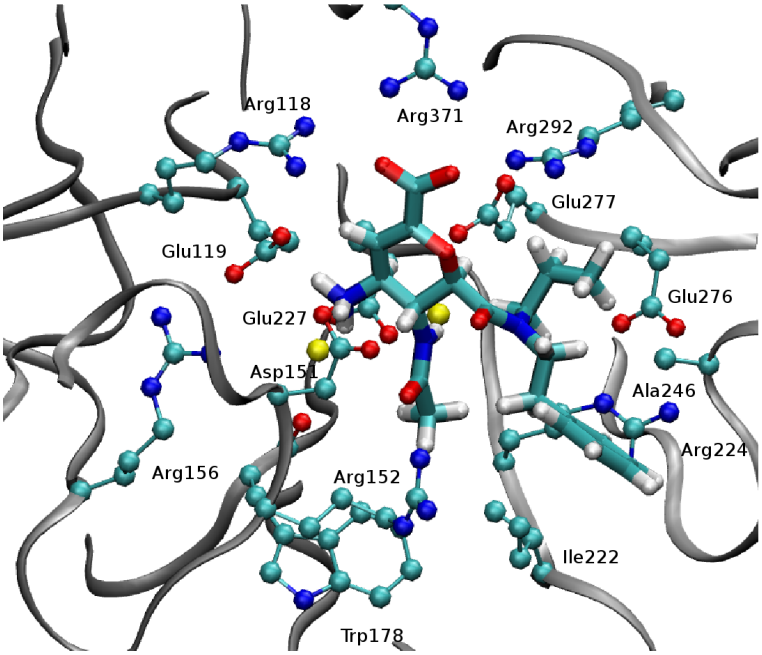


Figure 5.1: The binding site in the crystal structure of a N9 strain of influenza A complexed with the 6-carboxamide sialic acid analogue compound **10**.[149] Hydrogen atoms on the amino acid sidechains and waters are not shown for clarity. The oxygen atoms of the buried water molecules are shown in yellow.

Towards the top of the binding site, a triad of arginine side chains (Arg118, Arg292 and Arg371) interact strongly with the carboxylate moiety of the inhibitor, including a planar salt bridge with Arg371. The acetamido fragment (middle, bottom of picture 5.1) interacts through hydrogen bonds with Arg152 and a buried water molecule, and the methyl group makes favorable contacts with Trp178 and Ile222. The amino group of **10** (left corner of picture 5.1) occupies a small pocket where it experiences hydrogen bonding interactions with Glu119, Asp151, Arg156, Glu227 and a buried water molecule. The apolar substituents on the amide group (right corner of picture 5.1) fill two distinct small pockets. The phenethyl group of **10** fits into the 'trans' pocket between Ile222 and Ala246 while the propyl

group occupies the 'cis' pocket delimited by Glu276 and Arg224.

Different substitutions on the cis and trans part of the amide group are listed in table 5.1. It can be seen that the perturbations cover a wide range of binding affinity. The common scaffold of these inhibitors is shown in figure 5.2.

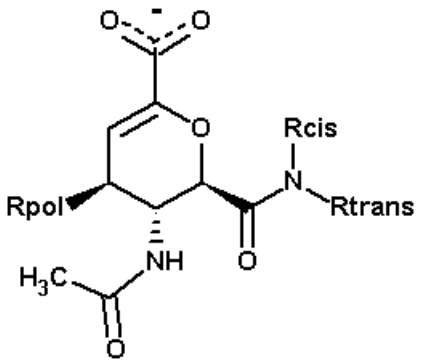


Figure 5.2: Structure of the neuraminidase inhibitors.

Table 5.1: Experimental activity of the sialic acid analogues against neuraminidase[78]

Compound	R_{trans}	R_{cis}	R_{pol}	IC_{50} (μM)
1	Me	H	NH_3^+	190
2	Et	H	NH_3^+	13
3	Me	Me	NH_3^+	2.4
4	Et	Et	NH_3^+	0.003
5	Me	H	$NHC(NH_2)^+$	7
6	Me	Me	$NHC(NH_2)^+$	0.025
7	Et	Et	$NHC(NH_2)^+$	0.001
8	$(CH_2)_2Ph$	Pr	$NHC(NH_2)^+$	0.005
9	$(CH_2)_2Ph$	H	NH_3^+	12
10	$(CH_2)_2Ph$	Pr	NH_3^+	0.005

In general, it appears more favourable to fill the cis pocket than the trans pocket, up to the Et or Pr substituent (**2** to **4**, **2** to **3**, **1** to **2**). This is presumably because this pocket is less solvent exposed and hydrophobic substituents placed there can experience stronger van der Waals forces than in the wider, more exposed trans pocket. It appears that there is no further benefit from adding substituents bulkier than Et in either pocket (**4** to **10**).

In addition, the amino group can be replaced by a guanadino group. The resulting compounds exhibit a stronger binding affinity. Crystallographic evidence suggests that the increases in binding affinity is partly due to the expulsion of a buried water molecule. This water would normally interact with the amino group, but cannot be accomodated in the binding site in the presence of the bulkier guanadino group.

5.3 System setup and simulation protocols

The PDB structure of a N9 neuraminidase complexed to compound **10** at a resolution of 2.0 Å was selected as the starting point for this study (PDB code 1BJI). Hydrogen atoms were added using the program reduce.[151] The ligands were modelled in the complex on the basis of the binding mode of **10**. As the experimental studies were conducted at a pH of 6.5, the histidines were assumed to be protonated, unless there was evidence that a hydrogen bond could be accepted from another residue. The protein was setup with the AMBER99 force field, inhibitors were setup with the GAFF force field and the atomic partial charges were derived using the AM1/BCC method[109] as implemented in the package AMBER8.[112] The system was energy minimized using the Sander module of AMBER8 and a generalised Born force field.[112] The backbone of the energy minimized protein was kept rigid for subsequent Monte Carlo simulations which were conducted with a modified version of the ProtoMS2.1 package.[147] To reduce the computational cost, only the protein residues that are within 15 Å of any heavy atom of compound **10** were retained. The resulting protein scoop consisted of 145 residues. For the explicit solvent simulations, the complex was hydrated by a sphere of TIP4P water molecules[138] of 22 Å radius and centred near the geometric centre of the ligand. To prevent evaporation, a half-harmonic potential with a 1.5 kcal $mol^{-1}Å^{-1}$ constant was applied to water molecules whose oxygen atom distance to the ligand centre of geometry was greater than 22 Å. A similar sphere of water was employed to solvate the ligands in the unbound state. The bond angles and torsions of the protein side chains within 10 Å of any heavy atom of the ligand and all the

bond angles and torsions of the ligand were sampled during the simulation with the exception of rings. The bond lengths of the protein and ligand were kept rigid. In addition random rigid body translations and rotation of the ligand were performed (with a step size of 0.03 Å and 0.1 degrees respectively).The total charge of the system was brought to zero by neutralizing 2 lysine residues lying in the outer (frozen) part of the scoop (residue numbers 273 and 432). A 10 Å residue based cutoff was employed in all simulations. In the generalised Born simulations, a cutoff of 20.0 Å for the calculation of the Born radii was applied. To increase the efficiency of the generalised Born simulations, the techniques developed in chapter 3 were employed. In particular, a threshold of 0.005 Å to the recalculation of the Born energy was applied. A simplified sampling potential with a residue based cutoff of 6.0 Å, Born radii cutoff of 12 Å and threshold 0.05 Å was adopted and generated configurations were considered for the calculation of thermodynamic properties every 10 steps of this potential. A series of perturbations that transform one ligand into another was devised.

For the explicit solvent simulations in the bound state, solvent moves were attempted with a probability of 85.7%, protein sidechain move with a probability of 12.8% and solute move with a probability of 1.4%. In the unbound state, solvent moves were attempted 98.4% of the time. The solvent was equilibrated for 20 million (M) configurations to remove any repulsive contact with the solute(s). The system was then equilibrated in one end state (typically corresponding to the largest ligand) for 20M further moves where solute, protein and solvent moves were attempted. The resulting configuration was distributed over 12 values of the coupling parameter λ (0.00,0.10,...,0.90,0.95,1.00) and equilibrated for 10M moves before collecting statistics for 30M moves. In agreement with the experimental binding free energy measurement protocol, the simulations were carried out at a temperature of 37 °C.[149]

In the implicit solvent simulations, solute moves were attemped 10% of the time, with the remainder being protein sidechain moves. In the unbound state, 2 thousand (K) moves of equilibration were performed before 200 K moves of data collection. In the bound state, the system was pre-equilibrated at one value of λ

for 600 K moves. The resulting configuration was distributed over the 12 values of λ and further equilibration was performed for 100 K moves. Data was collected over the remaining 900 K moves.

This series of inhibitors was also scored using common empirical scoring functions described in the literature. The program GOLD[12] was used to obtain scores with the scoring functions Goldscore,[21] Chemscore[19,20] and ASP.[22] The compounds were scored on the basis of the modelled binding mode in the energy minimised X-ray structure of neuraminidase used in the free energy study. To avoid artefacts due to use of a different force field, the ligands were locally minimised according to each scoring function, prior to scoring.

5.4 Explicit solvent simulations results

The calculated relative binding free energies with the explicit solvent protocol for the series of DANA derivatives is shown in table 5.2. The results follows the experimental trends, but the binding free energies of the stronger inhibitors are overestimated.

Table 5.2: Comparison between experimental and calculated relative binding free energies and relative solvation free energies with the explicit solvent protocol[a]

Pert	Exp[b]	$\Delta\Delta G_{bind}$	$\Delta\Delta G_{solv}$	ΔG_{prot}	ΔG_{wat}	ΔG_{vac}
1t3	-2.67	-5.25 $\pm$ 0.62	-1.58 $\pm$ 0.67	21.87 $\pm$ 0.34	27.12 $\pm$ 0.52	28.70 $\pm$ 0.43
1t2	-1.63	1.44 $\pm$ 0.41	1.24 $\pm$ 0.32	1.87 $\pm$ 0.29	0.43 $\pm$ 0.29	-0.81 $\pm$ 0.14
2t3	-1.04	-7.19 $\pm$ 0.77	-2.52 $\pm$ 0.78	19.74 $\pm$ 0.43	26.93 $\pm$ 0.64	29.45 $\pm$ 0.45
3t4	-4.09	-4.14 $\pm$ 0.69	1.46 $\pm$ 0.54	-8.78 $\pm$ 0.51	-4.64 $\pm$ 0.47	-6.10 $\pm$ 0.27
2t4	-5.13	-9.32 $\pm$ 1.06	-2.53 $\pm$ 1.08	12.04 $\pm$ 0.65	21.36 $\pm$ 0.84	23.89 $\pm$ 0.68
2t9	0.08	-2.56 $\pm$ 1.21	0.84 $\pm$ 0.93	-7.30 $\pm$ 0.80	-4.74 $\pm$ 0.91	-5.58 $\pm$ 0.20
4t10	0.25	-5.46 $\pm$ 1.38	2.34 $\pm$ 1.09	-3.12 $\pm$ 0.90	2.34 $\pm$ 1.04	0.00 $\pm$ 0.31
9t10	-4.80	-11.86 $\pm$ 1.33	-1.49 $\pm$ 1.32	13.33 $\pm$ 0.81	25.19 $\pm$ 1.06	26.68 $\pm$ 0.78
3t6	-2.78	24.00 $\pm$ 1.35	-7.27 $\pm$ 1.30	8.68 $\pm$ 0.82	-14.64 $\pm$ 1.32	-8.05 $\pm$ 0.89
5t6	-3.45	-5.61 $\pm$ 0.58	-1.62 $\pm$ 0.62	20.69 $\pm$ 0.27	26.30 $\pm$ 0.51	27.92 $\pm$ 0.35
5t7	-5.15	-7.34 $\pm$ 1.18	0.10 $\pm$ 1.04	15.12 $\pm$ 0.78	22.46 $\pm$ 0.88	22.36 $\pm$ 0.56
6t7	-1.70	-1.36 $\pm$ 0.66	1.14 $\pm$ 0.53	-4.54 $\pm$ 0.49	-3.18 $\pm$ 0.44	-4.32 $\pm$ 0.29
7t8	0.65	-3.97 $\pm$ 1.39	3.07 $\pm$ 1.04	-4.62 $\pm$ 0.99	-0.65 $\pm$ 0.97	-3.72 $\pm$ 0.37

a Figures in kcal mol^{-1}

b Relative free energies are calculated using the formula $\Delta\Delta G = \Delta G_2 - \Delta G_1 = RTln(K_1/K_2)$ with the approximation that the ratio of the IC_{50} is equal to the ratio of the dissociation constants.[161]

In an attempt to assess the reliability of the simulation results, four different thermodynamic cycles were closed for the binding and solvation free energies. The resulting hystereses are shown in figure 5.3.

The hystereses are not as low as the simulation results reported in the previous chapter. In particular, the closure for compounds **2**, **3** and **4** is a bit high for both the solvation and binding free energies.

A difficulty arises in the perturbation of compound **3** into compound **6**. Crystallographic evidence suggests that the bulkier guanadino group of **6** must expel a crystallographic water that is present when **3** is bound. Initially, the perturbation was conducted in the presence of the crystallographic water. Towards the end of the perturbation, the guanadino group is sufficiently large such that the water molecule must move out of the way. To estimate the free energy change, a different protocol than the one described in the methods section was devised. Each window was equilibrated for 30M moves and data collection was then performed for 50M moves. The crystallographic bound water molecule was treated as a so-

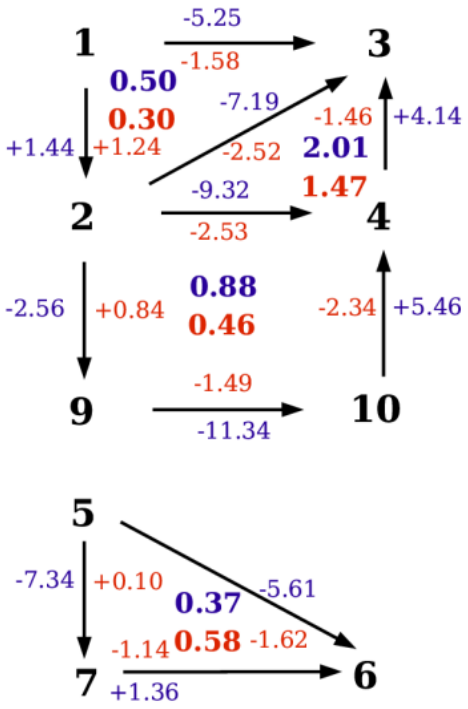


Figure 5.3: Explicit solvent protocol: the closure of four thermodynamic cycles for the calculated relative binding (blue) and solvation (red) free energies of the DANA analogues. All the figures are in kilocalories per mole.

lute molecule by ProtoMS. Large translational moves were allowed for this water molecule. It was hoped that this would permit the water to escape from the pocket where it usually lies, once the guanadino group is fully grown. The free energy change for the perturbation in the bound state was then found to be 8.68 $\pm$ 1.00 kcal mol^{-1}. The free energy change for the perturbation in the unbound state was calculated with the standard protocol and found to be -14.64 $\pm$ 1.32 kcal mol^{-1}. The relative binding free energy is thus 23.32 $\pm$ 1.65 kcal mol^{-1}. This quantity is very different from the experimental answer, which is -2.78 kcal mol^{-1}. Inspection of the trajectory snapshots and the energy components collected in the simulation show that the high free energy chance in the bound state is due to unfavourable Lennard Jones interactions between the crystallographic water and the guanadino group. Over the simulation time scale, the water is unable to leave the binding site and is trapped in a metastable state between the protein and ligand. Barillari and coworkers have studied a very similar system,[167] where the amino group is perturbed into a guanadino group. The only difference between the lig-

ands they studied and those presented here is the replacement of the amide moiety and the cis/trans substituents by a glycerol group. These groups interact with the protein on the other side of the binding pocket and do not interact directly with the amino or guanadino group. The replacement of the amino group by a guanadino group is also thought to displace crystallographic bound waters and the difference in binding free energy between the amino and guanadino derivatives is -2.3 kcal mol^{-1}. In light of the uncertainties in the experimental measurements, this is in good agreement with the measured change in binding free energy for our system (-2.78 kcal mol^{-1}). Barillari and coworkers employed a more elaborate simulation protocol where the whole protein is solvated into a box of TIP4P water and subject to periodic boundary conditions. In addition, small motion of the protein backbone was allowed. The perturbation was conducted in several steps, with the anihilation of crystallographic bound waters first (using the double decoupling methodology), followed by the perturbation of **3** into **6**. Barillari and coworkers reported that the calculated binding free energy was strongly dependent on the non bonded cutoff employed. With a non bonded cutoff of 10 Å, a relative binding free energy of 14.2 $\pm$ 1.2 kcal mol^{-1} was obtained. With a non bonded cutoff of 20 Å, the relative binding free energy dropped to -0.2 $\pm$ 1.2 kcal mol^{-1}. With a non bonded cutoff of 30 Å, the relative binding free energy was found to be -3.4 $\pm$ 1.1 kcal mol^{-1}.[167]

In light of the complexity of the protocol applied by Barillari et al., and the necessity to apply a large non bonded cutoff, which could not be done easily with the chosen method to solvate the protein, it was decided to adopt the results of Barillari et al. obtained at a cutoff of 30 Å for this particular perturbation.

Table 5.3 shows the free energy difference of the ligands with respect to compound **1**.

Table 5.3: Explicit solvent protocol: the Experimental and Calculated Binding free energies with respect to compound **1**.[a]

Compound	Perturbation pathway[b]	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
7	[**1t3**+**3t6**+**6t7**]	-10.01 $\pm$ 1.42	-7.15
4	[**1t3**+**3t4**];[**1t2**+**2t4**]	-8.64 $\pm$ 1.03	-6.76
8	[**1t3**+**3t6**+**6t7**+**7t8**]	-13.98 $\pm$ 1.99	-6.51
10	[**1t2**+**2t9**+**9t10**];[**1t3**+**3t4**+**4t10**]	-13.92 $\pm$ 1.75	-6.51
6	[**1t3**+**3t6**]	-8.65 $\pm$ 1.26	-5.45
3	[**1t3**]	-5.25 $\pm$ 0.62	-2.67
5	[**1t3**+**3t6**+**6t5**]	-3.04 $\pm$ 1.39	-2.00
9	[**1t2**+**2t9**]	-1.12 $\pm$ 1.28	-1.71
2	[**1t2**];[**1t3**+**3t2**]	1.69 $\pm$ 0.70	-1.63
1		0	0

a Figures in kcal mol^{-1}

b Figures obtained by summing free energy changes over different perturbations, and in some cases, averaging over two different pathways

The results for the explicit solvent simulations on this set of ligands is summarised in figure 5.4. At 3.38 kcal mol^{-1}, the MUE is relatively high. This is essentially because the binding energy of the two potent binders **10** and **8** is vastly overestimated. If these two compounds are excluded, the MUE drops to 2.21 kcal mol^{-1}. The results nonetheless follow closely the experimental trends and the coefficient of determination is 0.82 and the predictive index 0.93.

The poor quantitative agreement of some of the simulation results is of concern. In light of the observations of Barillari et al., it was decided that a non bonded residue cutoff of 10 Å might be not adequate for this sytem. The simulations were therefore repeated with a non bonded residue cutoff of 12 Å. A cutoff larger than 12 Å would require a different system setup, which includes a larger portion of the protein residues and a larger sphere of water to solvate the protein-ligand complex. The simulation results are reported in table 5.4.

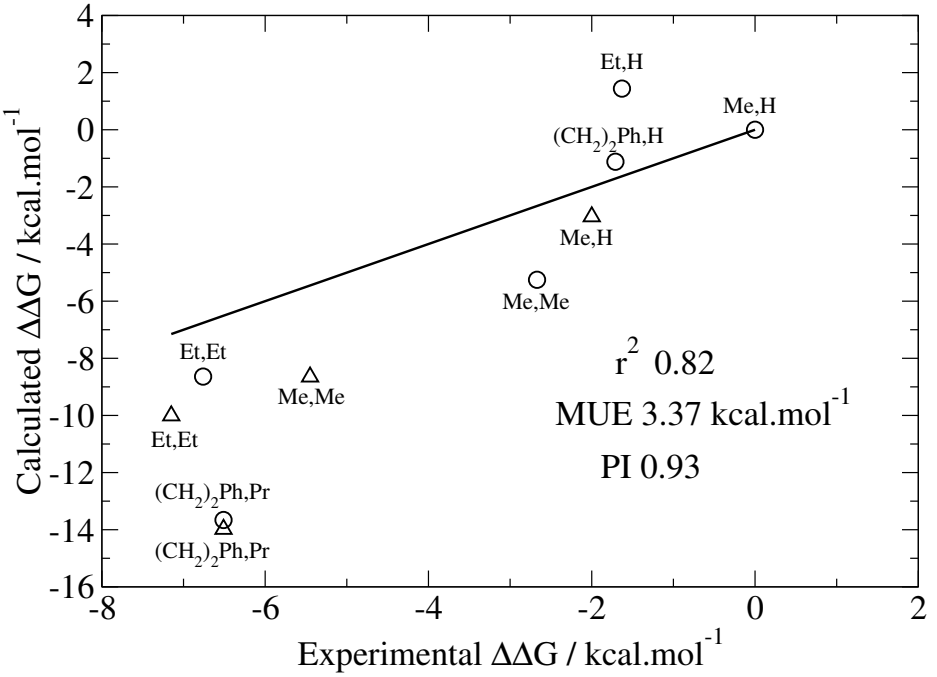


Figure 5.4: Summary of the explicit solvent protocol results. The amino ligands are represented with a circle, the guanadino ligands are represented with a triangle.

Table 5.4: Comparison between experimental and calculated relative binding free Energies and relative Solvation free Energies with the explicit solvent protocol and a non bonded cutoff of 12 Å [a]

Pert	Exp[b]	$\Delta\Delta G_{bind}$	$\Delta\Delta G_{solv}$	ΔG_{prot}	ΔG_{wat}
1t3	-2.67	-5.45 ± 0.61	-1.23 ± 0.67	22.02 ± 0.31	27.47 ± 0.52
1t2	-1.63	-3.44 ± 0.36	1.24 ± 0.34	-3.01 ± 0.18	0.43 ± 0.31
2t3	-1.04	-2.22 ± 0.74	-2.47 ± 0.77	24.76 ± 0.41	26.98 ± 0.62
3t4	-4.09	-6.28 ± 0.54	1.03 ± 0.52	-11.35 ± 0.30	-5.07 ± 0.45
2t4	-5.13	-5.80 ± 1.02	-2.15 ± 1.06	15.94 ± 0.62	21.74 ± 0.81
2t9	0.08	-0.57 ± 0.95	0.83 ± 0.96	-5.32 ± 0.78	-4.75 ± 0.94
4t10	0.25	-11.60 ± 1.26	2.52 ± 1.02	-9.08 ± 0.81	2.52 ± 0.97
9t10	-4.80	-10.84 ± 1.34	-1.99 ± 1.29	13.85 ± 0.86	24.69 ± 1.03
5t6	-3.45	-6.22 ± 0.58	-1.63 ± 0.64	20.07 ± 0.23	26.29 ± 0.53
5t7	-5.15	-12.84 ± 1.06	-0.29 ± 1.02	9.23 ± 0.63	22.07 ± 0.85
6t7	-1.70	-5.79 ± 0.58	0.41 ± 0.54	-9.70 ± 0.36	-3.91 ± 0.45
7t8	0.65	-4.09 ± 1.37	1.89 ± 1.09	-5.92 ± 0.90	-1.83 ± 1.03

a Figures in kcal mol^{-1}

b Relative free energies are calculated using the formula $\Delta\Delta G = \Delta G_2 - \Delta G_1 = RTln(K_1/K_2)$ with the approximation that the ratio of the IC_{50} is equal to the ratio of the dissociation constants.[161]

The closures of the thermodynamic cycles **1** to **2**, **2** to **3** and **3** to **1** are 0.21 and 0.00 kcal mol^{-1} for the binding and solvation free energy respectively. The closure for the thermodynamic cycle **2** to **3**, **3** to **4** and **4** to **2** are 2.70 and 0.71 kcal mol^{-1}. The closure for the cycle **2** to **4**, **4** to **10**, **10** to **9** and **9** to **2** are 5.99 and 1.53 kcal mol^{-1}. The closure for the cycle **5** to **6**, **6** to **7** and **7** to **5** are 0.83 and 0.93 kcal mol^{-1}. The hysteresis is once again high for the cycle involving compound **2**, **3** and **4** and particularly high for the 4 steps cycle, even considering the statistical error associated with each step. This suggests that these simulation results can not be interpreted with confidence. However, the two other cycles have reasonable hysteresis.

Table 5.5 shows the free energy difference of the ligands with respect to compound **1**. The results of Barillari et al were once again used for the perturbation of **3** to **6**.

Table 5.5: Explicit solvent protocol, cutoff 12 Å: the experimental and calculated binding free energies with respect to compound**1**.[a]

Compound	Perturbation pathway	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
7	[**1t3**+**3t6**+**6t7**]	-14.64 $\pm$ 1.38	-7.15
4	[**1t3**+**3t4**];[**1t2**+**2t4**]	-10.49 $\pm$ 0.95	-6.76
8	[**1t3**+**3t6**+**6t7**+**7t8**]	-18.73 $\pm$ 1.94	-6.51
10	[**1t2**+**2t9**+**9t10**]+[**1t3**+**3t4**+**4t10**]	-19.09 $\pm$ 1.59	-6.51
6	[**1t3**+**3t6**]	-9.89 $\pm$ 1.26	-5.45
3	[**1t3**]	-5.45 $\pm$ 0.61	-2.67
5	[**1t3**+**3t6**+**6t5**]	-2.63 $\pm$ 1.59	-2.00
9	[**1t2**+**2t9**]	-4.01 $\pm$ 1.02	-1.71
2	[**1t2**];[**1t3**+**3t2**]	-3.44 $\pm$ 0.66	-1.63
1		0	0

a Figures in kcal mol^{-1}

The results for the explicit solvent simulations on this set of ligands is summarised in figure 5.5. The MUE is even higher than for the simulations performed at a cutoff of 10 Å. This is once again because of compounds **10** and **8** whose binding free energy is vastly overestimated. The results still follow the experimental trends and the coefficient of determination is 0.79 and the predictive index

0.96.

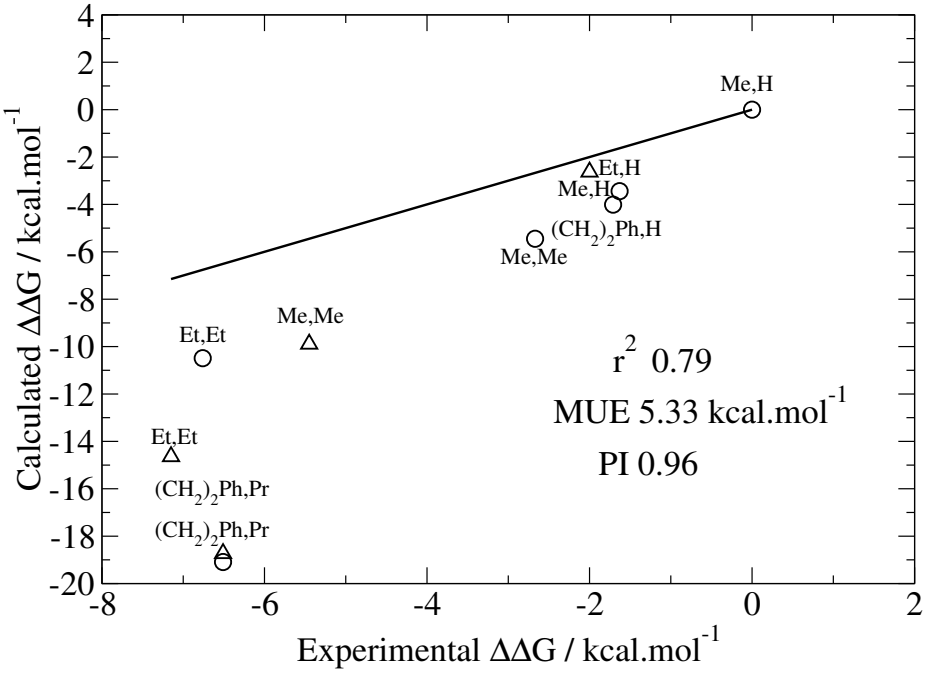
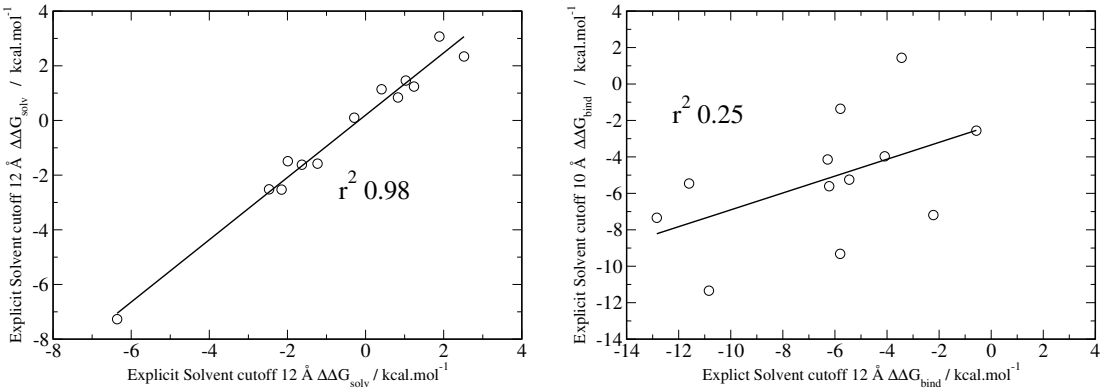


Figure 5.5: Summary of the explicit solvent protocol results with a non bonded cutoff of 12 Å. The amino ligands are represented with a circle, the guanadino ligands are represented with a triangle.

The free energies calculated with a cutoff of 10 or 12 Å are compared in figure 5.6. The solvation free energies are very similar, but the binding free energies exhibit marked differences. It was verified that this was not due to a lack of sampling by repeating some simulations. Similar results were obtained (within statistical sampling error) and this confirmed that the large differences cannot be attributed solely to incomplete sampling.

A detailed analysis of individual simulations is necessary to understand the origin of the very different binding free energies. The perturbation of compound **4** into **10** involves the addition of an extra phenyl and methyl group at the same time. The relative solvation free energy is very similar for the simulation carried out at a cutoff of 10 or 12 Å (+2.34 $\pm$ 1.02 and +2.52 $\pm$ 0.95 kcal mol^{-1}) but the relative binding free energies are very different (-5.46 $\pm$ 1.30 and -11.60 $\pm$ 1.18 kcal mol^{-1}). Figure 5.7 shows the free energy gradients recorded for the perturbations carried out in the bound and unbound state, at a cutoff of 10 and 12 Å.



(a) (b)

Figure 5.6: The correlation between predicted solvation and binding free energies by the two explicit solvent simulation protocols at a cutoff of 10 and 12 Å.
(a) Relative solvation free energies (b) Relative binding free energies

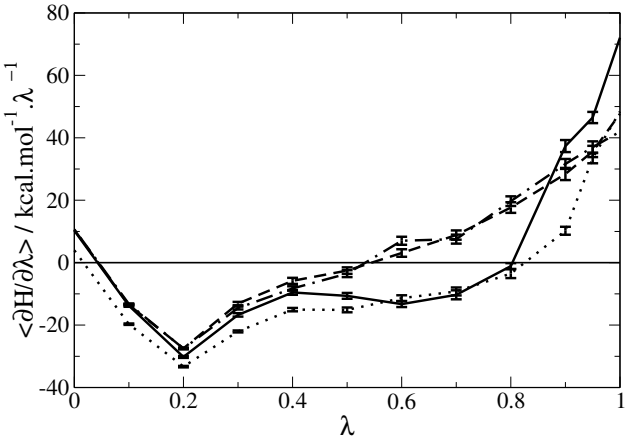


Figure 5.7: The free energy gradients recorded in the perturbation of **4** into **10**. The solid line corresponds to the perturbation carried out in the bound state with a cutoff of 10 Å. The dashed line is the perturbation in the unbound state with a cutoff of 10 Å. The dotted line is the perturbation in the bound state with a cutoff of 12 Å and the dashed-dotted line the perturbation in the unbound state with a cutoff of 12 Å.

It is seen that in the unbound state, the free energy gradients are identical to within statistical error. The gradients in the bound state, however, differ. The free energy gradients are more negative in the first half of the simulation when the perturbation is carried out with a cutoff of 12 Å. Most of the difference between the two simulations comes from the more rapid increase of the free energy gradients for the perturbation carried out with a cutoff of 10 Å in the later half of the simulation.

In an attempt to understant the origin of this difference, the simulation trajectory recorded at a value of λ 0.90 was visualised. Figure 5.8 shows an overlay of 10 snapshots saved during the simulation at the two different cutoff values.

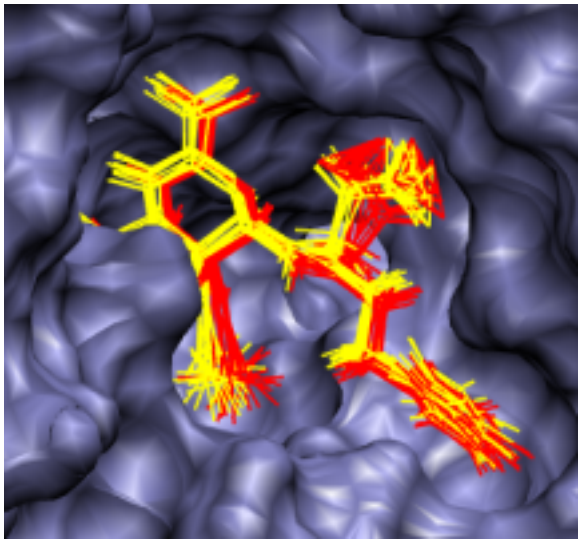


Figure 5.8: Overlay of 10 ligand snapshots sampled from a trajectory recorded at a value of λ set to 0.90. In red, snapshots from the simulation performed with a cutoff of 10 Å. In yellow, from the simulation performed with a cutoff of 12 Å. The solvent accessible surface area of the binding site is represented to indicate the position of the cis and trans binding pocket.

It is seen that at a cutoff of 10 Å, the ligand centre of geometry is shifted by about 0.5 Å with respect to the position of the ligand in the simulation carried out at a cutoff of 12 Å. Further, the propyl group occupies two alternative configurations in the cis pocket, while at a cutoff of 12 Å, no such behavior is observed. The position of the phenyl ring is essentially unchanged. Since, in the simulation performed at a cutoff of 10 Å, the system experiences a more rapid increase in its free energy gradient, and the ligand in this simulation is positioned closer to the edges of the cis pocket, it is possible that the difference in free energy gradients arises from unfavourable Lennard Jones contact with the amino acids that form the cis pocket. One way to verify this hypothesis is to plot the contribution of solute-protein Lennard Jones energy to the free energy gradients recorded during the simulation. This quantity is simply:

$$< \frac{\partial U_{LJ,sol-prot}}{\partial \lambda} >= \frac{< U_{LJ,sol-prot} >_{\lambda+d\lambda} - < U_{LJ,sol-prot} >_{\lambda-d\lambda}}{2\delta\lambda} \qquad (5.1)$$

Figure 5.9 plots this contribution as a function of the coupling parameter λ. The contribution of the solute-protein intermolecular Lennard Jones energy to the free energy gradients is very similar for the two systems until about λ = 0.70, at which point it increases more rapidly when the simulation is carried out at a cutoff of 10 Å. At a value of λ = 0.90, the difference between the two quantities is a little under 20 kcal $mol^{-1}.\lambda^{-1}$. This would account for about two-thirds of the difference in the free energy gradients seen in figure 5.7. The remaining difference could be attributed to other components in the force field, or complex coupling between different energy terms. It is worth mentioning that crystallographic evidence suggests that the cis pocket cannot accomodate cis substituents larger than a propyl group, and thus the positive solute-protein Lennard Jones contribution to the free energy gradients recorded at the end of the simulation indicates that the pocket has been filled. Since the Lennard Jones energy can become quickly strongly repulsive at short inter-atomic distances, a small displacement of circa 0.5 Å that brings the ligand closer to the cis pocket, could be enough to cause the more rapid increase in free energy gradients for the simulation conducted at a cutoff of 10 Å.

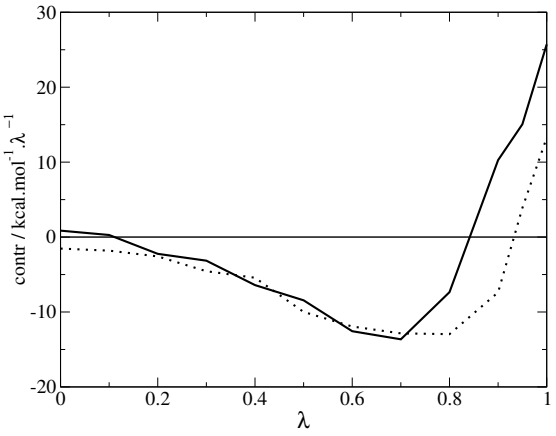


Figure 5.9: The contribution of the intermolecular Lennard Jones solute-protein energy to the free energy gradients recorded in the perturbation of **4** into **10**. The solid line is for the perturbation carried out at a cutoff of 10 Å and the dotted line for the perturbation carried out at a cutoff of 12 Å.

In the perturbation of compound **1** into **2**, identical relative solvation free energies are obtained (1.24 ± 0.30 and 1.24 ± 0.31 kcal mol^{-1}) , while the relative binding free energies differ markedly(1.44 ± 0.37 and -3.44 ± 0.32 kcal mol^{-1}). In the free energy gradients plots, reported in figure 5.10, it is seen that when the

perturbation is carried out with a non bonded cutoff of 12 Å, the free energy gradients are systematically more negative, and a large discrepancy between the two simulation protocols towards the latter half of the simulation occurs.

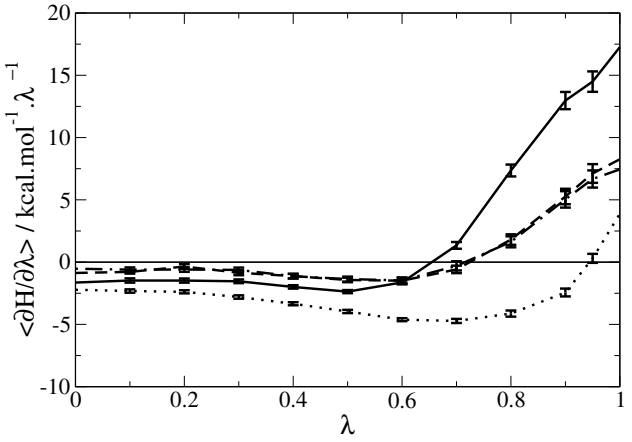


Figure 5.10: The free energy gradients recorded in the perturbation of **1** into **2**. The solid line corresponds to the perturbation carried out in the bound state with a cutoff of 10 Å. The dashed line is the perturbation in the unbound state with a cutoff of 10 Å. The dotted line is the perturbation in the bound state with a cutoff of 12 Å and the dashed-dotted line the perturbation in the unbound state with a cutoff of 12 Å.

The different binding free energies arise from the quite different interactions the ligand exhibit with the protein, as depicted in figure 5.11. When the simulation is run with a cutoff of 12 Å, the ethyl group of compound **2** fills the cis pocket. if the simulation is run with a cutoff of 10 Å, the ethyl group occupies the trans pocket. The ethyl group is too small to fill the larger trans pocket and it experiences less favourable intermolecular solute-protein Lennard Jones interactions.

This is confirmed by plotting the contribution of the intermolecular Lennard Jones solute-protein energy to the free energy gradients for this perturbation. From figure 5.12, it is clear that the ligand experiences more favourable Lennard Jones interaction if the cutoff is set to 12 Å. Furthermore, the difference in the contribution of the solute-protein intermolecular Lennard Jones energy to the free energy gradients accounts for practically all the differences in the free energy gradients between the two simulations.

Another interesting difference is observed in the explicit solvent simulations. When simulations are conducted with a cutoff of 12 Å, the residue arginine 371 occasionaly opens up to let a TIP4P water molecule interact with the other

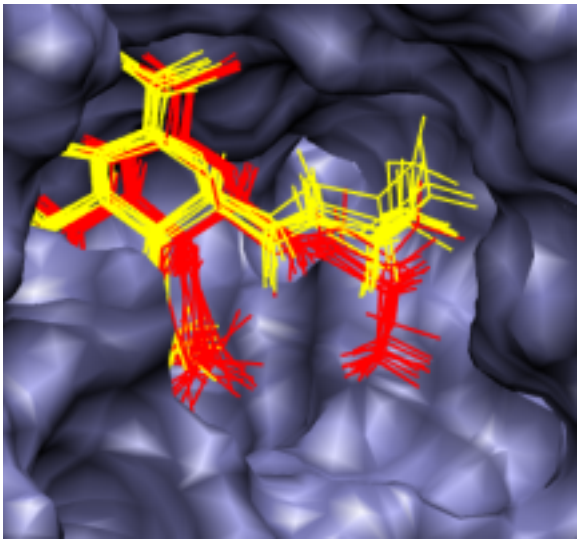


Figure 5.11: Overlay of 10 ligand snapshots sampled from a trajectory recorded at a value of λ set to 1.00 for the perturbation **1t2**. In red, snapshots from the simulation performed with a cutoff of 10 Å. In yellow, from the simulation performed with a cutoff of 12 Å. The solvent accessible surface area of the binding site is represented to indicate the position of the cis and trans binding pocket.

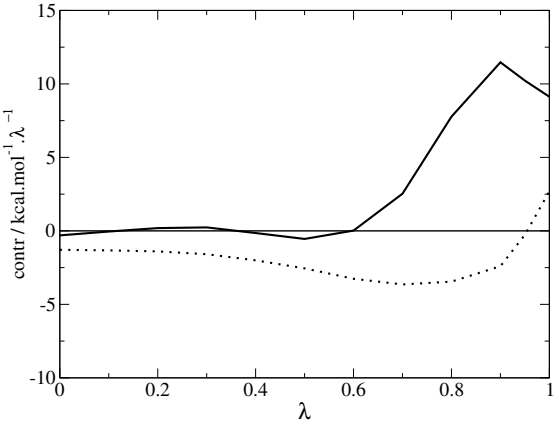


Figure 5.12: The contribution of the intermolecular Lennard Jones solute-protein energy to the free energy gradients recorded in the perturbation of **1** into **2**. The solid line is for the perturbation carried out at a cutoff of 10 Å and the dotted line for the perturbation carried out at a cutoff of 12 Å.

arginines and the carboxylate group of the ligand (see figure 5.13). These configurations are typically observed in 10 to 20 % of the trajectories that have been visualised. This behavior is however, not observed when the simulations are run with a cutoff of 10 Å, in which case, arginine 371 stays in a configuration that closely matches its position in the crystallographic structure.

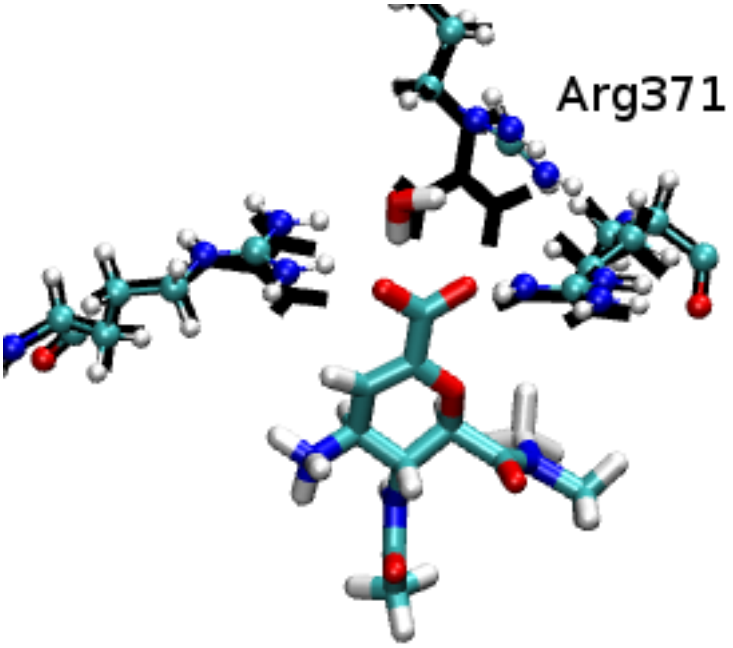


Figure 5.13: Opening of arginine 371 observed in trajectories generated with a non bonded cutoff of 12 Å. In gray, the position of the arginine triad in the crystallographic structure of neuraminidase (1BJI).

5.5 Generalised Born simulations results

The calculated relative binding free energies with the implicit solvent protocol for the same series of perturbations are presented in table 5.6. The results match closely the experimental trend and, surprisingly, are in much better quantitative agreement with experiment than was the case for the explicit solvent simulations.

Table 5.6: Comparison between experimental and calculated relative binding free energies and relative solvation free energies with the implicit solvent protocol[a]

Pert	Exp[b]	$\Delta\Delta G_{bind}$	$\Delta\Delta G_{solv}$	ΔG_{prot}	ΔG_{wat}
1t3	-2.67	-2.71 $\pm$ 0.55	-1.64 $\pm$ 0.51	24.35 $\pm$ 0.47	27.06 $\pm$ 0.28
1t2	-1.63	-0.21 $\pm$ 0.29	1.24 $\pm$ 0.19	0.22 $\pm$ 0.26	0.43 $\pm$ 0.13
2t3	-1.04	-2.32 $\pm$ 0.64	-2.63 $\pm$ 0.55	24.50 $\pm$ 0.55	26.82 $\pm$ 0.32
3t4	-4.09	-2.97 $\pm$ 0.45	0.95 $\pm$ 0.37	-8.12 $\pm$ 0.37	-5.15 $\pm$ 0.25
2t4	-5.13	-6.31 $\pm$ 0.72	-2.44 $\pm$ 0.80	15.14 $\pm$ 0.56	21.45 $\pm$ 0.45
2t9	0.08	2.09 $\pm$ 0.58	0.08 $\pm$ 0.26	-3.41 $\pm$ 0.56	-5.50 $\pm$ 0.16
4t10	0.25	2.22 $\pm$ 0.73	0.99 $\pm$ 0.40	3.21 $\pm$ 0.67	0.99 $\pm$ 0.30
9t10	-4.80	-8.19 $\pm$ 0.97	-0.96 $\pm$ 0.95	17.53 $\pm$ 0.74	25.72 $\pm$ 0.62
3t6	-2.78	-2.86 $\pm$ 1.13	-4.97 $\pm$ 1.21	-15.88 $\pm$ 0.79	-13.02 $\pm$ 0.81
5t6	-3.45	-2.95 $\pm$ 0.42	-1.87 $\pm$ 0.42	23.10 $\pm$ 0.34	26.05 $\pm$ 0.25
5t7	-5.15	-4.63 $\pm$ 0.93	-0.91 $\pm$ 0.71	16.82 $\pm$ 0.79	21.45 $\pm$ 0.50
6t7	-1.70	-2.23 $\pm$ 0.50	0.76 $\pm$ 0.37	-5.79 $\pm$ 0.43	-3.56 $\pm$ 0.25
7t8	0.64	0.84 $\pm$ 0.78	1.32 $\pm$ 0.37	-1.56 $\pm$ 0.67	-2.40 $\pm$ 0.39

a Figures in kcal mol^{-1}

b Relative free energies are calculated using the formula $\Delta\Delta G = \Delta G_2 - \Delta G_1 = RTln(K_1/K_2)$ with the approximation that the ratio of the IC_{50} is equal to the ratio of the dissociation constants.[161]

The closure of the four thermodynamic cycles, shown in figure 5.14 is reasonable, with the exception of the binding free energies for the 4 step cycle. The statistical sampling error for each step of this cycle is large (see table 5.6) and reflects the larger structural perturbations that have been attempted. It is therefore possible that a closure of 2 kcal mol^{-1} reflects the large statistical uncertainties of each individual simulation.

Table 5.7 shows the free energy difference of the ligands with respect to compound **1**.

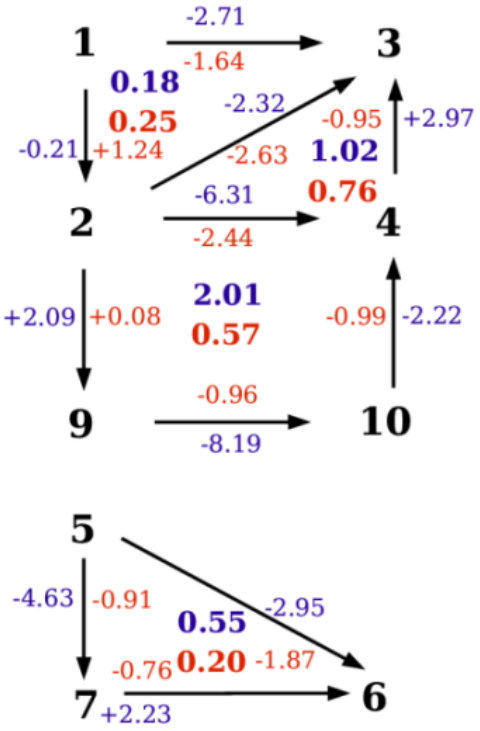


Figure 5.14: Implicit solvent protocol: the closure of four thermodynamic closure for the calculated relative binding (blue) and solvation (red) free energies of the DANA analogues. All the figures are in kilocalories per mole.

Table 5.7: Implicit solvent protocol: the experimental and calculated binding free energies with respect to compound **1**.[a]

Compound	Perturbation pathway	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
7	[**1t3**+**3t6**+**6t7**]	-7.80 $\pm$ 1.35	-7.15
4	[**1t3**+**3t4**];[**1t2**+**2t4**]	-6.10 $\pm$ 0.74	-6.76
8	[**1t3**+**3t6**+**6t7**+**7t8**]	-6.96 $\pm$ 1.56	-6.51
10	[**1t2**+**2t9**+**9t10**]+[**1t3**+**3t4**+**4t10**]	-4.89 $\pm$ 1.09	-6.51
6	[**1t3**+**3t6**]	-5.57 $\pm$ 1.26	-5.45
3	[**1t3**]	-2.71 $\pm$ 0.55	-2.67
5	[**1t3**+**3t6**+**6t5**]	-2.62 $\pm$ 1.33	-2.00
9	[**1t2**+**2t9**]	1.88 $\pm$ 0.65	-1.71
2	[**1t2**];[**1t3**+**3t2**]	-0.30 $\pm$ 0.56	-1.63
1		0	0

a Figures in kcal mol^{-1}

The results for the generalised Born simulations on this set of ligands are summarised in figure 5.15. The MUE at 1.01 kcal mol^{-1} is much better than that

obtained for the explicit solvent simulations. The coefficient of determination is 0.84 and is not significantly different from the explicit solvent results. The calculated predictive index stands at 0.95 and is nearly identical to that obtained with the explicit solvent protocol. Qualitatively, the explicit and implicit solvent protocols perform similarly. Quantitatively, the implicit solvent protocol performs significantly better.

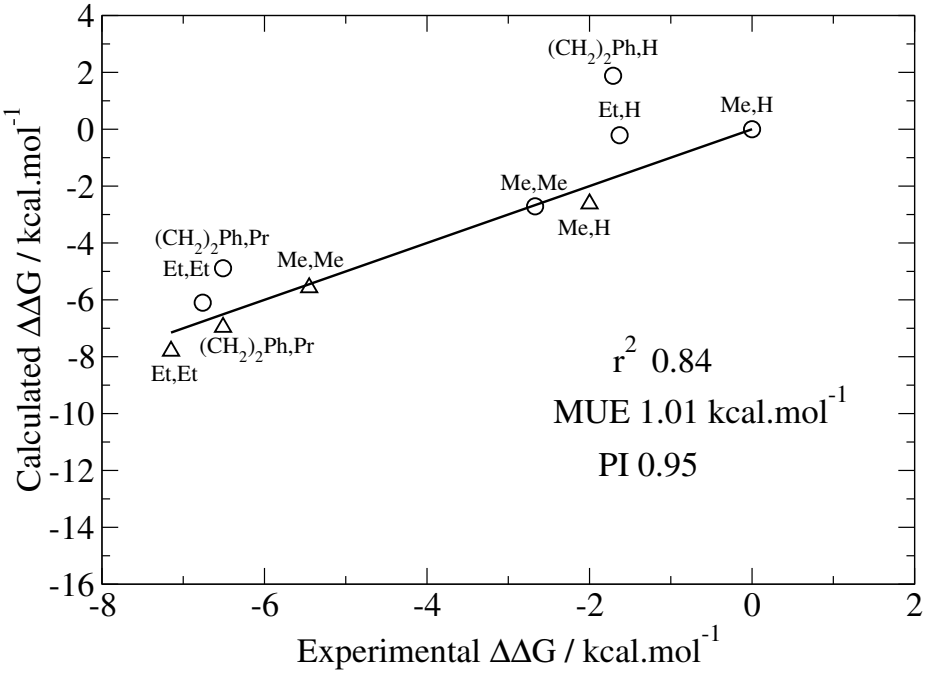


Figure 5.15: Summary of the implicit solvent protocol results. The amino ligands are represented with a circle, the guanadino ligands are represented with a triangle.

Since it was found in the previous section that the explicit solvent simulation results were sensitive to the non bonded cutoff, the same set of simulations was repeated with a cutoff of 12 Å. It was only necessary to run the simulations in the bound state, as the simulation results in the unbound state would be identical with a GBSA force field. The binding free energies thus obtained are summarised in table 5.8.

Table 5.8: Comparison between experimental and calculated relative binding free Energies with a cutoff of 12 and 10 Å [a]

Pert	Exp	$\Delta\Delta G_{bind}$ [b]	$\Delta\Delta G_{bind}$ [c]
1t3	-2.67	-3.00 $\pm$ 0.47	-2.71 $\pm$ 0.55
1t2	-1.63	-0.22 $\pm$ 0.29	-0.21 $\pm$ 0.29
2t3	-1.04	-2.59 $\pm$ 0.63	-2.32 $\pm$ 0.64
3t4	-4.09	-2.54 $\pm$ 0.50	-2.97 $\pm$ 0.45
2t4	-5.13	-5.96 $\pm$ 0.74	-6.31 $\pm$ 0.72
2t9	0.08	1.89 $\pm$ 0.60	2.09 $\pm$ 0.58
4t10	0.25	1.99 $\pm$ 0.72	2.22 $\pm$ 0.73
9t10	-4.80	-7.18 $\pm$ 1.01	-8.19 $\pm$ 0.97
3t6	-2.78	-1.11 $\pm$ 1.09	-2.86 $\pm$ 1.13
5t6	-3.45	-2.96 $\pm$ 0.47	-2.95 $\pm$ 0.42
5t7	-5.15	-5.71 $\pm$ 0.90	-4.63 $\pm$ 0.93
6t7	-1.70	-3.14 $\pm$ 0.47	-2.23 $\pm$ 0.50
7t8	0.64	2.12 $\pm$ 0.74	0.84 $\pm$ 0.78

a Figures in kcal mol^{-1}

b Implicit solvent protocol, cutoff 12 Å

c Implicit solvent protocol, cutoff 10 Å

The closure of the thermodynamic cycles **1** to **2**, **2** to **3** and **3** to **1** is 0.19 kcal mol^{-1}. The closure for the thermodynamic cycle **2** to **3**, **3** to **4** and **4** to **2** is 0.83 kcal mol^{-1}. The closure for the cycle **2** to **4**, **4** to **10**, **10** to **9** and **9** to **2** is 1.32 kcal mol^{-1}. The closure for the cycle **5** to **6**, **6** to **7** and **7** to **5** is 0.39 kcal mol^{-1}. These figures are all lower than those obtained with a cutoff of 10 Å and suggest the results can be interpreted with confidence. The relative binding free energies with respect to compound **1** are reported in table 5.9.

Table 5.9: Implicit solvent protocol with a cutoff of 12 Å: the experimental and calculated binding free energies with respect to compound **1**.[a]

Compound	Perturbation pathway[b]	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
7	[**1t3**+**3t6**+**6t7**]	-7.25 $\pm$ 1.27	-7.15
4	[**1t3**+**3t4**];[**1t2**+**2t4**]	-5.86 $\pm$ 0.74	-6.76
8	[**1t3**+**3t6**+**6t7**+**7t8**]	-5.13 $\pm$ 1.47	-6.51
10	[**1t2**+**2t9**+**9t10**]+[**1t3**+**3t4**+**4t10**]	-4.53 $\pm$ 1.10	-6.51
6	[**1t3**+**3t6**]	-3.84 $\pm$ 1.19	-5.45
3	[**1t3**]	-3.00 $\pm$ 0.47	-2.67
5	[**1t3**+**3t6**+**6t5**]	-1.15 $\pm$ 1.28	-2.00
9	[**1t2**+**2t9**]	1.67 $\pm$ 0.66	-1.71
2	[**1t2**];[**1t3**+**3t2**]	-0.32 $\pm$ 0.53	-1.63
1		0	0

a Figures in kcal mol^{-1}

b Figures obtained by summing free energy changes over different perturbations, and in some cases, averaging over two different pathways

The mean unsigned error is 1.32 kcal mol^{-1}. This is not very different from the value of 1.01 kcal mol^{-1}, obtained with a cutoff of 10 Å, particularly considering the very different results that were obtained for the explicit solvent simulations at two different cutoff values. The PI has increased to 0.97 because all the ligands except compound **9** have been correctly ranked. The results are summarised in figure 5.16.

In the previous section, it has been shown that increasing the non bonded cut-off from 10 to 12 Å did not improve the simulation results and that this caused the residue arginine 371 to exhibit unexpected configurations. In the generalised Born simulations, with a cutoff of 10 or 12 Å, no such behavior is observed and the arginine stays in the configuration it adopts in the crystal structure. This suggests that the previous behaviour is caused by an imbalance of electrostatic interactions between TIP4P water and this arginine. Long range electrostatics might indeed be a reason why the generalised Born results are much less sensitive to the non bonded cutoff than the explicit solvent results. Because the generalised Born energy is anti-correlated to the coulombic energy term, fluctuations in the long range coulombic term are dampened.[133] Another source of discrepancy might lie in the

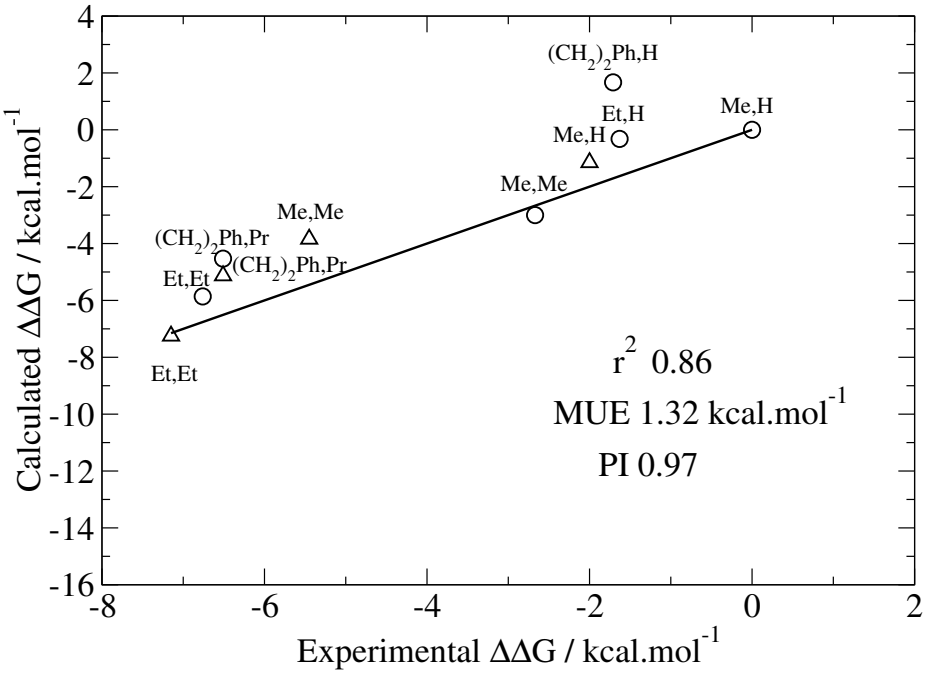
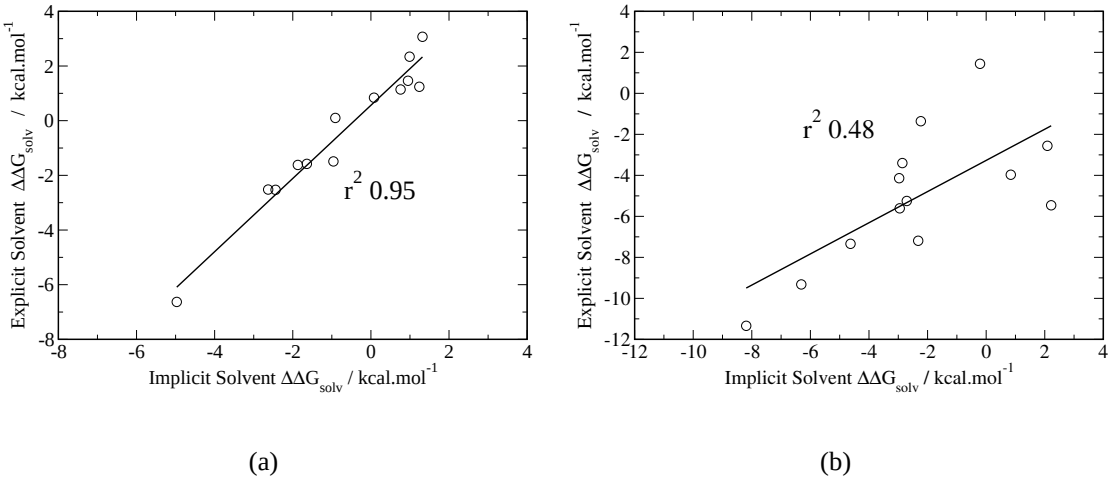


Figure 5.16: Summary of the implicit solvent protocol results obtained with a cut-off of 12 Å. The amino ligands are represented with a circle, the guanadino ligands are represented with a triangle.

solvation of the protein by a ball of TIP4P water molecules. While this treatment has been employed several times,[51–54] it is known to affect the calculated binding free energies by introducing boundary effects. If long range electrostatic effects are important in this system, a periodic box of water might be more appropriate. Note that this would increase the computational expense for the explicit solvent calculations by approximately one order of magnitude.

In figure 5.17, correlation plots between the generalised Born and TIP4P simulations for the solvation and binding free energies are reported. It is clear that both methodologies reproduce well the relative solvation free energies of the compounds in this set. The biggest discrepancy occurs for the perturbation **3t6**, where there is a difference of 1.6 kcal mol^{-1} between the two methods (bottom left corner of figure 5.17(a)). This perturbation involves the transformation of an amino group into a guanadino group and has a large associated statistical error with both solvation protocols. As most other perturbations involve the addition of extra non-polar groups, it is not unsurprising to observe very good agreement. In the perturbations of **4** to **10** and **7** to **8**, which consist of the addition of an extra phenyl and methyl group, the generalised Born simulations predict a relative solvation free

energy which is lower than the explicit solvent protocol by 1.3-1.8 kcal mol^{-1}. Although it is fair to note that the explicit solvent simulations have a high statistical sampling error (circa 1.0 kcal mol^{-1}), while the sampling error for the implicit solvent simulation is only 0.3-0.4 kcal mol^{-1}.



(a) (b)

Figure 5.17: The correlation between predicted solvation and binding free energies by the explicit and implicit solvent simulation protocols at a cutoff of 10 Å.
(a) Relative solvation free energies (b) Relative binding free energies

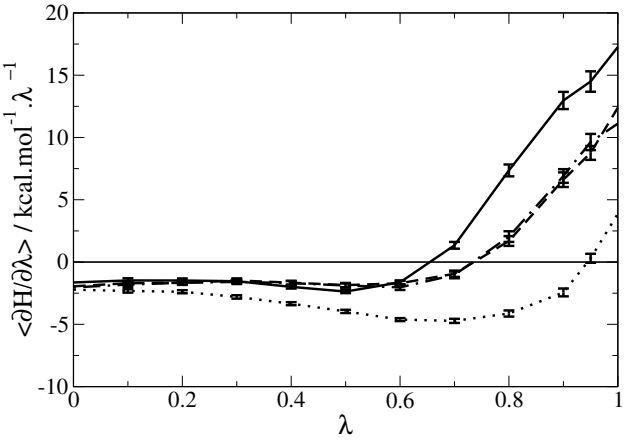


Figure 5.18: The free energy gradients recorded in the perturbation of **1** into **2** in the bound state. The solid line corresponds to the perturbation carried out in explicit solvent with a cutoff of 10 Å. The dashed line is the perturbation in implicit solvent with a cutoff of 10 Å. The dotted line is the perturbation in explicit solvent with a cutoff of 12 Å and the dashed-dotted line the perturbation in implicit solvent with a cutoff of 12 Å.

The strong decrease in correlation between the solvation and binding free energies warrants a detailed analysis of the simulation trajectories generated by the two methodologies. In figure 5.18, the free energy gradients for the perturbation

of **1** to **2** in the bound state in an implicit solvent model with a non bonded cutoff of 10 and 12 Å is reported. For visual emphasis, the results obtained with the explicit solvent simulation are also reported. The implicit solvent simulations are very similar, unlike the explicit solvent simulations. Interestingly, towards the end of the simulation, the free energy gradient increase is such that te final free energy gradients are intermediate between the two curves obtained for the explicit solvent simulations. Analysis of the simulation snapshots at a value of $\lambda = 1.0$ reveals the origin of the difference in the free energy gradients.

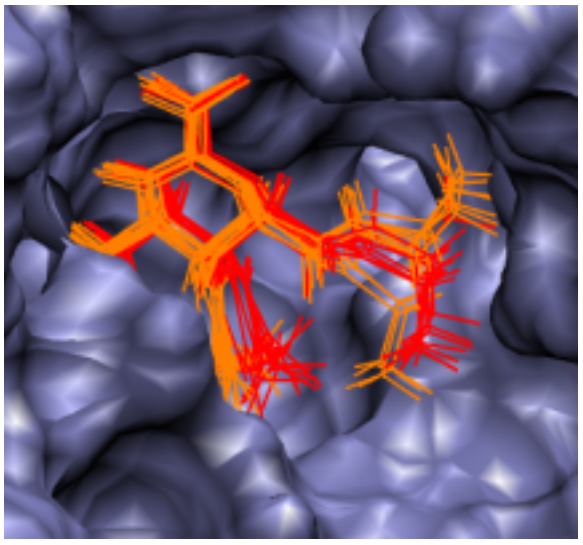


Figure 5.19: Overlay of 10 ligand snapshots sampled from a trajectory recorded at a value of λ set to 1.00 for the perturbation **1t2**. In orange, snapshots from the implicit solvent simulation performed with a cutoff of 10 Å. In red, from the explicit solvent simulation performed with a cutoff of 10 Å. The solvent accessible surface area of the binding site is represented to indicate the position of the cis and trans binding pockets.

Figure 5.19 shows an overlay of several ligand snapshots sampled regularly from the trajectory obtained with the implicit solvent methodology. For visual emphasis, the position of the ligand snapshots obtained during a simulation in explicit solvent at a cutoff of 10 Å is also reproduced. Recalling the results of the previous section, when running an explicit solvent simulation, the ethyl group is seen to occupy mainly a single pocket, depending on the cutoff employed. This resulted in very different free energy profiles. In the generalised Born simulations, the ethyl group on the ligand can sample both pockets during the simulations and the free energy profiles are not sensitive to the non bonded cutoff employed. Also, it can be seen that the position of the central ring of the ligand differs between the solvation

models. In the explicit solvent simulations, the ring is slightly tilted compared to the ring in the implicit solvent simulations and the amide group that bears the ethyl substituent is projected closer to the edges of the cis pocket by circa 0.8 Å.

For completeness, the contribution of the intermolecular Lennard Jones solute-protein energy to the free energy gradients for the generalised Born simulations performed at the two different cutoffs values is plotted in figure 5.20. For reference, the two explicit solvent profiles are also plotted.

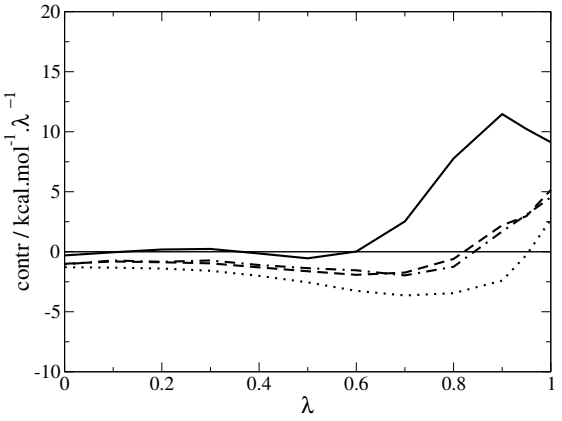


Figure 5.20: The contribution of the intermolecular Lennard Jones solute-protein energy to the free energy gradients recorded in the perturbation of **1** into **2**. The solid line is for the perturbation carried in explicit solvent with a cutoff of 10 Å and the dotted line at a cutoff of 12 Å. The dashed and dotted-dashed lines corresponds to the implicit solvent profiles, obtained at a cutoff of 10 or 12 Å respectively.

The following observations suggest that the origin of the differences in the free energy gradients between the implicit and explicit solvent approaches can be due to three factors. First, the solvation model can have such an influence on the potential energy surface that the two ligands adopt different configurations in the binding site. As a result, when the extra methyl group is grown, it experiences a different environment. The discrepancy between the two simulations would therefore be caused by force field effects. Second, in the explicit solvent simulations, conversion of the ligand between the two configurations can be hindered by the presence of several water molecules around the binding site. With simple Monte Carlo moves that randomly displace/rotate one water molecule at a time, it might be difficult for the solvent to let the ethyl group rotate freely. The origin of the differences between the free energy gradients would then be caused by an incomplete sampling of the thermally accessible states for the ligand in the binding site.

Third, a combination of the two.

The construction of the predictivity plot (figure 5.15) with respect to ligand **1** requires the perturbation of compound **3** into **6** to be performed. As was noted in the previous section, this requires the anihilation of a water molecule. Here no such difficulty is encountered since no water molecule has been modelled into the binding site with the generalised Born protocol and the perturbation does not show any conceptual difficulty. Intuitively, one would expect the implicit solvent simulation to yield results in disagreement with the observed change in binding free energy. This is because a crystallographic water molecule bridging interactions between the ligand and the protein should exhibit a behaviour very different from bulk water. The experimental change in binding free energy is -2.8 kcal mol^{-1}. With a simulation cutoff of 10 Å, the generalised Born simulation yields a result of -3.03 $\pm$ 1.27 kcal mol^{-1}, which is in very good agreement. With concern that this result might be due to luck, the simulation was extended for a further 900 K moves for each window. The final results, -2.86 $\pm$ 1.04 kcal mol^{-1} is not different. The same simulation was repeated with a different restart point and a cutoff of 12 Å. The obtained binding free energy was -0.93 $\pm$ 1.36. Extension of the simulation by 900 K more moves gives a free energy difference of -1.11 $\pm$ 1.12 kcal mol^{-1}. Thus the results with a cutoff of 12 Å are not as accurate as those with a cutoff of 10 Å. However they have to be considered in light of their large statistical uncertainty. In the previous section, the direct perturbation of **3** into **6** did not give good agreement with experiment, because the water was trapped between the ligand and protein and unable to escape to the bulk on the simulation time scale. Here as there is no explicit water, the guanadino group can be accomodated into the pocket without difficulty. It is tempting to argue that in the process of growing the guanadino group, a volume of high dielectric space has been replaced by a low dielectric space. Thus, to some extent, the desolvation of the pocket is taken into account by the generalised Born theory. It is surprising however that such a simple treatment of water expulsion would lead to a good agreement with experiment and in the absence of other systems to test the methodology, one must keep in mind that the good agreement between observed and calculated binding free energy change

might be fortuitous.

5.6 Influence of protein flexibility

Following the encouraging results obtained with COX-2 in the previous chapter, the impact of the protein flexibility on the free energy results obtained with the generalised Born protocol are investigated in this section. The fixed protein simulations were carried out for 300 K moves, with equilibration of 30 K at each window. A non bonded residue cutoff of 10 Å was used. The simulation duration was about 40-60% of the time taken by the simulations with protein flexibility.

Table 5.10: Comparison between experimental and calculated relative binding free energies, implicit solvent, cutoff 10 Å and rigid or flexible protein[a]

Pert	Exp	$\Delta\Delta G_{bind}$[b]	$\Delta\Delta G_{bind}$[c]
1t3	-2.67	-3.32 $\pm$ 0.40	-2.71 $\pm$ 0.55
1t2	-1.63	-1.13 $\pm$ 0.26	-0.21 $\pm$ 0.29
2t3	-1.04	-2.64 $\pm$ 0.53	-2.32 $\pm$ 0.64
3t4	-4.09	-3.68 $\pm$ 0.45	-2.97 $\pm$ 0.45
2t4	-5.13	-6.30 $\pm$ 0.69	-6.31 $\pm$ 0.72
2t9	0.08	-1.50 $\pm$ 0.47	2.09 $\pm$ 0.58
4t10	0.25	-1.41 $\pm$ 0.60	2.22 $\pm$ 0.73
9t10	-4.80	-8.24 $\pm$ 0.97	-8.19 $\pm$ 0.97
3t6	-2.78	-2.97 $\pm$ 1.18	-2.86 $\pm$ 1.13
5t6	-3.45	-3.53 $\pm$ 0.38	-2.95 $\pm$ 0.42
5t7	-5.15	-6.60 $\pm$ 0.84	-4.63 $\pm$ 0.93
6t7	-1.70	-3.38 $\pm$ 0.49	-2.23 $\pm$ 0.50
7t8	0.64	-2.94 $\pm$ 0.62	0.84 $\pm$ 0.78

a Figures in kcal mol^{-1}

b Implicit solvent protocol, no protein flexibility

c Implicit solvent protocol, protein flexibility

The closure of the thermodynamic cycles **1** to **2**, **2** to **3** and **3** to **1** is 0.45 kcal mol^{-1}. The closure for the thermodynamic cycle **2** to **3**, **3** to **4** and **4** to **2** is 0.02 kcal mol^{-1}. The closure for the cycle **2** to **4**, **4** to **10**, **10** to **9** and **9** to

2 is 2.03 kcal mol^{-1}. The closure for the cycle **5** to **6**, **6** to **7** and **7** to **5** is 0.31 kcal mol^{-1}. The cycle involving the larger substituents **9** and **10** always exhibit large deviations from good closure. This could be either because the sampling is insufficient or because there are inconsistencies in the models of each end state at the different perturbations. However, we do not find any such inconsistency in the input files.

The relative binding free energies with respect to compound **1** are reported in table 5.11.

Table 5.11: Implicit solvent protocol and rigid protein with a cutoff of 10 Å: the experimental and calculated binding free energies with respect to compound **1**.[a]

Compound	Perturbation pathway[b]	Calc $\Delta\Delta G_{bind}$	Exptl $\Delta\Delta G_{bind}$
7	[**1t3**+**3t6**+**6t7**]	-9.67 $\pm$ 1.34	-7.15
4	[**1t3**+**3t4**];[**1t2**+**2t4**]	-7.22 $\pm$ 0.67	-6.76
8	[**1t3**+**3t6**+**6t7**+**7t8**]	-12.61 $\pm$ 1.48	-6.51
10	[**1t2**+**2t9**+**9t10**]+[**1t3**+**3t4**+**4t10**]	-9.64 $\pm$ 0.98	-6.51
6	[**1t3**+**3t6**]	-6.29 $\pm$ 1.25	-5.45
3	[**1t3**]	-3.32 $\pm$ 0.40	-2.67
5	[**1t3**+**3t6**+**6t5**]	-2.76 $\pm$ 1.31	-2.00
9	[**1t2**+**2t9**]	-2.63 $\pm$ 0.53	-1.71
2	[**1t2**];[**1t3**+**3t2**]	-0.91 $\pm$ 0.46	-1.63
1		0	0

a Figures in kcal mol^{-1}

b Figures obtained by summing free energy changes over different perturbations, and in some cases, averaging over two different pathways

The MUE is 1.90 kcal mol^{-1}, which is higher than when the calculations are performed with protein flexibility enabled. This is mainly because the binding affinity of **8** and **10** is now largely overestimated. In general, the simulations now favour the largest compounds, and the binding free energies of compound **8**, **9** and **10** are now more negative by about 4 to 5 kcal mol^{-1}. Other less bulky substituents exhibit the same trend, but to a lesser extent (**7** more negative by about 2 kcal mol^{-1}, **4** more negative by about 1 kcal mol^{-1}, **3** and **2** more negative by about 0.60 kcal mol^{-1}.

This origin of this gain in binding affinity can be directly attributed to the rigidity of the protein sidechains. In a simulation with protein side chain flexibility enabled, as the ligands become larger, they occupy a larger portion of the binding site and they limit the number of configurations the protein side chains surrounding the binding site can adopt. This configurational restriction should translate in a loss of entropy, disvafouring the larger ligands. As this penalty is not considered if the protein side chain are considered rigid, the larger substituents are more favoured. The protein binding site was energy minimised in presence of the largest compound **10** in this series before conducting the Monte Carlo simulations and no steric clashes between the ligands and the protein side chains occur. This may also bias the simulation results towards the larger compounds.

In spite of these differences, the coefficient of determination is 0.80 and the predictive index still stands at a very high value of 0.96, and the qualitative ranking of the inhibitors is still excellent.

The results are summarised in figure 5.21.

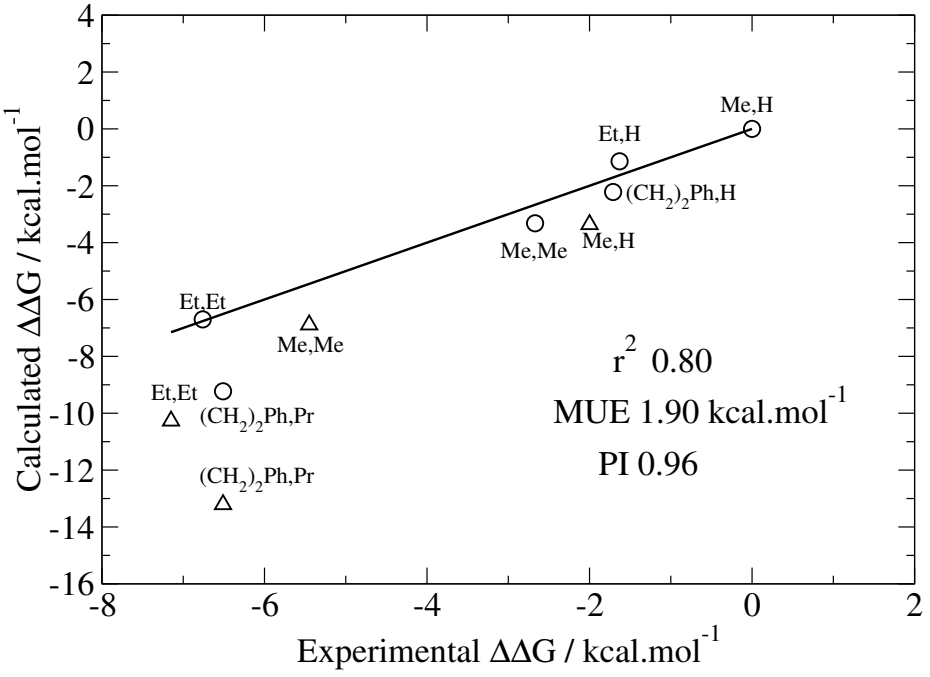
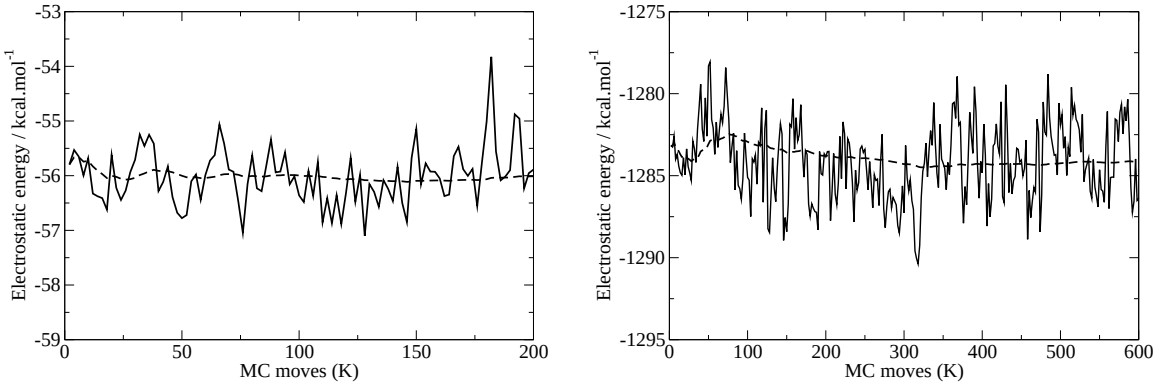


Figure 5.21: Summary of the results of the implicit solvent protocol with a rigid protein. The amino ligands are represented with a circle, the guanadino ligands are represented with a triangle.

5.7 Importance of configurational averaging

Some workers have suggested incorporating solvation effects into empirical scoring functions by calculating the solvation free energy of a ligand, protein and ligand-protein complex.[165,168] In most empirical scoring functions, a single configuration of the ligand bound in the protein binding site is usually considered. In figure 5.22, the total electrostatic energy (Coulombic and generalised Born energy) of compound **1** in the unbound state and bound to the protein is plotted as a function of the number of Monte Carlo moves.



(a) (b)

Figure 5.22: The fluctuations in the electrostatic energy during the simulation of compound **1**
(a) Isolated in solution (b) Bound to neuraminidase

It is important to remember that the configurations generated are those that are thermally accessible to the system at a temperature of 37 °C. Because the ligand is relatively stable in the binding site, and was manually docked such that it reproduces the binding mode of an analogue, the vast majority of the ligand configurations would be equivalent to acceptable docking results and could have been used to obtain a score based on that single configuration. From the plots above, it is seen that the solvation energies fluctuate significantly. Even in the unbound state, which consists of the ligand isolated in solution, the electrostatic energy can fluctuate by 1 to 2 kcal mol^{-1}. Since all the points along these trajectories would be a suitable candidate for scoring and yet the electrostatic energy fluctuates significantly, any binding energy score obtained from a single snapshot analysis would

arguably be unreliable. The MM/PBSA method may avoid these issues to some extent as it typically averages the solvation energy of 100-200 hundred snapshots extracted from an explicit solvent trajectory. It is not clear however if such a low number of snapshots would be enough to obtain precise results.

5.8 Computational cost and convergence

To assess the efficiency of each methodology employed to rank the compounds studied in this chapter, the convergence of the mean unsigned error, the predictive index and the closure of the 4 thermodynamyc cycles is plotted in the following figures. Results are reported for the simulations performed at a cutoff of 10 Å only. Similar behaviour is observed at a cutoff of 12 Å.

For the explicit solvent simulations, after 10 hours of simulation, the mean unsigned error stabilises around 2.5 kcal mol^{-1}. However, it steadily increases after 18 hours. This suggest that all the calculated individual free energy differences may not be fully converged. The opposite behavior is observed with the implicit solvent simulations, and the mean unsigned error peaks at 2 kcal mol^{-1} after about 3 hours and then steadily decreases to 1.08 kcal mol^{-1} after 9 and a half hours, when the simulations were stopped. Ideally, these simulations should have been run longer. The MUE is different from that reported in the previous section (1.01 kcal mol^{-1}), because the perturbation of **3** into **6** was conducted for twice as long as the timings shown in this graph. For a fair comparison, only the free energy results for the first half of this perturbation have been used to construct this plot. If no protein flexibility was allowed, the mean unsigned error is seen to rapidly oscillate around 1.75 kcal mol^{-1} after about 3 hours.

The convergence of the predictive indices is plotted in figure 5.24. The PI obtained with the explicit solvent simulations is stable after about 8 hours of simulation. Both implicit solvent protocols yield relatively stable PIs very quickly (in about one hour).

It is worth recalling that neither figure 5.23 or 5.24 could be used to decide when the simulations have run for long enough, as in a practical application, these quantities cannot be calculated without the *a priori* knowledge of the experimental

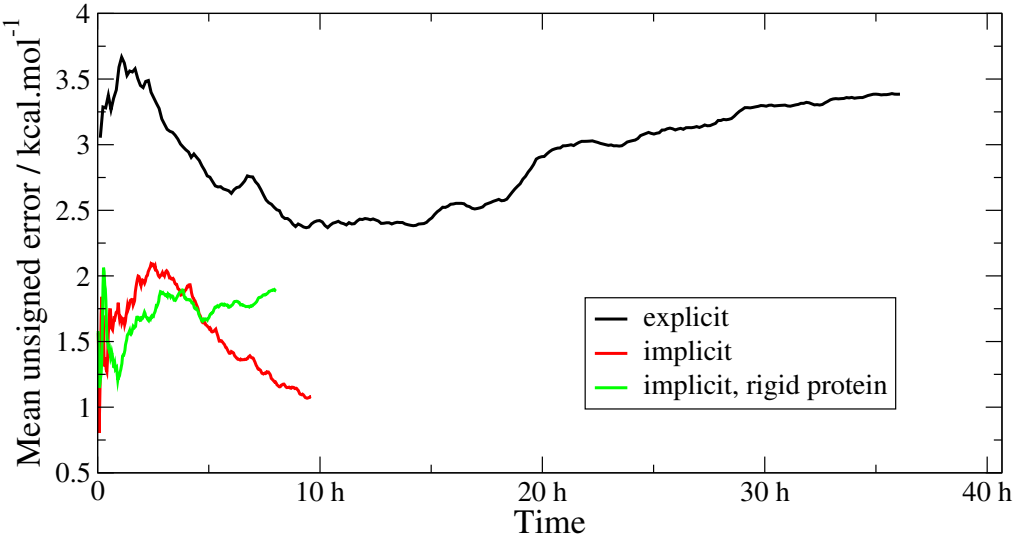


Figure 5.23: The convergence of the mean unsigned error as a function of the time taken to complete a single simulation at one value of λ.

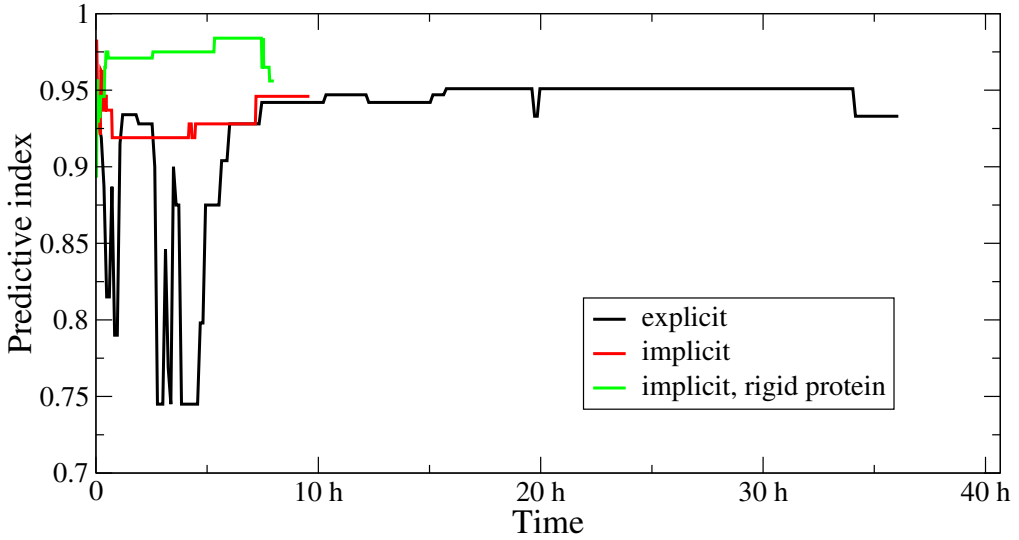


Figure 5.24: The convergence of the predictive index as a function of the time taken to complete a single simulation at one value of λ.

binding free energies. In figure 5.25, the hysteresis of the 4 thermodynamic cycles is plotted as a function of time, for the three different protocols. Some cycles converge more readily than others. For example, a low hysteresis is rapidly achieved with every protocol for the cycle involving compounds **1**, **2** and **3**, while significant fluctuations of the hysteresis are observed for the cycle between compounds **2**, **3** and **4**. In addition, the deviations are large (greater than 1 kcal mol^{-1}) when the simulations were stopped. Also, in the explicit solvent simulations, the hysteresis of this cycle can be seen to reduce steadily to 0.5 kcal mol^{-1} during the first 20

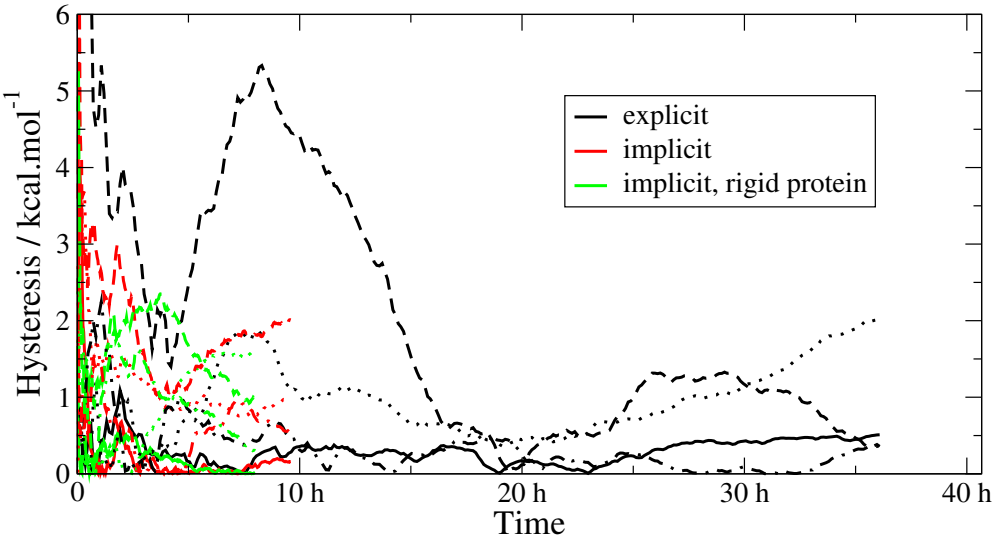


Figure 5.25: The convergence of the closure of the thermodynamic cycles for the binding free energies as a function of the time taken to complete a single simulation at one value of λ. The solid lines are for the cycle involving compounds **1**, **2**, **3** . The dotted lines for the cycle involving compounds **2**, **3** and **4**. The dashed lines for the cycle involving compounds **2**, **4**, **9** and **10**. The dashed-dotted lines for the cycle involving compounds **5**, **6** and **7**.

hours of simulation, before increasing again up to 2.0 kcal mol^{-1} at the end the simulation.

Despite the difficulty in obtaining well behaved hysteresis on this system, it is worth mentionning that qualitative and quantitative results can be obtained much more quickly, owing presumably to a cancellation of errors between individual free energy differences.

5.9 Comparison with empirical scoring functions

Predictive indices for the series of neuraminidase inhibitors have been computed using the Chemscore, GoldScore and ASP scoring function. The results are presented graphically in figures 5.26, 5.27 and 5.28. Since it is unclear how the scores predicted by these methods can be related to binding free energies, quantitative descriptors such as the mean unsigned error or the coefficient of determination were not calculated.

The predictive index of Chemscore is 0.00. This means that the ordering of the compounds is random. This is essentially due to the fact that chemscore penalises greatly the guanadino compounds with respect to the amino derivatives.

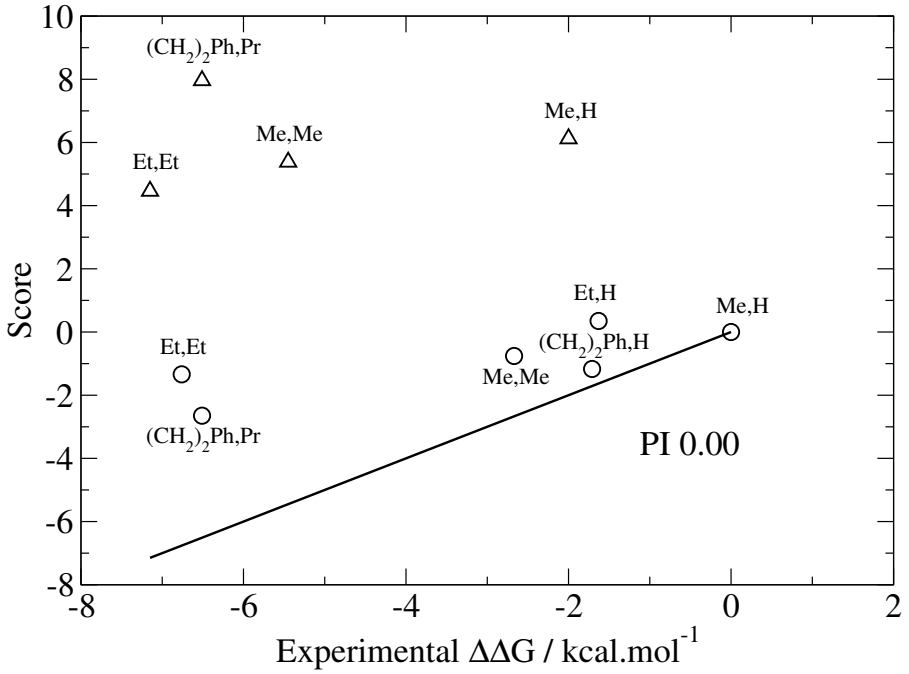


Figure 5.26: Chemscore: Calculated score and experimental binding free energy of a series of neuraminidase inhibitors. All the data is relative to compound **1**

Even within the same family, the discrimination is poor. For example, compound **9** has a score similar to compound **4** while in reality the latter is a stronger binder by more than 5 kcal mol^{-1}.

By contrast, Goldscore performs fairly well on this set. The main discrepancy is the relatively high score of the phenyl derivatives which causes the affinity of compound **9** to be overestimated relative to the other derivatives. The free energy methods discussed in the previous sections generally did not yield a large binding affinity increase for the introduction of a phenyl group, in agreement with experimental trends. The plot observed here suggest that Goldscore essentially favours compounds with the largest number of atoms. Finally, the difference of affinity between the amino and guanadino groups is well reproduced by Goldscore.

The results obtained with ASP are very similar to those from Goldscore. The main difference is that the guanadino compounds are more favoured with ASP than Goldscore. The predictive index stands at 0.77

The results presented here contrast sharply with those obtained for the series of COX-2 inhibitors, where Chemscore was significantly superior to either Goldscore or ASP. This suggest that micro/nano molar inhibitors cannot be ranked consistently across different targets by a scoring function.

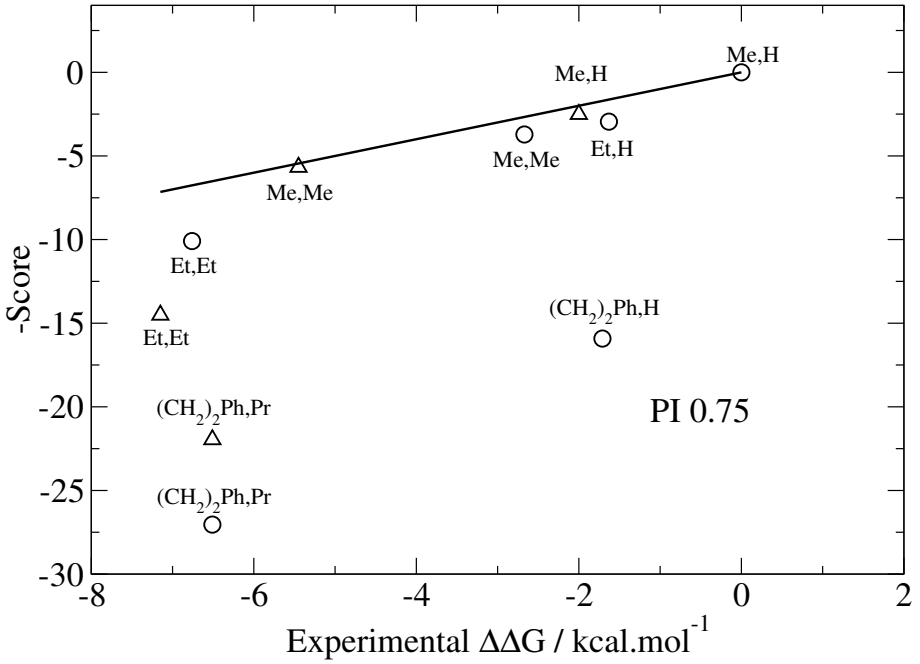


Figure 5.27: Goldscore: Calculated score and experimental binding free energy of a series of neuraminidase inhibitors. All the data is relative to compound **1**. The negative of the Goldscore is plotted such that if the method can explain the variation of the relative binding free energies, a positive correlation would be observed.

5.10 Conclusion

The relative binding free energies of a series of neuraminidase inhibitor derivatives of DANA have been calculated by means of explicit solvent (TIP4P) and implicit solvent (generalised Born) free energy simulations. The results of the implicit solvent simulations, with a mean unsigned error of 1.01 kcal mol^{-1} and a coefficient of determination of 0.84 are in quantitative agreement with the experimental measurements. The results obtained with the explicit solvent simulations, with a mean unsigned error of 3.34 kcal mol^{-1} deviate significantly from the experimental data. However, both methodologies predict the qualitative trends in the binding affinity for the series of inhibitors and yield a predictive index of 0.96 and 0.95 respectively. The explicit solvent simulation results are shown to be very sensitive to the non bonded cutoff and increasing this to 12 Å leads to a mean unsigned error of 5.33 kcal mol^{-1}, in poorer quantitative agreement with experiment. It is shown that a different non bonded cutoff leads to a sampling of different protein-ligand states in the binding site, which explains the observed large discrepancies. By contrast, the generalised Born simulations exhibit a much reduced sensitivity to the

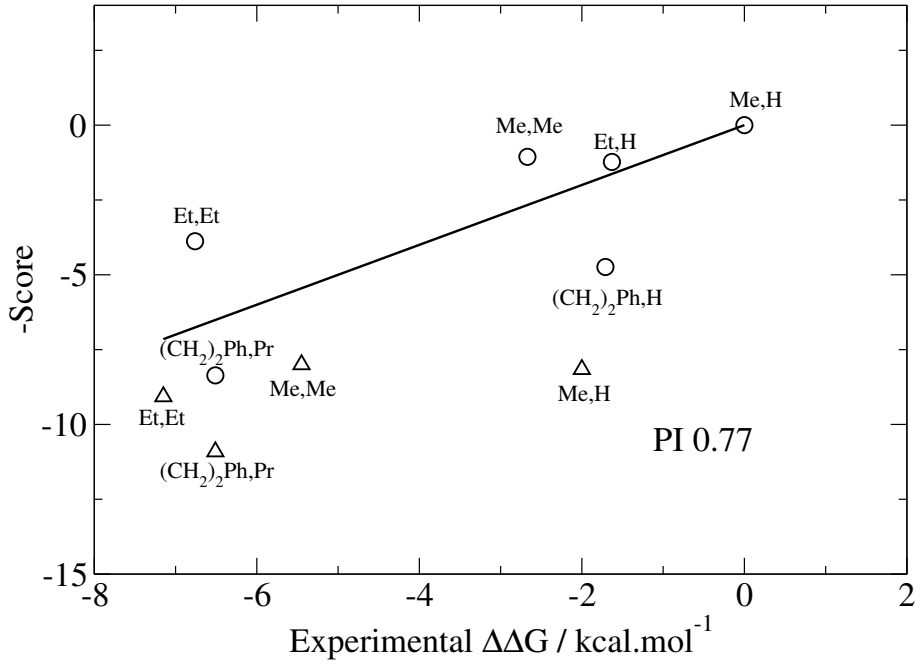


Figure 5.28: ASP: Calculated score and experimental binding free energy of a series of neuraminidase inhibitors. All the data is relative to compound **1**, and the negative of the score is plotted for reasons similar to the Goldscore method.

non bonded cutoff. The origin of the discrepancies between explicit and implicit solvent protocols is attributed to force field effects and the different ensembles of protein-ligand states that are formed by the two methodologies. In light of the high degree of exposure of the binding site to the solvent, and the presence of crystallographic water molecules that mediate interactions between the protein and the ligand, the performance of the generalised Born simulations is truly remarkable. It is then shown that neglecting protein side chain flexibility does not modify the qualitative ranking of the inhibitors in this series, although the quantitave agreement is worsened due to a systematic increase in binding affinity for the largest compounds. It is unclear if the simulation protocol allowed sufficient Monte Carlo moves to obtain fully converged results. High, relatively stable PIs are obtained in about 8 hours of simulation with the explicit solvent protocol and in just 1-2 hours with the implicit solvent protocols (flexible or rigid protein side chains). Finally, the ability of commonly used empirical scoring functions to rank these compounds correctly has been assessed by calculating predictive indices for Chemscore, Goldscore and ASP. The PIs of 0.00, 0.75 and 0.77 are significantly lower and very different from those recorded on the previous system cyclooxygenase-2, suggest-

ing that the empirical scoring functions tested here do not produce results of a consistent quality.

Chapter 6

Alternative pathways in free energy calculations

"Imagination is more important than knowledge."

Albert Einstein

6.1 Introduction

One important limitation that most rigorous free energy methods suffer is that, in order to calculate the relative binding free energy of two ligands, it is necessary for the two molecules to be very similar in two aspects. First, they should have a similar potential energy surface such that the FEP or TI equations converge the free energy difference readily. Second, they should be structurally similar such that it is simple to devise a change in the internal coordinates of one ligand that convert it into the other species (with the appropriate modifications of force field parameters). In an effort to relax this second constraint, alternative methods of coupling two molecular species in the calculation of their relative free energies are investigated in this chapter.

6.2 Single and dual topology methods

In a most general fashion, when implementing relative free energy calculations in a computer program, it is necessary to devise a scheme to transform the poten-

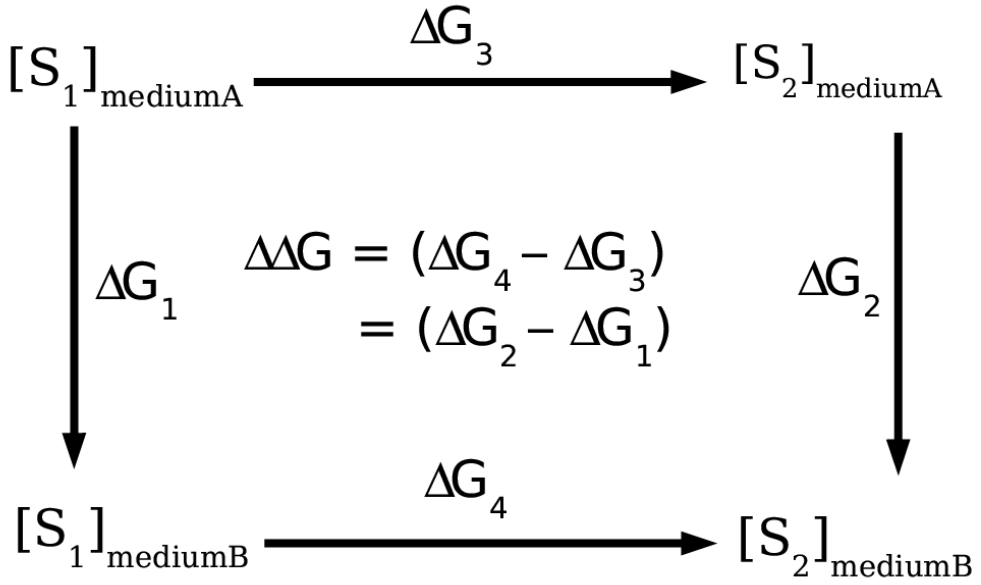


Figure 6.1: A general thermodynamic cycle that relates the difference in free energy between S_1 and S_2 in two media A and B. S_1 and S_2 could be two small molecules and medias A and B, water and vacuum, in which case the double free energy difference will correspond to the relative hydration free energy of S_2 with respect to S_1. If the media A and B represents a solvated protein and pure water, then the double free energy difference will correspond to the relative binding free energy of S_2 with respect to S_1. While the horizontal processes corresponding to ΔG_3 or ΔG_4 are often measured experimentally, the vertical processes corresponding to ΔG_1 or ΔG_2 are usually easier to calculate in a computer simulation.

tial energy function of a system S_1 into the potential energy function of system S_2. Calculated single free energy differences are often not directly comparable to experiment if they are not related before to a reference state and this is usually accomplished through the construction of a thermodynamic cycle. A simple and general thermodynamic cycle is highlighted in figure 6.1.

To calculate the free energy differences in figure 6.1 by free energy perturbation (equation 1.28) or thermodynamic integration (equation 1.30) it is necessary to define the potential energy function that represents the interactions of S_1 or S_2 with its surrounding medium. Typically, the two systems are coupled to each other through the introduction of a parameter λ and a suitable form for a potential energy function is:

$$U(\lambda) = U_0 + \Delta U(\lambda) \tag{6.1}$$

where U_0 represents the energy terms that are not related to S_1 or S_2 (e.g, the interactions between solvent molecules). The second term depends on λ which

often varies between 0 and 1 such that $\Delta U(0) = U(S_1)$ and $\Delta U(1) = U(S_2)$. The exact nature of the coupling is arbitrary as long as the end states (the first and last states defined by λ) corresponds rigorously to the two systems of interest.

One method to accomplish this is to define each of the force field terms in $\Delta U(\lambda)$ as a linear combination of the values of the force field term of system S_1 and S_2. For example, the angle stretching term can be expressed as:

$$\begin{aligned} Uang(\lambda) &= K_\theta(\lambda)[\theta - \theta_{eq}(\lambda)]^2 \\ K_\theta(\lambda) &= \lambda K_\theta(S_2) + (1-\lambda)K_\theta(S_1) \\ \theta_{eq}(\lambda) &= \lambda\theta_{eq}(S_2) + (1-\lambda)\theta_{eq}(S_1) \end{aligned} \tag{6.2}$$

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

$$\begin{aligned} U_{coul}(\lambda) &= \frac{q_i(\lambda)q_j}{4\pi\varepsilon_0 r_{ij}(\lambda)} \\ q_i(\lambda) &= \lambda q_i(S_2) + (1-\lambda)q_i(S_1) \\ r_{ij}(\lambda) &= \lambda r_{ij}(S_2) + (1-\lambda)r_{ij}(S_1) \end{aligned} \tag{6.3}$$

Equation 6.3 emphasises that geometric terms as well as force field terms vary in the coupling of S_1 and S_2. A difficulty with this method is encountered when the two systems do not have the same number of atoms. In this case it is necessary to introduce a dummy atom in one of the end state. In the end state where it should not exist, the dummy atom should not contribute at all to the intermolecular energy. It is however often necessary to retain some intramolecular energy terms associated with the dummy atom. In the absence of bond or angle terms, the dummy atom would be able to dissociate completely from the molecule it is attached to. This phenomenon invariably leads to divergence of the calculated free energy difference and is therefore undesirable.[169,170] A solution is to associate a bond or angle term with the dummy atom throughout the perturbation. The influence of that extra term on the calculated free energy will normally cancel out in the double free

energy difference. Because this method requires the presence of the same number of particles in the two systems, it is often called the single topology method.

Another drawback of the single topology method is that it is necessary to specify internal coordinate changes such that the system S(λ) matches the topology of S_1 or S_2 in the two end states. This operation can be trivial (for example, perturbing an hydrogen into a methyl group requires the elongation of a C-H bond to a C-C bond and the introduction of three dummy atoms linked to the hydrogen atom) or very difficult if one tries for example to couple two molecules with a completely different topology.

It is possible to define a coupling scheme by adopting a different approach. Instead of linearly perturbating the force field parameters of a system S(λ), one can define simultaneously S_1 and S_2 in the medium. The potential energy function becomes:

$$U(\lambda) = U_0 + \lambda U(S_2) + (1-\lambda)U(S_1) \tag{6.4}$$

And for example the coulombic energy of atoms i from S_2, i' from S_1 with an atom j from the medium would be:

$$U_{coul}(\lambda) = \lambda \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + (1-\lambda)\frac{q_{i'} q_j}{4\pi\varepsilon_0 r_{i'j}} \tag{6.5}$$

And it is apparent that the coupling with λ occurs by scaling the interaction energy terms instead of a combination of force field parameters. The two systems S_1 and S_2, while present in the simulation, should not experience any interaction with each other. This is done by ignoring any pair-pair energy terms involving them. Furthermore, in the implementation chosen here, the intramolecular non bonded energy of S_1 and S_2 is not coupled to λ. As a result, a fully decoupled system will experience the complete intramolecular energy terms and none of the intermolecular energy terms, which would be equivalent to having transferred the system to an ideal gas phase. Because in this simulation protocol, the two perturbed systems are present as distinct molecules the method is called dual topology.[31] The protocol is described in figure 6.2.

An advantage of the dual topology method is that it is not necessary to define

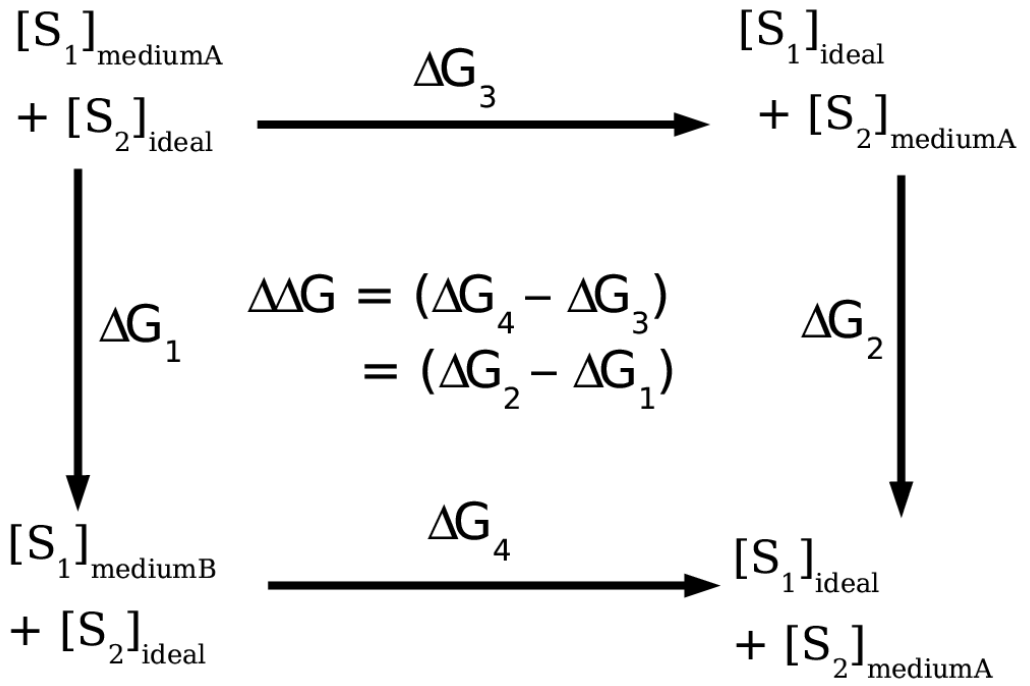


Figure 6.2: A thermodynamic cycle constructed with the dual topology method that relates the difference in free energy between S_1 and S_2 between two medias A and B. Note that if medium A or B corresponds to an ideal state ('vacuum like'), then ΔG_3 or ΔG_4 is 0.

dummy atoms or to devise means to perturb the geometry of S_1 into S_2. A disadvantage is that the number of particles to be simulated is increased, but that is rarely an issue as the number of particles defining the surrounding medium is often in vast excess of those part of the perturbed systems.

Because the two methods accomplish the perturbation of S_1 into S_2 differently, they will yield different single free energy differences. The double free energy difference, however, should be identical because the free energy is a state function. This has indeed been observed by computer simulations.[170]

6.3 Softening the intermolecular interactions

In the dual topology method, the functional form for the Lennard Jones energy term would be:

$$U(\lambda) = U_0(LJ) + \lambda U(S_2)_{LJ} + (1-\lambda)U(S_1)_{LJ} \tag{6.6}$$

and the LJ terms involving interactions with S_N and the medium are scaled by

λ or (1-λ) respectively. If λ is equal to 0 or 1, then one of the systems does not experience any interaction with the medium, meaning it can for example, overlap completely with some molecules belonging to the medium. On the other hand, at a value of λ = 0.001 or λ = 0.999, any such interaction will be reduced by a factor of one thousand. This could be deemed sufficient to smoothly decouple S_N from the surrounding medium, but recalling the functional form of the Lennard Jones interaction energy

$$U_{LJ} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{6.7}$$

because of the high exponent on the repulsive term, two non bonded systems that approach closely will experience a very high energy and should they fully overlap, the equation will diverge. This behaviour will not be altered by a simple linear scaling of the calculated energy.

Dual Topology simulations run with a standard Lennard Jones equation almost invariably yield divergent free energy profiles close to the end states, where one of the system must be turned off completely.[171,172]

This problem could in principle happen in single topology calculations with dummy atoms. This is however rarely experienced because it has become common practice to retract the non interacting dummy atoms inside the van der Waals radius of a nearby non dummy atom. As a result, the dummy atom are 'protected' from bad overlaps.[173]

One method to overcome the so called "Lennard Jones end point singularity" problem is to make use of a modified Lennard Jones equation.

Beutler et al.[171] and Zacharias et al.[172] proposed the following equation:

$$U_{LJ,soft,\lambda} = (1-\lambda)4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{(\lambda\alpha_{soft} + r_{ij}^2)^6}\right) - \left(\frac{\sigma_{ij}^{6}}{(\lambda\alpha_{soft} + r_{ij}^2)^3}\right)\right] \tag{6.8}$$

Equation 6.8 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

equation allows for the smooth annihilation of an atom i belonging to S_1. For the atoms i' of S_2, the parameter λ is simply substituted by (1-λ). Equation 6.8 depends on a parameter α_{soft} that control the degree of 'softness' of the potential. The actual value of the parameter is often adjusted to obtain a smooth free energy change.[174]

When the Lennard Jones interactions are softened such that atomic overlaps between non bonded particles becomes possible, it is necessary to use a modified functional form for the coulombic equations as well, otherwise, it might be feasible for two atoms of opposite charge to adopt exactly the same coordinates and experience an infinitely attractive coulombic energy.[17,175]

$$U_{coul} = \frac{(1-\lambda)q_i q_j}{4\pi\varepsilon_0\sqrt{(\lambda + r_{ij}^2)}} \tag{6.9}$$

The softening of the electrostatic interactions calculated by Ewald summation has been recently proposed.[176]

6.4 Solvation free energy calculations

The dual topology method and the separation-shifted scaling softcore described previously were implemented in a modified version of the program ProtoMS21.[147] Equations 6.8 and 6.9 were slightly modified.

$$U_{LJ,soft,\lambda} = (1-\lambda)4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}^{12}}{(\lambda\delta\sigma_{ij} + r_{ij}^2)^6}\right) - \left(\frac{\sigma_{ij}^{6}}{(\lambda\delta\sigma_{iJ} + r_{ij}^2)^3}\right)\right] \tag{6.10}$$

$$U_{coul} = \frac{(1-\lambda)^n q_i q_j}{4\pi\varepsilon_0\sqrt{(\lambda + r_{ij}^2)}} \tag{6.11}$$

The parameter α_{soft} was replaced by σ_{ij} in equation 6.10. This has the advantage that the softness does not have to be specified *a priori*. Furthermore, the use of σ_{ij} has the advantage that systems experiencing stronger Lennard Jones interactions will be automatically softened further. Should this not be sufficient, the

degree of softness can still be controlled through the parameter δ, which is set to 1 by default. The exponent n in equation 6.11 was introduced so that the rate of softening of the coulombic can be controlled as well.

6.4.1 Relative solvation free energy of ethane and methanol

To test the correctness of the implementation, the calculation of the relative solvation free energy of ethane and methanol was performed using single and dual topology methods.

Simulation protocol

Models of ethane and methanol were constructed using parameters from the GAFF force field[107] and atomic partial charges were derived using the AM1/BCC method.[108] For the single topology method, two hydrogen atoms on one methyl group of ethane were retracted with the coupling parameter λ and transformed into dummy atoms, while the bond length between the two carbon atoms was coupled with λ such that it matches the equilibrium bond length of two carbon atoms of ethane (λ = 0) and the carbon-oxygen bond of methanol (λ = 1). The solute(s) were placed at the center of a cubic box of dimension 25x25x25 Å which was filled with 533 TIP4P water molecules. A switching residue based cutoff of 10 Å was used and the intermolecular energies were feathered over the last 0.5 Å. The system was equilibrated for 50 million (M) moves at a temperature of 25 °C and a pressure of 1 atmosphere in the NPT ensemble. To keep the solute(s) centred in the box, translational motion of the solute was removed (but rotational motion was conserved). Solvent moves were attempted 99%, solute moves 0.9% and volume moves 0.1% of the time. Preferential sampling was employed with a constant of 200.0. The resulting configuration was distributed over 11 simulations of evenly spaced coupling parameter λ (0.00, 0.10 ,.., 0.90, 1.00). Each simulation was equilibrated for 5M moves and statistics were collected for 25 M subsequent moves. Free energy differences were calculated using the Replica Exchange Thermodynamic Integration method[59] and moves between neighbouring replicas were attempted every 200 K moves.

The single topology method required the perturbation of ethane into methanol in the gas phase. This was done by running 11 simulations of the evenly spaced coupling parameter λ for 500 K moves each.

In the dual topology simulations, unless otherwise noted, the rigid body rotations of the two solutes were coupled together. This should ensure that the two solutes will stay close to each other and is expected to help converge the free energies. The same principle is applied to the translation of the whole solutes, but there is no translational motion in this system.

The errors reported for each individual free energy are calculated in the following manner. For each calculation performed at a value of λ, the distribution of forwards free energy gradients (formed from the energy difference between the simulation performed at (λ + dλ) and λ) as calculated by ProtoMS for each block of the simulation are collected. A 95 % confidence interval is calculated for this distribution of forwards energy. The same procedure is applied for the backwards free energy ((λ - dλ) - λ). In a worst case scenario, the error estimate on the forwards free energy gradients is added to the forwards free energy gradients while the error estimate on the backwards free energy gradients is substracted. The same procedure is repeated by substracting and adding the error estimates to the forwards and backwards free energy gradients. This give an upper and lower bound to the value of the free energy gradients collected in this simulation. The actual free energy gradients are taken as the average of the forwards and the opposite of the backwards free energy gradients. This procedure is repeated for every value of λ. The three different free energy gradient profiles are then integrated to yield three different values of free energy change. The error on the estimated free energy is then taken as the average of the difference between the two free energies that included the maximum error estimates.

In principle, this procedure should yield a maximum upper bound on the statistical error associated with the free energy. However, this would be true only if all the regions of phase space that contribute significantly to the free energy difference have been visited with the correct probabilities.

Another more established method to calculate an error interval on the free energy difference is to calculate the free energy change for each block of simulation.

If during each block of simulation, the whole of phase space had been thoroughly sampled, then the same free energy difference would be obtained. In practice, they are likely to differ however. By plotting the distribution of the free energies calculated in each block of the simulation, a 95 % confidence interval can be obtained from this distribution. Prior to performing this analysis, the data from N blocks of K moves can be reduced to L blocks of (K*(N/L)) moves. This can be useful if the number of moves K was too short, in which case a correlation between two subsequent blocks would exist. This correlation would have an adverse effect on the calculation of the 95 % confidence interval, whose derivation is based on the assumption that the input data is not correlated.

Simulation results

In an effort to reliably assess the convergence of the calculated free energies, a free energy simulation was performed five times for several protocols. All the simulations used a different random number seed and yielded slightly different free energies. The errors for each simulation were estimated by calculating error bounds on the free energy gradients, or by performing a block average analysis of the distribution of free energies. These were calculated with an interval of 100K or 500K moves for each block. The results are presented in table 6.1. For comparison, the experimental relative solvation free energy difference between ethane and methanol is -6.90 kcal mol^{-1}.[115]

Table 6.1: Relative solvation free energy of ethane and methanol[a]

run	$\Delta\Delta G_{solv}$	Error$_{grads}$	Error$_{block500K}$	Error$_{block100K}$
Single Topology[b]				
1	-5.95	0.36	0.18	0.10
2	-5.79	0.37	0.20	0.11
3	-5.99	0.37	0.19	0.11
4	-6.19	0.36	0.15	0.10
5	-5.82	0.35	0.17	0.11
Average	-5.95 $\pm$ 0.15			
Dual Topology, no softcore				
1	-11.12	1.55	1.86	1.29
Dual Topology, softcore, δ = 0.25, n=1				
1	-7.63	0.61	0.57	0.35
Dual Topology, softcore, δ = 1, n = 1				
1	-6.20	0.42	0.21	0.13
2	-5.98	0.41	0.26	0.14
3	-6.34	0.43	0.20	0.13
4	-5.97	0.42	0.19	0.12
5	-6.02	0.42	0.17	0.12
Average	-6.10 $\pm$ 0.15			
Dual Topology, softcore, δ = 1, n = 1, do not sync rotations				
1	-6.30	0.36	0.24	0.14
2	-5.90	0.36	0.29	0.16
3	-5.87	0.34	0.23	0.13
4	-6.09	0.36	0.20	0.13
5	-6.01	0.38	0.19	0.12
Average	-6.03 $\pm$ 0.16			
Dual Topology, softcore, δ = 1, n = 0				
1	-5.97	0.39	0.18	0.13
2	-6.17	0.39	0.25	0.13
3	-5.92	0.38	0.21	0.13
4	-6.06	0.39	0.24	0.14
5	-5.87	0.39	0.23	0.14
Average	-6.00 $\pm$ 0.11			

a For the average of 5 runs, the error estimate is the 95 % confidence limit of the mean, obtained from the independent simulations. All the figures are in kcal mol^{-1}

b The free energy change for the perturbation in vacuum has been substracted. This quantity was 2.69 $\pm$ 0.01 kcal mol^{-1}

Table 6.2: Relative solvation free energy of ethane and methanol (continued) [a]

run	$\Delta\Delta G_{solv}$	Error_{grads}	$\text{Error}_{block500K}$	$\text{Error}_{block100K}$
Dual Topology, softcore, δ =1, n =2				
1	-6.58	0.45	0.21	0.14
2	-6.15	0.44	0.24	0.14
3	-6.05	0.45	0.21	0.13
4	-6.25	0.43	0.20	0.13
5	-5.91	0.44	0.26	0.15
Average	-6.19 $\pm$ 0.24			
Dual Topology, softcore, δ =2, n =1				
1	-6.17	0.69	0.41	0.24
2	-6.17	0.69	0.43	0.24
3	-6.68	0.70	0.40	0.22
4	-5.85	0.68	0.33	0.21
5	-5.61	0.66	0.41	0.22
Average	-6.10 $\pm$ 0.38			

The single topology simulation results are within 1 kcal mol^{-1} of the experimental figure. The error intervals obtained from the different methods under or overestimate the confidence limit obtained for 5 independent simulations. The block averaging methods do not agree with each other. With a block size of 100K moves, every particle in the box has been moved on average 187 times, but since preferential sampling was enabled, the water molecules closest to the solute, which will cause most of the free energy change, have been moved more often than this. There is no easy way to decide if a block was long enough such that each value in the distribution is uncorrelated. Nevertheless, a block size of 100K (or 500K) would be higher than those used in most free energy studies. Thus the methodology exhibits sensitivity to the block lengths, which make an estimation of the reliability of the results problematic.

The dual topology simulations fails to give a converged answer if the solute-solvent interaction energies are not softened. This is due to the very large fluctuations in the free energy gradient at the end of the λ coordinate, shown in figure 6.3. The fluctuations are at least two orders of magnitude higher than for other values of the coupling parameter. This is because at λ = 1.0, the ethane molecule should be

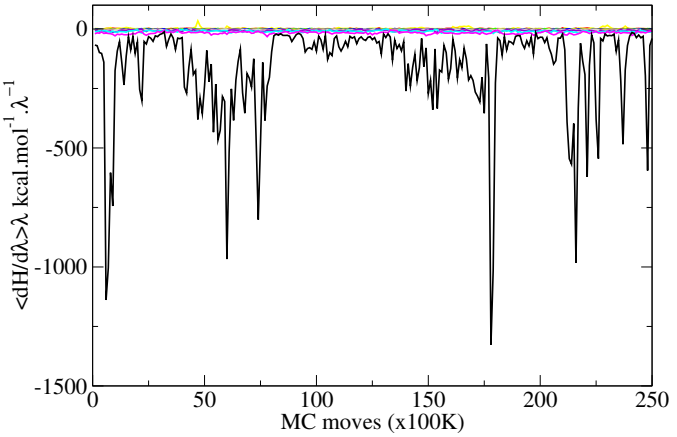


Figure 6.3: The fluctuations in the free energy gradient during the simulation carried out at $\lambda = 1.0$ with the dual topology method and no softening of the solute-solvent intermolecular interactions. For comparison, the gradients obtained at all other values of λ are also shown.

completely decoupled from the solvent. As it is bigger than the methanol molecule, it can overlap partially with neigbouring water molecules. When the energy of the system is calculated at $\lambda = 0.999$ to form the free energy gradients, a strongly repulsive Lennards Jones energy is obtained. Thus, occasionaly some configurations will exhibit very large gradients. If rare configurations contribute significantly to the free energy gradients, a very long simulation time will be necessary to obtain converged results. Note that the errors calculated by the block averaging methods are higher than those obtained from the free energy gradients method. However neither method provides error bounds that would overlap with the single topology results.

By contrast, most dual topology simulation with a softcore enabled yield results that agree to within error estimates with the single topology results. If the softcore parameter δ is too small, the Lennard Jones interactions are still too hard close to $\lambda = 1.0$ and most of the change in free energy gradients occurs there, with large variations, making the calculation imprecise. In ProtoMS2.1, a Monte Carlo move of the solute consists of the translation and rotation by a random amount around the molecule axis, followed the random perturbation of the internal degrees of freedom of the molecule. As mentioned in the previous section, the dual topology implementation permits the translation and/or rotation of the two molecules whose relative free energy change is of interest to be synchronised. Because in the end states, one of the two solute molecules has been completely decoupled from its

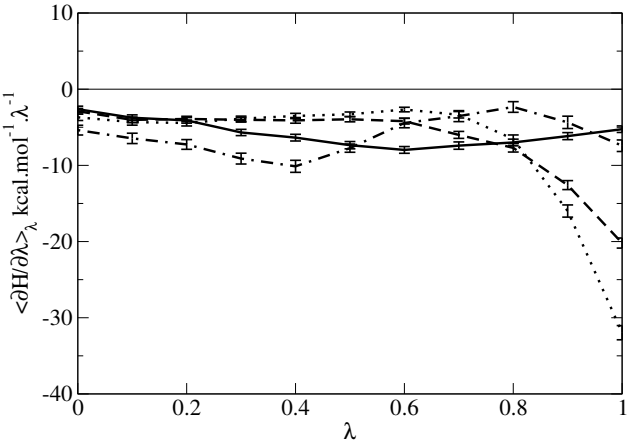


Figure 6.4: The free energy gradients for the perturbation carried out with various softcore parameter sets. For the solid line (n=0, δ=1.0), for the dashed line (n=1, δ=1.0), for the dotted line (n=2, δ=1.0), for the dashed-dotted line (n=1, δ=2.0).

environment, a rigid rotation of the molecule does not affect the potential energy of the system. Therefore, the free energy change between the two end states should be identical, whether or not the solute rigid body rotations were synchronised. This appear to be verified in this system, where the dual topology simulations with no synchronisation of the solute rotations yield identical answers (within error estimates) to the simulations performed without synchronisation. However, over a sufficiently long simulation time scale, the solvent environment should be insensitive to solute rotations, so it could be argued that the validity of the constraint has not been demonstrated.

Varying the softcore parameters δ or n modifies the value of the free energy gradients recorded at different values of λ. However, the final free energy difference is not affected and integration of the three different profiles yield the same free energy change (within statistical error). This is of course a consequence of the fact that the free energy is a state function. By setting n to 0, the solute-solvent coulombic energy term is linearly scaled (and shifted by λ) during the simulation. This means the coulombic energy of methanol is restored more evenly throughout the simulation, while with a parameter n of 1 or 2, it would appear more abruptly at the end of the simulation. This should cause the free energy gradients to vary more rapidly towards the end of the simulation. This is observed in the plot of the free energy gradients for different values of the softcore parameters in figure 6.4. When δ is increased from 1.0 to 2.0, the free energy gradients profile exhibits a

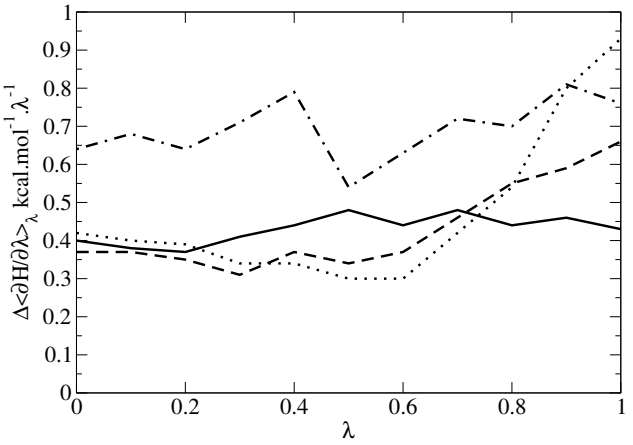


Figure 6.5: The error on the free energy gradients for the perturbation carried out with various softcore parameter sets . For the solid line (n=0,δ=1.0), for the dashed line (n=0,δ=1.0), for the dotted line (n=2,δ=1.0), for the dashed-dotted line (n=1,δ=2.0).

more complex behavior, with two inflexion points at $\lambda = 0.40$ and $\lambda = 0.80$. This more complex free energy gradient profile may be more inaccurately integrated by numerical integration schemes than the other plots.

In figure 6.5, the calculated error (obtained with the error calculation method based on the distribution of the free energy gradients) in the free energy gradients along λ for the three different protocols is plotted. The plot illustrate that when the rate of change of the free energy gradients increase, so does the calculated error interval. The simulations carried out with the softcore parameter $n = 0$ and $\delta = 1.0$ has the smoothest profile. The error on the free energy gradients is also similar across λ, while the decrease in free energy gradients observed at the end of the simulation correlates with an increase of the error interval.

Increasing the softcore parameter δ to 2.0 causes the error obtained by gradients and block analysis to increase. In figure 6.5 the fluctuations in the free energy gradients for a simulation conducted with parameters ($\delta = 1.0$, n=1) and ($\delta = 2.0$, n=1) can be compared. It is seen that the error on the free energy gradients is systematically higher across the entire range of the coupling parameter λ.

In addition, a plot of the replicas exchanged at different values of the coupling parameter λ during simulations performed with different softcore parameter sets are shown in figure 6.6. When the simulation is performed with the parameter set (n=0, δ=1.0) the replicas exchange freely across λ. When n is increased to 1 or

2, the replicas at λ 0.8-1.0 do not exchange as well with the rest of the system and tend rather to exchange between themselves. This suggest that the equilibrium distributions in the range of the coupling parameter λ 0.8-1.0 differ more from those at lower values of λ. This is because the coulombic energy of methanol is restored more abruptly at the end of the simulation when n is set to 1 or 2, compared with n set to 0. Finally, the RETI plot is more sparse across all values of λ when δ is set to 2.0, meaning that the acceptance rate for the exchange of two replicas is lower on average than if the simulation was conducted with δ set to 1.0. This is likely to be observed because, as δ increases, the solutes are made softer, which allows the surrounding solvent molecules to occupy a larger portion of the volume of space that would be occupied by the hard molecule. For reference, the RETI plot obtained with the single topology method is also shown in figure 6.6. All the replicas can exchange well and visit a wide range of λ values.

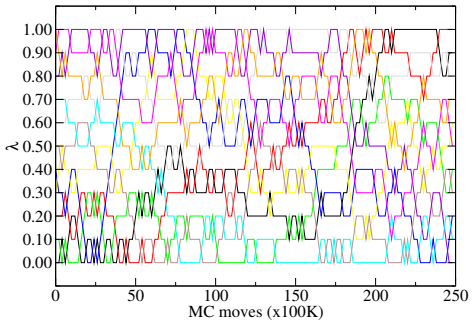
From all the observations reported in this section, the following observations can be drawn.

1. The shape of the free energy profile and the ease by which a free energy difference is calculated can be dramatically controlled through the parameter n and δ.

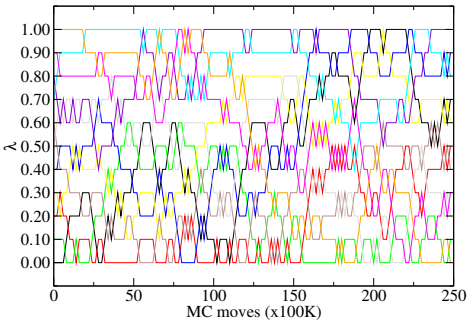
2. If a non-polar molecule is perturbed into a polar molecule, it is better to spread the increase in coulombic energy across the whole coupling parameter, rather than restore it abruptly at the end of the simulation. This can be done by setting n to a small integer.

3. If the value of δ is too small, the soft-core will be too hard and behaviour similar to a dual topology simulation without softcore will be observed (e.g, large fluctuation in the free energy gradients at the end states). If the value of δ is too large, the softcore will be too soft and this will make the free energy calculation harder as the volume of the two solute molecules will vary more.

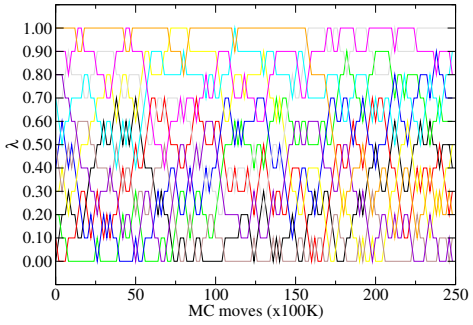
4. The error analysis method based on the free energy gradients tends to overestimate the true uncertainty in the simulation outcome. The error analysis method based on block averaging is sensitive to the number of blocks used



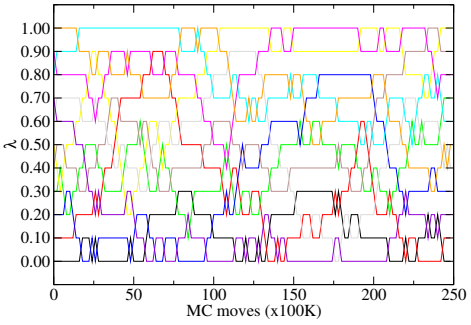
(a) n=0, δ=1.0



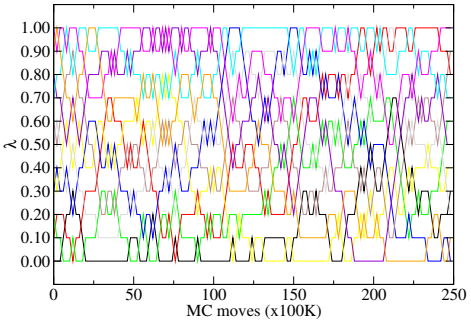
(b) n=1, δ=1.0



(c) n=2, δ=1.0



(d) n=1, δ=2.0



(e) single topology

Figure 6.6: The exchange of replicas during the free energy simulation for different values of the softcore parameter set.

(often chosen arbitrarily) and tends to underestimate the true uncertainty in the simulation outcome.

Finally, from this set of results the author formed an opinion of the conditions an ideal free energy gradient profile should fill:

1. The free energy gradients profile should be smooth and flat over the coupling

parameter λ. The free energy difference is obtained from integration of this profile, and a smoother profile can be integrated by a trapezium or Simpson rule accurately with less points than a complex, rapidly changing profile.

2. The errors on each individual free energy gradient should be as small as possible. This will depend on how rapidly the total energy of the system change with a small increment (or decrement) of λ by $d\lambda$.

3. The equilibrium density of state of the systems run at different values of the coupling parameter λ should be as similar as possible. This will allow the RETI method to exchange replicas at different values of the coupling parameter λ easily. If that is not the case, the benefits of the RETI method will be lessened. Inevitably, there are differences in the density of states at λ = 0.0 and λ = 1.0. But it is better to ”spread” this change evenly at each window than to have all the difference occur at a single value of λ.

There should be some degree of equivalence between all these conditions. For example, a smooth free energy profile is likely to mean that replicas can exchange readily. The observations reported in this section should prove useful to investigate optimum parameter sets for the softcore across different systems.

6.4.2 Relative solvation free energy of benzene, ethane and methanol

In addition to the calculation of the relative solvation free energy of ethane and methanol, the relative solvation free energy of benzene and methanol was calculated with the dual topology approach. Such a perturbation would be difficult (but not impossible) to set up for a single topology method as the geometry of the two solute molecules is quite different (planar ring and linear molecule). By contrast, the setup of such calculation with the dual topology method is no more complex than the setup of the perturbation of ethane to methanol. Because the method requires less user interaction, it could be more readily automated, which would be a requirement to make high throughput free energy simulations practical. The relative solvation free energy of benzene to ethane was also calculated by the same approach. This allows us to close a thermodynamic cycle involving benzene, ethane

and methanol and thus permit the degree of convergence of the simulation results to be assessed.

A protocol identical to the previous section was employed for these calculations with the exception that the system was solvated with 526 TIP4P waters, instead of the 533 TIP4P water used in the previous section.

The simulation results are summarised in table 6.3. For reference, the experimental relative solvation free energy of benzene to ethane is +2.9 kcal mol^{-1} and the experimental relative solvation free energy of benzene to methanol is -4.0 kcal mol^{-1}.[115]

Table 6.3: Relative solvation free energy of benzene, ethane and methanol[a]

run	$\Delta\Delta G_{solv}$	Error_{grads}	$\text{Error}_{block500K}$	$\text{Error}_{block100K}$
Benzene to methanol				
Dual Topology, softcore, δ = 1, n = 1				
1	-4.08	0.73	0.37	0.23
2	-3.28	0.68	0.31	0.20
3	-3.55	0.74	0.52	0.28
4	-3.64	0.72	0.41	0.24
5	-3.87	0.71	0.36	0.21
Average	-3.68 $\pm$ 0.29			
Dual Topology, softcore, δ = 1, n = 1, do not sync rotations				
1	-3.44	0.65	0.35	0.20
2	-3.62	0.67	0.37	0.21
3	-3.38	0.65	0.32	0.20
4	-3.86	0.64	0.38	0.23
5	-3.77	0.63	0.26	0.17
Average	-3.61 $\pm$ 0.20			
Dual Topology, softcore, δ = 2, n = 1				
1	-4.38	0.80	0.38	0.24
Benzene to ethane				
Dual Topology, softcore, δ =1, n =0				
1	+2.93	0.51	0.30	0.17
2	+2.88	0.47	0.25	0.15
3	+3.32	0.47	0.26	0.15
4	+2.87	0.50	0.29	0.16
5	+2.96	0.48	0.20	0.14
Average	+2.99 $\pm$ 0.18			

a For the average of 5 runs, the error estimate is the 95 % confidence limit of the mean, obtained from the independent simulations. All the figures are in kcal mol^{-1}

In the perturbation of benzene to methanol, simulations carried out with or without synchronisation of the solute rigid body rotations give the same free energy difference. In the later case, it is interesting that the different error analysis methods and the confidence interval on the distribution of free energies obtained from independent simulations are systematically smaller. Intuitively, the opposite results would have been expected as it was thought that synchronisation of the rigid

body rotations of the two solutes should lead to better convergence of the calculated free energy changes. Also, increasing δ to 2.0 increase the error measure on a single simulation, but to a lesser extent than in the previous system. The most precise predicted free energy change is -3.61 $\pm$ 0.20 kcal mol^{-1}, which slightly underestimates the experimental figure.

The perturbation of benzene to ethane is easier to conduct and yields a free energy change of +2.99 $\pm$ 0.18 kcal mol^{-1}, which is in excellent agreement with the experimental figure.

Taking the free energy change of ethane to methanol from the previous section as -6.03 $\pm$ 0.16 kcal mol^{-1}, the closure of the thermodynamic cyle is 0.57 kcal mol^{-1}. This quantity appears reasonable, given the error bounds on each individual simulation. Since the predicted relative solvation free energy of benzene to ethane is in good agreement with the experimental figure, and the predicted relative solvation free energy of benzene to methanol and ethane to methanol underestimate the experimental figure by 0.3 and 0.9 kcal mol^{-1} respectively, this suggests that the main origin of the discrepancy with the experiment is in the model of methanol. Since the GAFF Lennard Jones parameters for methanol are typical of existing biomolecular force fields, and that other force fields such as OPLS achieve very good agreement with the experimental relative solvation free energy of ethane to methanol,[177] the source of the discrepancy is likely to be found in the atomic partial charges generated by the AM1/BCC method.[108]

6.5 Binding free energy calculations

6.5.1 Relative binding free energy of celecoxib analogues

System setup

The methodology employed in the previous section was applied to the relative binding free energy calculation of the two analogues of celecoxib **8** and **1**, bound to cyclooxygenase-2 (see figure 6.7). This system was selected because it was observed to yield relatively precise binding free energies with the single topology method, by implicit or explicit solvent approaches, and was extensively studied

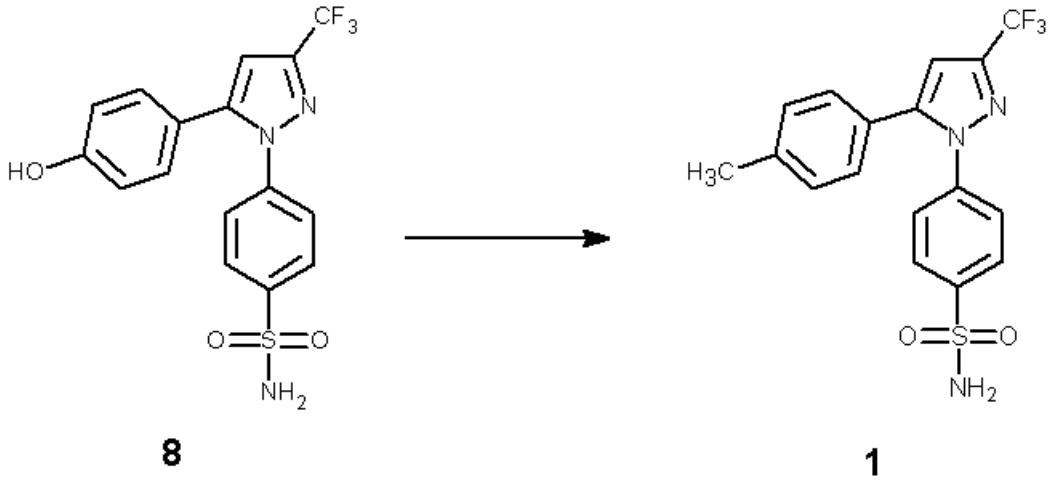


Figure 6.7: The hydroxyl analogue **8** is perturbed into celecobix (**1**) .

in chapter 3 and 4. The single topology simulations were performed with a protocol similar to that reported in chapter 4. In the dual topology simulations, the rigid body rotation and translation of the two solutes were coupled together. This is because, unlike in the relative solvation free energy calculations reported in the previous section, the solute translational motion must be enabled to yield correct sampling of the binding site, and without such coupling, one of the two solutes could float freely off the binding site at the end states, which would lead to divergence of the free energy gradients.[178] The protein-ligand complex or the ligand solvated in explicit water was equilibrated at a value of $\lambda = 0.50$ for 30M moves. Each simulation carried out at a value of λ was equilibrated for 10M moves and data was collected for 60M moves. This is twice the number of moves performed with the single topology method but was judged necessary because the free energy gradients were seen to fluctuate more readily (see next subsection).

Explicit solvent simulations

Because the computational expense for each binding free energy calculation was higher than the solvation free energy calculations, each simulation protocol was repeated only three times. The results are presented in table 6.4.

Table 6.4: Relative binding free energy of **8** and **1**[a]

run	$\Delta\Delta G_{solv}$	Error$_{grads}$	Error$_{block1M}$	Error$_{block500K}$
Single Topology, unbound state				
1	18.04	0.36	0.22	0.20
2	18.56	0.37	0.20	0.18
3	18.59	0.36	0.19	0.16
Average	18.40 ± 0.51			
Single Topology, bound state				
1	14.99	0.15	0.10	0.08
2	15.24	0.16	0.09	0.09
3	14.95	0.16	0.09	0.08
Average	15.06 ± 0.26			
Dual Topology, softcore, δ = 1.5, n = 0, unbound state				
1	5.57	1.18	0.69	0.59
2	5.44	1.19	0.82	0.65
3	4.40	1.21	0.86	0.67
Average	5.14 ± 1.08			
Dual Topology, softcore, δ = 1.5, n = 0, bound state				
1	2.92	0.82	0.57	0.45
2	1.27	0.82	0.43	0.36
3	3.02	0.85	0.50	0.41
Average	2.40 ± 1.65			

a For the average of 3 runs, the error estimate is the 95 % confidence limit of the mean, obtained from the independent simulations. All the figures are in kcal mol^{-1}

The single topology results are well behaved, in the bound and unbound state. A relative binding free energy of -3.34 $\pm$ 0.57 kcal mol^{-1} can be estimated from the results. This is similar to the value that was obtained previously in chapter 4. By contrast, the dual topology simulation results are disappointing. The error estimates are larger, and the spread of the simulation results make the calculation imprecise. This is even though the simulation length of each individual window was twice as long as for the single topology simulations. From the 3 independent simulations, a relative binding free energy of -2.74 $\pm$ 1.97 kcal mol^{-1} is estimated. Thus the two methods appear to give the same answer, although the imprecision on the dual topology method is such that definitive conclusions are difficult to draw.

Figure 6.8 shows the free energy gradients collected for this perturbation with both methodologies. It can be seen that the free energy gradient profile obtained with the single topology method is smooth and reproduced to within statistical error by independent runs, except perhaps at $\lambda = 1.0$. By contrast, in the dual topology method the free energy gradients vary sharply, particularly for the perturbation in the unbound state between λ 0.4 and λ 0.6. Even though this represents a fairly large change in gradients, its impact on the free energy change is small. This is because the area under the curve between λ 0.4 to 0.5 is similar to the area between λ 0.5 and 0.6. The former contributes negatively to the free energy change while the later contritubes positively. Of bigger concern is that the statistical error for individual free energy gradients are 5 to 10 times larger (this is not evident on figure 6.8 because of the scale of the plots) and the free energy gradient profile not as reproducible. This is even though twice as many Monte Carlo moves were employed in the dual topology simulations.

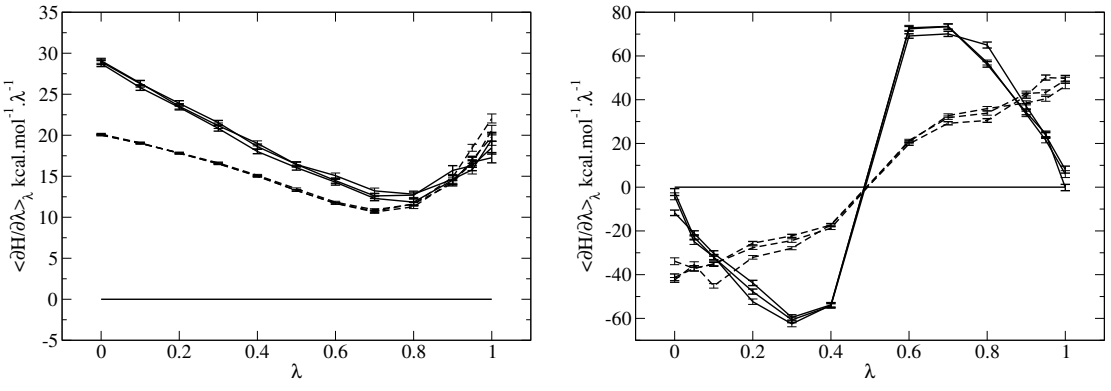
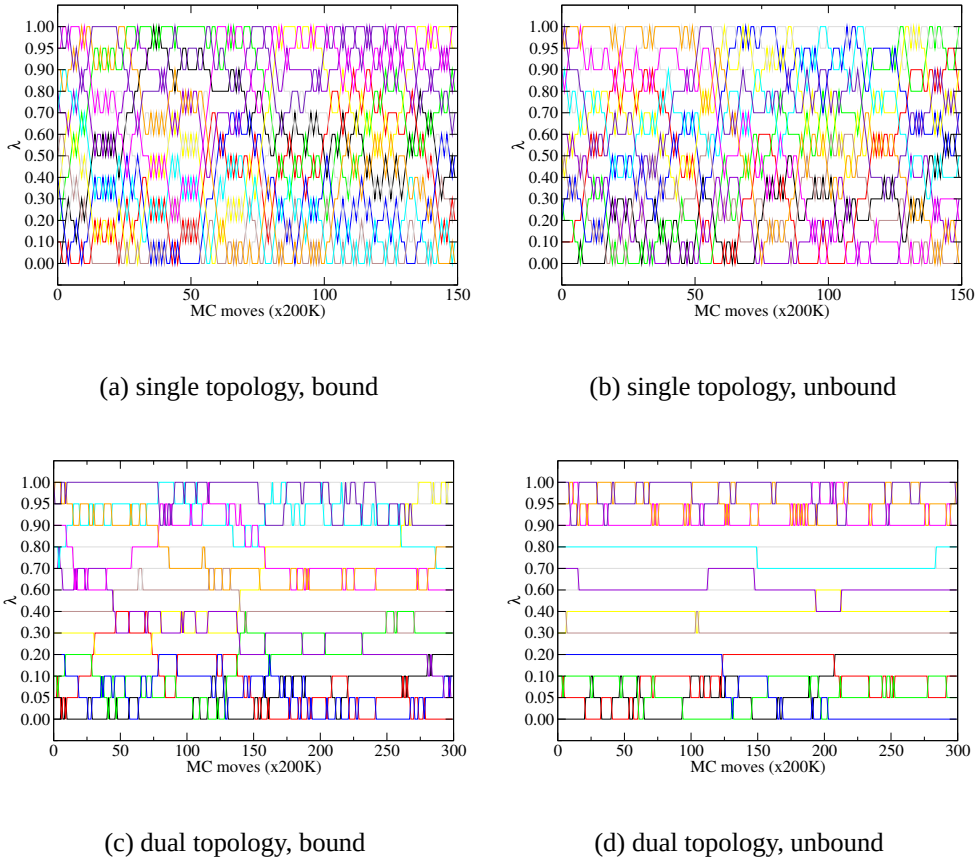


Figure 6.8: Free energy gradients collected during the perturbation of **8** into **1** in the bound (solid line) and unbound (dashed line) state. Three independent simulations are plotted and the error bars are shown.
(a) Single topology (b) Dual topology, $\delta = 1.5$, $n = 0$

Figure 6.9 highlights that the perturbation of **8** into **1** is considerably more difficult with the dual topology approach. With the single topology method, replicas can exchange well across the whole coupling parameter because the perturbed states are relatively similar. This is no longer the case in the dual topology method and the rate of exchange of replicas drops dramatically and is such that it is very

difficult to exchange replicas across the middle of the coupling parameter.



(a) single topology, bound

(b) single topology, unbound

(c) dual topology, bound

(d) dual topology, unbound

Figure 6.9: The exchange of replicas during the perturbation of compound **8** into **1** in the bound and unbound state with the single and dual topology methodology.

These results constrast sharply with those reported in the section on relative solvation free energy calculations of ethane and methanol, where a comparable rate of exchange of replicas was achieved with the dual and single topology methods, provided the softcore parameters were optimised for the system.

It was hypothethised that much of the difficulty in the dual topology simulations arise from the fact that the intermolecular coulombic and Lennard-Jones energy have to be decoupled simultaneously. In an effort to verify this assumption, a more complex thermodynamic cycle was devised. As can be seen in figure 6.10, the cycle involves the electrostatic discharging of **8**, followed by Lennard Jones decoupling of **8** while the interactions of **1** with the system are turned on simulatenously. The last step consists of charging **1**. When this series of simulations is run in the bound and unbound state, the difference of the sum of the free energy along each pathway will yield the relative binding free energy of **8** and **1**. Because the

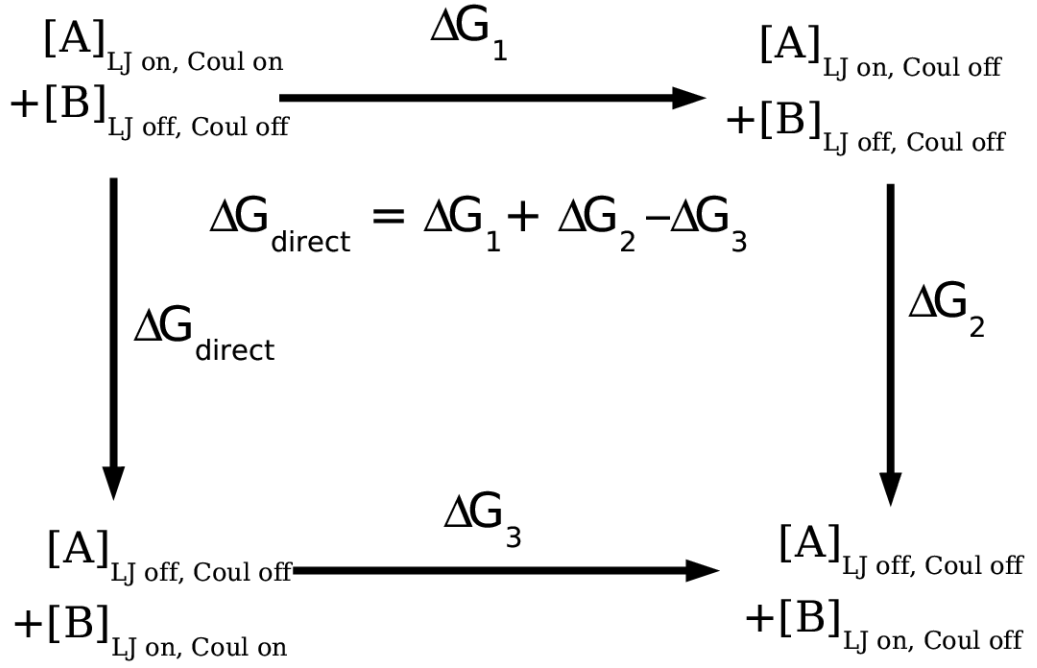


Figure 6.10: A thermodynamic cycle that breaks down the perturbation of ligand A into ligand B into three different steps. The term LJ and Coul refers to the ligands intermolecular Lennard Jones and the ligand intermolecular and intramolecular Coulombic energy respectively. To obtain a binding free energy, the cycle must be applied twice, in aqueous environment and in the solvated protein.

charging/discharging steps involve only one ligand, they were performed with the single topology method, while the Lennard Jones decoupling was performed with the dual topology method. All the simulations were equilibrated for 10M moves and data was collected for 30M moves. The results are presented in table 6.5.

Table 6.5: Relative binding free energy of **8** and **1**[a] by the three step cycle.

run	$\Delta\Delta G_{solv}$	Error$_{grads}$	Error$_{block1M}$	Error$_{block500K}$
Discharging **8**, unbound state				
1	165.15	0.61	0.54	0.41
2	165.51	0.61	0.37	0.29
3	165.15	0.63	0.40	0.31
Average	165.27 $\pm$ 0.35			
Discharging **8**, bound state				
1	162.65	0.39	0.24	0.19
2	162.44	0.37	0.26	0.20
3	162.25	0.36	0.22	0.19
Average	162.45 $\pm$ 0.34			
Exchange of **8** and **1** δ = 1.0, unbound state				
1	1.95	1.25	1.08	0.85
2	0.63	1.03	0.64	0.53
3	0.93	1.06	0.89	0.72
Average	1.17 $\pm$ 1.16			
Exchange of **8** and **1** δ = 1.0 , bound state				
1	2.00	1.33	0.90	0.78
2	1.07	1.31	1.00	0.80
3	2.08	1.22	0.96	0.74
Average	1.72 $\pm$ 0.94			
Discharging of **1**, unbound state				
1	147.26	0.58	0.34	0.27
2	147.29	0.58	0.44	0.34
3	147.95	0.60	0.36	0.28
Average	147.50 $\pm$ 0.66			
Discharging of **1**, bound state				
1	147.71	0.36	0.23	0.19
2	147.83	0.38	0.23	0.18
3	148.06	0.37	0.21	0.17
Average	147.87 $\pm$ 0.30			

a For the average of 3 runs, the error estimate is the 95 % confidence limit of the mean, obtained from the independent simulations. All the figures are in kcal mol^{-1}

From these results, a binding free energy difference of -2.64 $\pm$ 1.73 kcal mol^{-1} can be estimated. This once again agrees with the previous dual topology and sin-

gle topology results, but the results are still very imprecise. The steps involving discharging of either ligands yield very large free energy changes. This is because the intramolecular coulombic energy of the ligand is turned off and the magnitude of this term is larger than the intermolecular coulombic energy of the ligand. Despite the large free energy change, the results are fairly precise. Smaller free energy changes could be obtained if only the intermolecular coulombic energy was turned off. This would have the advantage that at the beginning or end of the cycle, one ligand would be in the ideal state. This is however, not possible with the single topology method in its current implementation in ProtoMS21. In addition, it is the exchange of the two uncharged ligands with the dual topology method that yield the most imprecise results. The free energy gradients plot for this perturbation, seen in figure 6.11(a) is not very different from the free energy gradient plot obtained with a direct perturbation of **8** into **1**. This shows that most of the difficulty in the calculations arise from the intermolecular Lennard Jones energy. This more elaborate thermodynamic cycle yields additional insights into the binding of compounds **8** and **1**. From the differences in free energy along each step of the pathway, it can be seen that the largest contribution to the binding free energy comes from the less favourable electrostatic environment in the protein for **8**. This is not unexpected as the hydroxy group of **8** cannot donate or receive a hydrogen bond with the position it adopts in the binding site of COX-2. Compound **1** may also experience stronger Lennard Jones interactions in the binding site of the protein, but the benefits are small, and uncertain given the error estimates.

With the following observations in mind, the direct simulation of **8** into **1** was repeated with the parameter δ adjusted to 1.25.

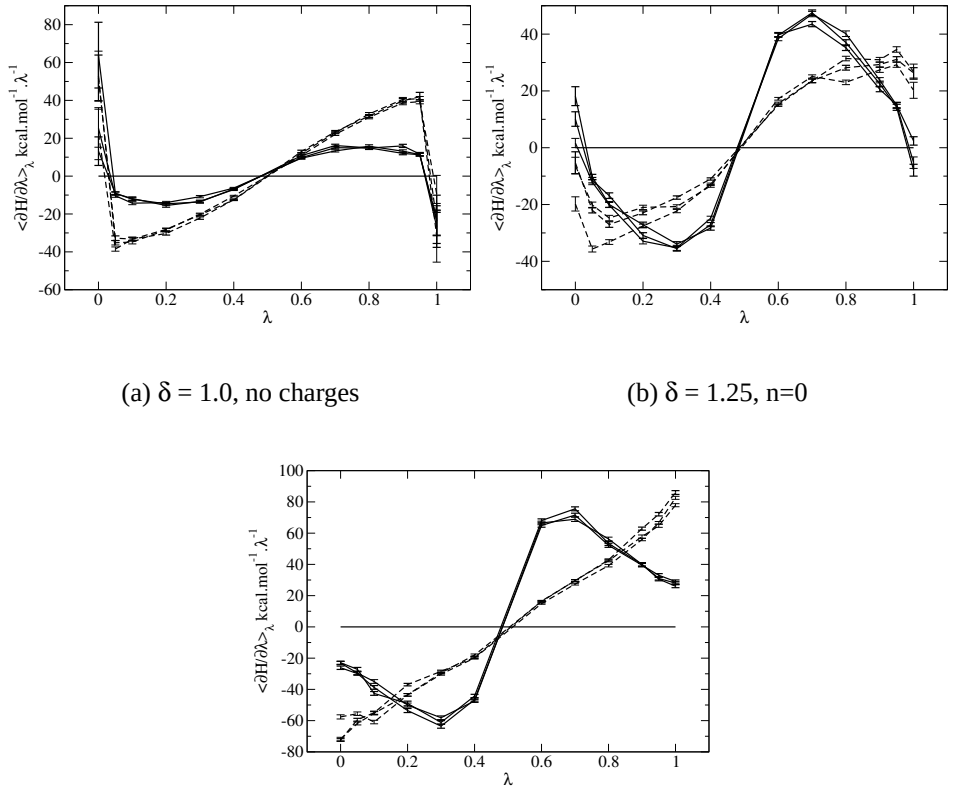
Table 6.6: Relative binding free energy of **8** and **1**[a]

run	$\Delta\Delta G_{solv}$	Error_{grads}	$\text{Error}_{block1M}$	$\text{Error}_{block500K}$
Dual Topology, δ = 1.25, n = 0, unbound state				
1	4.69	0.97	0.60	0.48
2	3.91	0.93	0.54	0.48
3	4.26	0.92	0.57	0.48
Average	4.29 ± 0.66			
Dual Topology, δ = 1.25, n = 0, bound state				
1	2.87	0.77	0.42	0.36
2	2.20	0.79	0.55	0.42
3	0.87	0.76	0.50	0.38
Average	1.98 ± 1.71			
Dual Topology, δ = 1.5, n = 0, no torsions, unbound state				
1	5.51	1.18	0.71	0.59
2	3.71	1.16	0.65	0.55
3	5.14	1.16	0.84	0.65
Average	4.79 ± 1.59			
Dual Topology, δ = 1.5, n = 0, no torsions, bound state				
1	1.80	0.73	0.48	0.38
2	-0.68	0.74	0.38	0.31
3	0.13	0.75	0.46	0.38
Average	0.42 ± 2.12			

a For the average of 3 runs, the error estimate is the 95 % confidence limit of the mean, obtained from the independent simulations. All the figures are in kcal mol^{-1}

The simulations performed with the softcore parameter δ set to 1.25 would yield a binding free energy difference of -2.31 $\pm$ 1.83 kcal mol^{-1}. As seen in figure 6.11(b), the free energy gradient profile is very different from that one recorded with δ set to 1.0 or 1.5 and shows how this parameter strongly affect the free energy gradients. It was thought that the ensemble of configurations the ligands would adopt when they are interacting with a protein/aqueous environment could be very different from those it would adopt in an ideal gas state. If this is indeed the case, one could expect to have difficulties in obtaining converged free energy gradients, particularly towards the end states. To examine this hypothesis, the simulations were repeated with no sampling allowed on the torsional degrees of freedom of the ligands. As a result, the two ligands are much more rigid (with only rigid body

rotation/translation and bond angles being sampled) and cannot adopt configurations different from those observed in the binding site. Surprisingly, this does not make the perturbations any easier, with the simulation results still very inconsistent. The free energy gradients profile, shown in figure 6.11(c) is similar to that obtained when torsional flexibility is allowed (figure 6.8).



(a) δ = 1.0, no charges

(b) δ = 1.25, n=0

(c) δ = 1.5, n=0, no torsions

Figure 6.11: Free energy gradients collected during the perturbation of **8** into **1** in the bound (solid line) and unbound (dashed line) state with three different protocols for the dual topology simulations. Three independent simulations are plotted and the error bars are shown.

The observations presented in this section lead us to the following conclusions:

1. The dual topology methodology performs poorly in comparison to the single topology method for the calculation of the relative binding free energy of similar ligands. This behaviour was not observed when perturbing small molecules to calculate their relative solvation free energy. The ligands are much larger than these small molecules and exhibit several degrees of freedom. In the single topology simulations, the perturbed states overlap very

well with the reference states because most of the internal degrees of freedom of the ligand are sampled synchronously. By contrast, in the dual topology simulations, the two molecules, representatives of the end states, have their internal degrees of freedom sampled independently. This reduces the overlap of the equilibrium density of states of the two molecules and increases the number of Monte Carlo moves required to obtain good convergence of thermodynamic properties.

2. This problem is also magnified by the softening of the intermolecular interactions of the whole solute. In particular, because the Lennard Jones term is softened, the accessible volume of the ligands varies significantly throughout the simulation and additional sampling is required to cover these regions.

Implicit solvent simulations

In the previous section, it was seen that the perturbations in the unbound state were often as difficult or more difficult to converge reliably free energy differences. The implicit solvent approach, successfully applied in the previous chapters of this thesis, avoids these difficulties. It is thus of interest to test the dual topology method in combination with a Generalised Born Surface Area force field. The current implementation of dual topology in ProtoMS21 does not support surface area calculations for the dual topology solutes. Because we are mainly interested in comparing single and dual topology results on this system, rather than relating the calculated quantities to experimental observables, surface area calculations were not included in the implicit solvent simulations (for either single or dual topology).

The protocol employed to carry out these simulations was similar to that reported in chapters 3 and 4, except that, following pre-equilibration at $\lambda = 0.50$, each simulation performed at one value of the coupling parameter λ was further-equilibrated for 300 K moves and data was collected for 1.8 M moves.

The simulation results are summarised in table 6.7.

Table 6.7: Relative binding free energy of **8** and **1**[a]

run	$\Delta\Delta G_{solv}$	Error$_{grads}$	Error$_{blockA}$	Error$_{blockB}$
Single Topology , unbound state				
1	19.59	0.09	0.03	0.02
2	19.63	0.08	0.03	0.03
3	19.63	0.09	0.02	0.03
Average	19.62 ± 0.04			
Single Topology, bound state				
1	17.44	0.17	0.15	0.12
2	17.59	0.17	0.09	0.07
3	17.64	0.17	0.18	0.12
Average	17.56 ± 0.17			
Dual Topology, unbound state				
1	6.01	0.14	0.07	0.05
2	6.00	0.13	0.05	0.04
3	6.02	0.12	0.08	0.04
Average	6.01 ± 0.02			
Dual Topology, δ = 1.25, n = 0, bound state				
1	3.86	0.87	0.74	0.56
2	3.65	0.82	0.76	0.52
3	3.30	0.85	0.66	0.49
Average	3.60 ± 0.48			
Dual Topology, δ = 1.25, n = 0, bound state, rigid protein				
1	4.07	0.76	1.04	0.61
2	3.58	0.81	1.06	0.63
3	3.04	0.83	0.95	0.63
Average	3.56 ± 0.87			

a For the average of 3 runs, the error estimate is the 95 % confidence limit of the mean, obtained from the independent simulations. All the figures are in kcal mol^{-1}

The single topology are very well behaved and precise. The perturbation of **8** into **1** in the unbound state was the major source of imprecision in the explicit solvent simulations. Understandably, no such difficulties are seen when the aqueous environment is represented by an implicit medium. The relative binding free energy is found to be -2.06 $\pm$ 0.18 kcal mol^{-1}.

The dual topology simulation in the unbound state yields a very precise free energy change of 6.01 $\pm$ 0.02 kcal mol^{-1}. Recalling that this figure corresponds

to the relative solvation free energy (without a surface area dependent term in this calculation), it is interesting to compare this quantity with what would be obtained with the single topology results. This requires us the substraction of the free energy change for the perturbation of **8** to **1** in vacuum from the free energy change for the perturbation in the GB force field. The former is found to be 13.54 $\pm$ 0.08 kcal mol^{-1} while the later is reported in table 6.7 and is 19.62 $\pm$ 0.04 kcal mol^{-1}. The relative solvation free energy would then be 6.08 $\pm$ 0.09 kcal mol^{-1}, which agrees with the dual topology results to within statistical error.

With an estimated free energy change of 3.60 $\pm$ 0.48 kcal mol^{-1}, the perturbation in the bound state is more precise than the results obtained previously. However, the error estimates from the single run are similar to those obtained with an explicit solvent simulation and it is likely that lower spread of the results is an artefact of the small sample size (3). The relative binding free energy is found to be -2.41 $\pm$ 0.49 kcal mol^{-1}, which agrees with the single topology results.

With an implicit model of water, the possibility of performing a free energy simulation with a rigid protein environment has been hinted at in chapter 4 and 5. Here such a calculation is reported with the dual topology method and is found to yield essentialy identical results. Disappointingly, the errors are as large or larger than when protein side chain flexibility is taken into account. This give strength to the view that the major difficulty in the dual topology method lies in the thorough sampling of the configurational space available to the ligand.

6.5.2 Relative binding free energy of diclofenac and celecoxib

System setup

In the previous section we have demonstrated that equivalent answers can be obtained with the dual and single topology methods for the relative binding free energy calculation of two congeneric inhibitors. The single topology method was seen to yield more precise answers with less computational resources and should therefore be favoured when applicable. However, by construction, the single topology method requires the internal degrees of freedom of one solute to be perturbed into those of another. This is a relatively simple task when the two solutes of in-

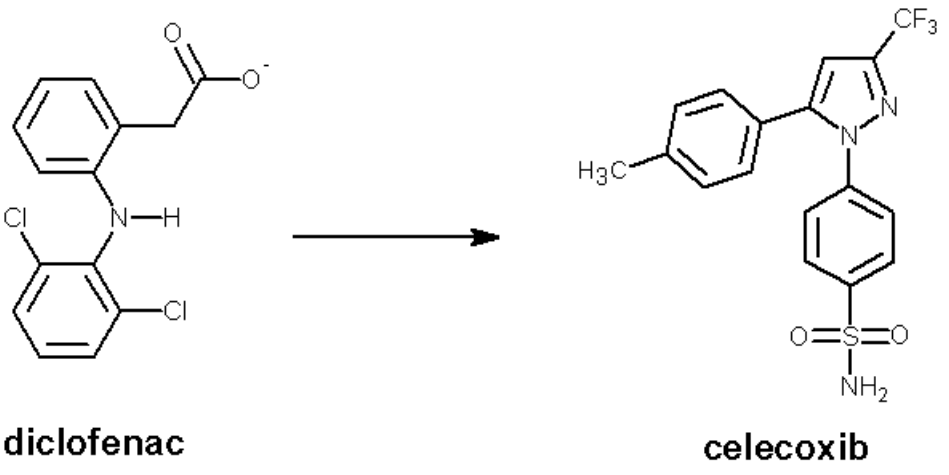


Figure 6.12: The drug diclofenac is perturbed into the drug celecoxib.

terest are structurally similar, but becomes increasingly difficult if the two solutes exhibit different chemical topologies. By contrast, the dual topology method requires no such information and the difficulty of setting up the simulation is not a function of the structural differences between the two solutes (although it could certainly affect the rate of convergence of the thermodynamic properties). As a proof of principle, we consider the perturbation of the drug diclofenac into the drug celecoxib, illustrated in figure 6.12. These two drugs have been used for the treatment of pain and are important pharmaceutical compounds. The potency of these drugs is often obtained by measuring an IC_{50}, which represents the concentration necessary to inhibit 50% of the enzymatic activity. IC_{50} can be related to binding free energies, using the Cheng-Prusoff equation.[161] Unfortunately, the measurement of one IC_{50} is very sensitive to the experimental protocol employed and the reported IC_{50} of the same ligand can vary by several orders of magnitude, depending on the assay conditions.[179,180] Thus it is difficult to relate the IC_{50} of inhibitors reported in different studies and it might be best to avoid converting those to an absolute binding free energy. Both diclofenac and celecoxib are known inhibitors of COX2. Their IC_{50} against COX2, obtained with the same assay conditions was established at 0.075 μM and 0.34 μM respectively.[181] This suggest very roughly that their relative binding free energy should be in the range of $\pm$ 1-2 kcal mol^{-1}.

A monomer of COX-2 complexed to the drug diclofenac was extracted from

the PDB structure 1PXX.[182] The structure was then aligned onto a monomer of COX-2 complexed to the brominated analogue of celecoxib (PDB structure 1CX2).[150] Celecoxib occupies a larger volume of the binding site and diclofenac is seen to occupy mainly the southern edges of the binding pocket. There is little backbone motion between the two different binding sites, and several amino acid side chains adopt the same conformation in the two binding sites. The main differences are seen between His90 and Arg120. In addition, the orientation of the hydroxyl group of Tyr348 and Ser530 differs. In the structure of diclofenac, two crystallographic waters are seen to interact with the ligand, and one of them is making strong hydrogen bonds to the carboxyl group of diclofenac. No crystallographic waters are present in the complex of the brominated analogue of celecoxib with COX-2. This is not evidence that these two water molecules have been expelled from the binding site however, as it might be an artefact of the structure refinement by the crystallographers.

The explicit solvent simulation methodology might prove difficult in this system as it is not clear how the two crystallographic waters should be considered. An implicit solvent framework simplifies the task but may lead to qualitatively wrong answers if specific water-solute interactions are important for the binding of diclofenac. Since the present focus of this work is on the feasibility of such free energy perturbation, concerns about accuracy of the model of the protein-ligand interactions were ignored.

Models of diclofenac (λ = 0.0) and celecoxib (λ = 1.0), in the conformation they adopt when bound to COX-2, were loaded in a modified version of ProtoMS21. An equilibrated protein scoop of COX-2, used in chapter 4 and 5 was further equilibrated for 600 K moves at a value of λ = 0.50. The resulting configuration was distributed over 12 simulations at different values of the coupling parameter (λ = 0.00, 0.10, .., 0.90, 0.95, 1.00). In the bound state, each simulation was equilibrated for 600 K moves prior to 1.8 M moves of data collection. The position of the heavy atoms in the backbone of the protein was frozen. In the unbound state, each simulation was equilibrated for 2K moves before collecting data for 200 K moves. The parameters for the GB force field, the non bonded interactions and the move probabilities were identical to those used previously. Simulations were

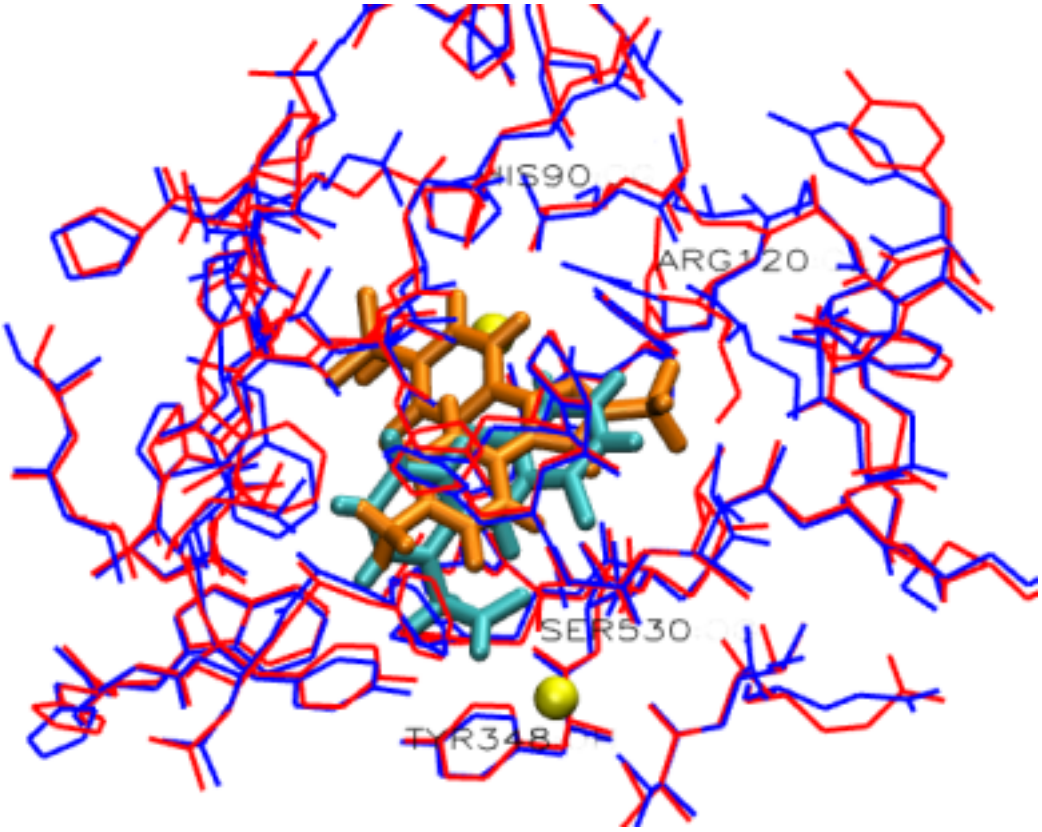


Figure 6.13: Overlay of the binding site of diclofenac (PDB code 1PXX) and the binding site of the brominated analogue of celecoxib (PDB code 1CX2). In blue, the binding site of diclofenac and in cyan diclofenac. In red, the binding site of the celecoxib and in orange, celecoxib. In yellow, two crystallographic waters present in the binding site that interact with diclofenac.

the protein is rigid were also considered. In this case, only solute moves were attempted. Each simulation was equilibrated for 50 K moves and data collected for 500 K moves.

Results

Table 6.8: Relative binding free energy of diclofenac and celecoxib[a]

run	$\Delta\Delta G_{solv}$	Error$_{grads}$	Error$_{blockA}$	Error$_{blockB}$
Dual Topology, unbound state				
1	42.10	0.10	0.05	0.03
2	42.11	0.10	0.05	0.03
3	42.14	0.10	0.07	0.03
Average	42.12 $\pm$ 0.03			
Dual Topology, δ = 1.25, n = 0, bound state				
1	28.55	1.07	1.01	0.78
2	27.55	1.31	0.89	0.71
3	28.20	1.11	0.70	0.55
Average	28.10 $\pm$ 0.85			
Dual Topology, δ = 1.25, n = 0, bound state, rigid protein				
1	-31.39	1.01	1.50	0.89
2	-28.23	1.22	1.64	0.85
3	-27.94	1.00	1.26	0.73
Average	-29.19 $\pm$ 3.22			

a For the average of 3 runs, the error estimate is the 95 % confidence limit of the mean, obtained from the independent simulations. All the figures are in kcal mol^{-1}

The simulation results are summarised in table 6.8. The free energy change in aqueous environment is 42.12 $\pm$ 0.03 kcal mol^{-1}. This quantity is very large and understood because diclofenac is negatively charged while celecoxib is neutral. Thus solvation of diclofenac is thermodynamically favoured.

The free energy change in the protein environment is 28.10 $\pm$ 0.85 kcal mol^{-1}. This suggest a relative binding free energy of -14.02 kcal mol^{-1}. As noted in the system setup section, there are uncertainties in relating measured IC_{50} to binding free energies. However, any difference of more than a few kcal mol^{-1} would be unreasonable, and thus celecoxib is much too favoured by the present force field.

In addition, a non bonded cutoff of 10 Å is probably not sufficient to handle correctly the change in long range electrostatic interactions, which almost certainly will be significant as one net charge is annihilated in the present perturbation. Finally, when perturbing closely structurally related ligands, one can hope to benefit from cancellation of errors in the force field parameters, particularly the torsional potentials. This would no longer be the case when dealing with different ligands and thus, obtaining accurate answers would be a challenge for modern biomolecular force fields. Rather than focusing on the accuracy of the present simulations, we wish to emphasis their precision. The spread of the results is less than 1 kcal mol^{-1}, suggesting that reasonably precise estimates of free energy changes can be obtained with the current protocol. In addition, the error on these calculations is similar to those reported in the perturbation of **8** into **1**. By contrast, building a zmatrix that would allow diclofenac to be perturbed into celecoxib would be a feat and almost impossible to automate by the single topology method.

Finally, we note that if no protein sidechain motion is allowed, the free energy change varies dramatically, to favour celecoxib by 71.3 ± 3.22 kcal mol^{-1}. This is because residues Tyr348 and Ser350 have to reorient their hydrogen bond donating groups to interact with diclofenac. In addition, some degree of plasticity of the binding site is necessary to accommodate the change of shape of the ligand. If no such motions are allowed, and because the protein model was built from the crystallographic structure of COX2 complexed with an analogue of celecoxib, it is not suprising that celecoxib would be more favoured, although the magnitude of the free energy change is unexpected. Although this was not attempted in this work, it might be interesting to perform the same calculation with the two ligands modelled into the binding site obtained from the complex of diclofenac and COX2 (PDB code 1PXX). If sufficient sampling has been performed, the free energies should be similar.

The present results emphasise that the good predictive power obtained in the free energy calculations reported in chapter 4 and 5, where no protein flexibility is allowed, might not be observed in a general case. Presumably, protein flexibility becomes increasingly important if one is interested in structurally diverse, non congeneric sets of ligands.

6.6 Conclusion

An implementation of the dual topology method in ProtoMS21 has been described. It allows the calculation of free energy differences by a different approach from the more standard, single topology method. Because a naive implementation of the dual topology method suffers from numerical instabilities at the simulation of the end states (at λ 0.00 or λ 1.00), it is necessary to combine the methodology with a softening of the intermolecular interactions (particularly, the Lennard Jones term). The protocol is applied to the calculation of the relative solvation free energy of ethane and methanol, solvated in a box of TIP4P water. Equivalence between the single and dual topology methods is demonstrated. In addition, it is shown that the softcore parameters can be optimised such that the imprecision in the calculated free energy change is minimised. Different methods to estimate errors on the calculated free energy changes from a single simulation are compared to the confidence interval obtained from the spread of five independent simulations. None of these methods are found to reliably estimate the true uncertainty. The dual topology method is then applied to the calculation of the relative solvation free energy of benzene, ethane and methanol. The results appear well converged and suggest that the main source of discrepancy between theory and experiment lies in the atomic partial charges of methanol, obtained by the AM1/BCC methodology. The perturbation of benzene into ethane or methanol is achieved trivially with the dual topology approach, while with the single topology method, it would have been difficult to convert a linear molecule into a planar, cyclic molecule. The dual topology method is then applied to the calculation of the relative binding free energy of the two inhibitors of cyclooxygenase-2, compound **8** and **1**. It is shown that the single topology method yields much more precise free energy changes than the dual topology method. This is because these two congeneric inhibitors are structurally very similar and the perturbation of the ligand is localised onto a small part of the molecule. Because the dual topology method samples the internal degrees of freedom of the ligands independently, the free energy difference is a function of a much larger number of degrees of freedom and requires more configurational averaging to yield results of similar precision. The single topology method should

be thus preferred when it is applicable. In addition, it is shown that the optimum softcore parameters can be system dependent and that the binding of the two ligands can be decomposed into more complex thermodynamic cycles involving a mixed use of single and dual topology to gain additional insights into the components of the relative binding free energy. Finally, the dual topology method can be combined with the implicit solvent techniques described previously. This eliminates completely any inprecision on the contribution of the unbound state to the calculated relative binding free energy, although the simulation of the bound state is no easier (on this system). The dual topology method is then applied to the calculation of the relative binding free energy of two very structurally different inhibitors of COX2, diclofenac and celecoxib. The simulations fail to calculate an accurate free energy difference. This can be due to an incorrect treatment of long range electrostatics, lack of sampling of the protein degrees of freedom, and improper torsional parameters for the ligands. However, the results are essentially as precise as those obtained in the perturbation of the two congeneric, celecoxib analogues, suggesting that the present methodology can be applied to investigate the binding of structurally different compounds. Finally, it is shown that while the incorporation of protein flexibility was not very important when calculating the relative binding free energy of similar ligands, it appears necessary to deal with structurally different compounds.

Chapter 7

Concluding remarks

“A poet once said ”The whole universe is in a glass of wine”. We will probably never know in what sense he meant that, for poets do not write to be understood. But it is true that if we look at a glass closely enough we see the entire universe. There are the things of physics: the twisting liquid which evaporates depending on the wind and weather, the reflections in the glass, and our imaginations adds the atoms. The glass is a distillation of the Earth’s rocks, and in its composition we see the secret of the universe’s age, and the evolution of the stars. What strange array of chemicals are there in the wine? How did they come to be? There are the ferments, the enzymes, the substrates, and the products. There in wine is found the great generalization: all life is fermentation. Nobody can discover the chemistry of wine without discovering, as did Louis Pasteur, the cause of much disease. How vivid is the claret, pressing its existence into the consciousness that watches it! If our small minds, for some convenience, divide this glass of wine, this universe, into parts - physics, biology, geology, astronomy, psychology, and so on - remember that Nature does not know it! So let us put it all back together, not forgetting ultimately what it is for. Let it give us one more final pleasure: drink it and forget it all!”

Richard P. Feynman

In the pharmaceutical industry it is commonly viewed that free energy calculations are too time consuming to be of practical use as a tool for drug design. This is mainly because multiple simulations of several tens of millions of Monte Carlo moves are necessary to obtain binding free energies with a precision similar to those measured experimentally. The severe requirement on the amount of sampling is due in part to the necessity of averaging the free energy over the degrees of freedom of several thousands solvent molecules that solvate the protein-ligand complex.

This research set out with the idea of combining the successful theories of implicit solvation with the rigorous statistical mechanics framework that allows the calculation of free energy changes. By treating the solvent as a continuous medium, the complexity of the systems is greatly reduced and the number of Monte Carlo moves required to obtain converged free energies is dramatically reduced. The method should therefore provide a means to obtain precise free energies more rapidly, hence allowing a larger number of compounds to be studied in the same amount of time. There is considerable evidence in the literature that implicit theories of solvation, when properly parameterised, yield solvation free energies in good agreement with explicit solvent simulations. However, few published studies have attempted to calculate binding free energies in protein-ligand complexes solvated implicitly. It is therefore important to test the accuracy of the calculated binding free energies with this protocol, and compare them with the results obtained by more established methods.

In chapter 2, a protocol for the calculation of relative binding free energies in an implicit solvent was proposed. The protocol relies on the AMBER[34] and GAFF[107] force fields and a GBSA theory of implicit solvation.[96] Because no suitable parameterisation of a GBSA force field that covers chemical groups commonly encountered in drug-like molecules was available for the GAFF force field, initial efforts were focused on the derivation of appropriate parameter sets. This work followed standard parameterisation protocols available in the literature.[117,118] The validation of the top performing parameter sets was carried out by calculating potentials of mean force for the association of several small molecules in solution. The small molecules were chosen so as to encompass a wide variety of intermolec-

ular interactions. Interestingly, deficiencies in standard parameterisation protocols were identified and means to overcome those were suggested. This work has thus produced optimum parameter sets for a GBSA/GAFF force field and highlighted methodological issues in GBSA force field parameterisation. The results of this study were published under the title “The parameterization and validation of generalized Born models using the pairwise descreening approximation”, in the Journal of Computational Chemistry.[100]

In chapter 3, the combination of Generalised Born algorithms with Monte Carlo sampling for the calculation of relative binding free energies in protein ligand complexes was considered. Initial results proved disappointing. Because of the non local nature of the GBSA energy, the two methods do not integrate well and lead to a dramatic loss of efficiency when the methodology is applied to larger systems. These issues were resolved by introducing approximations in the calculation of the GB energy and specialised Monte Carlo acceptance tests that permit the thermodynamic properties of the system in a GBSA force field to be calculated while using simpler, more efficient, theories of solvation. The approximations were carefully tested and found to have a minor or negligible impact on the calculated binding free energies. The resulting protocol was only 4-5 times slower than Monte Carlo simulations in vacuum, which is typical of the efficiency of molecular dynamics simulation of proteins in a GBSA forcefield. In addition to delivering a methodology for rapid binding free energy calculations, it is expected that the method will be useful for conducting Monte Carlo simulations of protein folding. The results of this study were published under the title “Efficient generalized Born models for Monte Carlo simulations”, in the Journal of Chemical Theory and Computation.[183]

In chapters 4 and 5, the methodology developed in the two previous chapters was applied to two protein-ligand systems. The two systems differed widely in the features of their binding site and the nature of the protein-ligand interactions. Cyclooxygenase-2 has a buried, hydrophobic binding site while neuraminidase has a polar, solvent exposed binding site, with crystallographic waters mediating interactions between the ligand and the protein. In the former case, the ability of a Generalised Born methodology to treat desolvation of a ligand and the binding site

was assessed. In the later case, it was unclear if an implicit treatment of water could yield meaningful results in such a polar, solvent exposed system. The implicit solvent methodology was found to give predictions of high quality, both qualitatively and quantitatively. Small errors due to the improper treatment of ligand desolvation were observed in the case of COX-2. However, these could be easily corrected by a simple protocol, where small pockets of high dielectric present in the binding site were filled with non interacting spheres. While crude, this protocol made the implicit solvent simulations as accurate as the explicit solvent simulations. Because this approach amounts to a better calculation of the atomic Born radii, the protocol could be improved by implementing more elaborate Born radii calculations available in the literature.[140]

When applied to neuraminidase, the implicit solvent protocol was found to yield superior predictions as compared to the explicit solvent protocol. It was found that the ligands would experience dramatically different interactions with the protein when alternative theories of solvation were employed. As a result, different relative binding free energies were obtained. The impact of protein flexibility on the calculated relative binding free energies was investigated. It was found that while the neglect of the protein internal degrees of freedom gave poorer quantitative agreement, the qualitative ranking of the inhibitors in both series was essentially unchanged. The convergence of the predictions with each protocol was examined and it was concluded, retrospectively, that high quality predictions could have been obtained in 1 to 2 hours with the implicit solvent protocol and neglect of protein flexibility, 5-7 hours with the implicit solvent protocols and about 10 hours with the explicit solvent simulations, provided sufficient CPUs are available to run calculations on the whole series simultaneously. In both cases, with the protocols employed in this work, this would amount to 130-150 CPUs. This figure is not unreasonably large. Through advances in distributed computing technologies it is likely that such a number of CPUs could be obtained at a reasonable cost. At the time of writing, the cluster of the University of Southampton has 900 available CPUs.

It would be unwise to formulate definitive conclusions from results obtained on only two different protein ligand systems, but in the hands of the author the

implicit solvent methodology has proven a robust and efficient alternative to the more established protocols that make use of explicit solvation. Further work could focus on more elaborate implicit treatments of solvation, but given the good performance of the present implicit models, it might be better to extend the testing to a larger set of protein-ligand complexes and concentrate on better methodologies only if problematic systems are encountered.

The bulk of this thesis has focused on the introduction of efficient implicit solvent models so as to calculate rapidly relative binding free energies of protein ligand complexes. At the end of chapters 4 and 5 it was concluded that high quality predictions could be obtained in a few hours of simulation, with reasonable computational requirements[1]. Thus arguments against the adoption of free energy calculations in the pharmaceutical industries should not be targeted at their high computational cost. A major problem associated with this technology is that relative binding free energies are usually calculated between structurally similar ligands. As a result, free energy studies are often performed on a congeneric series of compounds. This is partly because the master equations that govern the calculation of free energy differences converges more readily when applied to similar systems, but also because of the difficulty of devising schemes to inter-convert two structurally unrelated ligands into each other. In chapter 6, alternative methods to convert one ligand into another were explored. In the single topology method, the internal degrees of freedom and force field parameters of one ligand are gradually modified to match the geometric and parametric values of the other ligand. This make it difficult to transform one ligand into a structurally unrelated ligand. In the dual topology method, the two ligands are simulated simultaneously and their interaction energy with the surrounding environment is scaled such that at either ends of the simulation, only one ligand is present in the surrounding environment, while the other is in a ideal thermodynamic state. Because this method does not attempt to modify the internal degrees of freedom of one ligand to match those of the other, it is more generally applicable. Numerical instabilities are encountered in the simulation of the end states, but these can be overcome through the introduc-

[1] the dramatic improvements over the past decade appear more linked to the increase in available computing resources than to fundamental advances in the theories of free energy calculation...

tion of a softening of the intermolecular energy terms. In addition constraints that prevent the ligand being decoupled from drifting out of the binding site have to be employed. The author stresses that most dual topology simulations published in the literature are actually hybrid single/dual topology, where a portion of the ligand is common to both molecules. Such methods suffer from the same difficulties encountered in the single topology approach (complex system setup), and cannot be applied if the two ligands do not share common structural features. By contrast, the approach developed here is completely general.

This method was initially applied to the calculation of the relative solvation free energy of ethane, methanol and benzene. After demonstrating that equivalent answers could be obtained with either approach, the method was applied to the calculation of the relative binding free energy of a set of COX-2 inhibitors. Two were congeneric inhibitors and two others were structurally different. In the later case, protein side chain flexibility had a dramatic impact on the calculated relative binding free energies. Thus, the good predictions observed in chapter 4 and 5 for models with no protein flexibility is presumably due to the fact that only congeneric inhibitors were considered. A more general binding free energy calculation protocol will almost certainly have to address protein flexibility.

The generality of the dual topology method comes with a price. Relative binding free energies were found to converge more slowly than with the single topology method. However, as demonstrated in the previous sections, computational expense should no longer be considered the primary issue associated with free energy calculations. Future work will consist in the application of the methodology to the calculation of the relative binding free energy of different scaffolds to a protein. In the pharmaceutical industry, identification of an appropriate scaffold that binds to a protein interest (“hit”) is often more difficult than the optimisation of a micromolar inhibitor (“lead optimisation”). The ability of a free energy method to identify promising scaffolds from a set of decoys would be a significant advance. Another interesting application of the methodology would lie in the refinement of ligand poses obtained by molecular docking. Typical docking programs sometimes identify different binding mode for one ligand. These binding modes cannot be discriminated by the crude empirical scoring functions employed by the docking

algorithms. However, free energy based methods may be successful in identifying the correct binding mode(s).

Other workers have recently proposed a variety of binding free energy estimation schemes which use QM/MM methods to incorporate polarisation effects in the calculated binding free energies.[184,185] These schemes currently score a single pose and thus depart from the statistical mechanic route to free energy calculation, and one way to improve on these studies could be to incorporate configurational sampling in the QM/MM calculations. The author should like to point out that sufficient sampling has to be performed to obtain precise binding free energies with a classical forcefield. By increasing the cost of the potential energy evaluation, there exists the risk of compromising on the amount of sampling carried out. Thus quantum mechanical treatments of ligand binding may not be the best route to high throughput relative binding free energy calculations.

Another effort currently pursued in the field is the calculation of absolute binding free energies. This typically requires decoupling of the complete ligand from the protein and aqueous environment, although distance based PMFs methods have been proposed.[92,186] The ability to predict an absolute binding free energy would be a significant advance. However, in the context of a structure based drug design project, free energy methods are more likely to make an impact after some hits have been identified. Once setup, biological assays can be performed relatively rapidly and it is thus likely that a molecular modeller would have access to a ligand structure and binding affinity before starting a free energy calculation project. Thus, absolute binding free energy predictions might not be the best way to advance applications of free energy calculations in drug design.

Before the advent of routine relative binding free energy calculations, a number of other issues that have not been considered in this work will also have to be addressed. First, the accuracy of biomolecular force fields may have to be improved. It is now becoming typical to parameterise force fields against free energies of solvation of small molecules. Perhaps in the future, such parameterisation could be extended to include relative binding free energies, in balanced protein-ligand datasets.

Second, there is a need to develop Monte Carlo moves that allow for large scale modifications of the simulated system. RETI is one example of such method as it allows possibly widely different configurations to be exchanged between simulations. However, the availability of Monte Carlo moves that focus sampling where it is needed, for example a particular protein side chain, known to adopt different conformers depending on the nature of the ligand complexed into the binding site, would go some way towards extending the reliability of free energy calculations.

Third, the ligand setup has to be fully automated. Modern force fields provide large libraries of torsional angle potentials and rely on quantum mechanical methods to obtain atomic partial charges. The main obstacle to automated ligand setup might lie in the automatic generation of a suitable zmatrix. However, this is simply a technical difficulty that can be solved with sufficient programming skill.

Fourth, the protein setup has to be fully automated. The assignment of the protonation state of acidic residues can prove difficult without any *a priori* knowledge of the local pkA of these residues. Yet they can have a significant impact on the calculated relative binding free energies by modifying the electric field surrounding the ligand. Molecular dynamic simulations at a constant pH represent one step toward the resolution of this problem by recognising that the protonation state of acidic residues is conformation dependent.[187–189] In addition, a more elaborate treatment of long range electrostatics would be desirable. The adoption of an efficient Ewald sum based method in Monte Carlo simulations would go some way towards the resolution of this problem.

Lastly, free energy calculation protocols (number of windows, extent of sampling..) are often decided arbitrarily or after exploratory work. In a high throughput context, this would not be practical and it would be useful to design intelligent protocols that determine the amount of sampling necessary to obtain converged results.

In conclusion, the broad aims of this research have been satisfied. The combination of implicit solvation with a rigorous statistical mechanics framework has been shown to be a competitive alternative to the traditional approach of explicit solvation. The application of various free energy methodologies to different protein-ligand systems has shown that the computational expense of a free energy

calculation should no longer prevent its wider adoption. This research has also attempted to make free energy calculations more generally applicable, although extensive validation of the proposed methodology has still to be carried out. Finally, before free energy simulations can be employed routinely as a scoring functions by the pharmaceutical industry, a number of other issues will have to be addressed.

Appendix A

Solving the integrals of chemical problems

"Chemistry is a trade for people without enough imagination to be physicists."

Arthur C. Clarke.

Systems that are of interest to chemists are usually modelled by hundreds or thousands of atoms and the integrals in equation 1.8 are very complex and multi dimensional. Their evaluation requires the use of numerical methods that will be described in this section.

A.1 The curse of dimensionality

Suppose we wish to evaluate the volume V of the unit sphere S in a space of dimension k. By unit sphere we mean that a point X of coordinates $(x_1, ..., x_k)$ from R^k belongs to the sphere S if the relation A.1 is true.

$$\sum_{i=1}^{k} x_i^2 \leq 1 \tag{A.1}$$

This problem can be represented by the integral

$$V = \int_S dx_1 ... dx_k \tag{A.2}$$

which solves to

$$V = \frac{\pi^{k/2}}{\Gamma(\frac{k}{2}+1)} \tag{A.3}$$

where Γ is the gamma function.

Assume we wish to solve this integral using a numerical approach. We can reformulate the problem and seek to calculate the ratio of the volume of the unit sphere to its bounding rectangle. Mathematically this means solving equation A.4.

$$I = \frac{V}{V_R} = \frac{\int_{[-1,1]^k} I_S(x_1,...,x_k)dx_1...dx_k}{\int_{[-1,1]^k} dx_1...dx_k} \tag{A.4}$$

Where $I_S(X)$ is 1 if X belongs to S and 0 otherwise. The analogy with equation 1.8 should be obvious. To evaluate A.4 we could use standard quadrature techniques such as the trapezium rule or Simpson's rule. All these methods involves the uniform spreading of points over $[-1,1]^k$ and averaging of the integrand over these points. If we decide to do so, and spread m points on each of the k dimensions, we require the evaluation of m^k points. The exponential increase in the number of points required to perform the quadrature means that the evaluation of multi-dimensional integrals is not practical using this approach (even with only 10 points per dimension, one has to perform 10 billion evaluations for a problem in 10 dimensions).

An alternative approach to tackle this problem is to use a Monte Carlo method.[190] Instead of uniformly spreading m^k points, N points are randomly and uniformly distributed over $[-1,1]^k$. The integrand is then estimated by the average of the N points.

$$I_{est} = \frac{1}{N}\sum_{i=1}^{N} I_S(X_i) \tag{A.5}$$

The law of large numbers guarantees that in the limit of an infinity of points the estimated integrand converge to the exact result. The central limit theorem also shows that the standard error associated with the estimated integrand I_{est} is of the order of $N^{-1/2}$.[190] Of great practical interest is that the number of points N used

is now under control. Thus one can always decide to draw more samples to reduce the error until it is satisfactory. Crucially, the error is independent of the dimension of the integral. The quadrature techniques discussed in the previous section usually gives a better approximate with less points than Monte Carlo techniques for low values of k, but as the dimension k increases they are eventually outperformed by the Monte Carlo approach. Thus at first sight it appears that Monte Carlo methods do not suffer from the 'curse of dimensionality'.

Since the volume of the rectangle $[-1,1]^k$ is 2^k, we can write

$$\frac{V}{V_R} = \frac{\pi^{k/2}}{\Gamma(\frac{k}{2}+1)2^k} \tag{A.6}$$

Table A.1 shows the value of the ratio V/V_R as a function of the dimension k. It is clear that the volume of the unit sphere S in proportion to the rectangle becomes extremely small as the dimension k increases.

Table A.1: Value of V/V_R as a function of the dimension k

k	Proportion
1	1.00×10^{0}
2	7.85×10^{-1}
3	5.24×10^{-1}
5	1.64×10^{-1}
10	2.49×10^{-3}
50	1.54×10^{-27}
100	1.87×10^{-69}

This has severe implications for the Monte Carlo approach. For high dimensions, because the random points are distributed uniformly, the integrand I_S is 0 for most of the samples. In fact, for k=50, even after a few million samples, it is very likely that the estimated integrand will still be 0. While it is true that 0 is not a bad estimate of I for high dimensions, this is not much comfort if we are interested in properties that depends on the exact volume of the sphere. It is clear that even though the Monte Carlo does not suffer from the curse of dimensionality, its application to multi-dimensional integral can be unsuccessful. That only a tiny fraction

of the volume of phase space has a non zero integrand is a common occurence when dealing with the integrals of statistical mechanics. In order to have a chance, we need to introduce methods that take into account the shape of the integrand.

A.2 Importance Sampling

The simple Monte Carlo approach introduced in the previous section fails to provide an accurate estimate of the integrand of equation A.4 when dealing with a large dimension k because the probability that a random point X selected uniformly from $[-1,1]^k$ belongs to the sphere S is extremely small. The solution is to select random points X from a non uniform distribution $\pi(X)$ suitably chosen to favour the selection of points in the region where the integrand $I(X)$ is significant. The bias introduced by the non uniform selection of points is then corrected according to the following equation

$$I_{est} = \frac{1}{N} \sum_{i=1 \, from \pi}^{N} \frac{I(X_i)}{\pi(X_i)} \tag{A.7}$$

It is clear that equation A.5 is a particular case of equation A.7 where π is the uniform distribution. The choice of a good distribution function π depends obviously on the function I. The ratio $I(X)/\pi(X)$ should be approximately constant over the range of integration. This is because regions that have a large integrand I(X) should have a large weight $\pi(X)$ so that many samples are drawn from this region, while regions with a small integrand should be sampled infrequently. In addition it should be easy to draw elements from the distribution π.

Consider the following example. We wish to estimate the quantity I

$$\begin{aligned} f(x) &= 3x^2 \\ I &= \int_0^1 \frac{f(x)}{\pi(x)} \pi(x) dx \end{aligned} \tag{A.8}$$

using Monte Carlo integration and the following importance distributions

$$\begin{aligned} \pi_0 &= 1 \\ \pi_1 &= 2x \\ \pi_2 &= 3x^2 \\ \pi_3 &= 4x^3 \end{aligned} \tag{A.9}$$

Note that according to our criterion, π_2 is the best importance sampler, since the ratio $\pi_2(x)/f(x)$ is constant, but this is because $\pi_2(x) = f(x)$.

In general, in order to draw a random number in the interval [a,b] according to a distribution g(x), the following method can be used.[190]

1. evaluate G(x) = $\int_a^b g(x)dx$
2. solve u = G(x) to obtain $x = G^{-1}(u)$

So we just have to generate a uniform random number u between [a,b] and then apply $x = G^{-1}(u)$ to generate samples from the given distribution g(x).

Table A.2 shows the average estimated integrand and the standard deviation after 10 simulations of N=100 samples with each importance sampler.

Table A.2: Estimates of I from eq A.8 by importance sampling

Function	Average	Deviation
π_0	1.027	0.111
π_1	0.986	0.031
π_2	1.000	0.000
π_3	0.999	0.036

It is manifest that π_0 performs worse than the other importance samplers. Application of π_2 yields systematically the correct answer for any selected point. Note however, that in order to draw samples from π_2 on [0,1] we had to solve $\int_0^1 \pi_2(x)dx$ which is precisely the integral we are trying to estimate. In general, the smallest variance can be obtained by selecting $\pi(x) = c \times f(x)$.[190] On a realistic application, the shape of the integrand can be quite complex and finding a very efficient

function π from which samples can be drawn easily can be as difficult as solving the integral of interest.

If we study equation 1.8 we see that an appropriate distribution function would be $\pi(x) = \exp(-\beta U(x))/Z_{N,NVT}$ which is in fact the Boltzmann distribution. If the product $A(x)\exp(-\beta U(x))/Z_{N,NVT}$ is not dominated by A, then such importance distribution is likely to provide a good estimate of 1.8 with a reasonable number of samples. Unfortunately we can not manipulate directly the Boltzmann distribution because the knowledge of the normalisation factor $Z_{N,NVT}$ would requires us to enumerate all the states the system of interest can adopt.

A.3 Markov Chains

In the previous section we have shown that standard Monte Carlo integration can be markedly improved when samples are drawn from a distribution π which has been selected such that it increases the likelihood of picking samples in the region where the integrand of the function of interest is high. In statistical mechanics, the integrals we are dealing with are very large (several hundreds of dimensions) and have a very complex shape, with several separated regions contributing to the integrand. Finding *a priori* a suitable weighting function π is almost as difficult as solving the integral. The purpose of this section is to show how this can be accomplished.

A.3.1 Definition

A stochastic process is a procedure which entails the generation of a number of states according to probabilistic rules. A Markov process is a form of stochastic process where the probability of generating a new state is dependent only on the state the process is currently in. The sequence of states that is generated by repeated applications of a given Markov process forms a Markov chain and the act of generating a new state in the Markov chain is called a trial.[191] A Markov process can be represented by a matrix. Equation A.10 models a 3 states Markov process Π where p_{ij} is the probability to make a transition from state i to state j and p_{ij}^n is the probability to make this transition in exactly n steps.

$$\Pi = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \tag{A.10}$$

Because the elements of the matrix Π are probabilities, each element π_{ij} should be a non negative number and the sum of each row should be equal to 1. Because we are interested in properties of a special class of Markov chains we need our Makov chain specified by Π to obey these other properties :[191]

1. *The Markov Chain must be positive recurrent.* A state y is recurrent if the probability of, starting in y, and eventually returning to y, is equal to 1. Otherwise y is said to be transient. If the Markov chain can returns to the state y in a finite number of steps, the state is said to be positive recurrent. An alternative way to see the difference between a transient state and a recurrent state is that recurrent states are infinitely visited while transient states are only visited a finite number of times in the limit of an infinite number of trials. The Markov chain is positive recurrent if all of its states are positive recurrent.

2. *The Markov Chain must be irreducible.* This means that the probability of connecting any two states x,y after a number of trials n is non zero.

3. *The Markov Chain must be aperiodic.* The chain must not cycle through a finite number of sets of states.

If a state y is positive recurrent and aperiodic, it is ergodic and if all the states of the chain fulfill this condition, the Markov chain is said to be ergodic. If the Markov chain is ergodic, then the following equation is true :

$$\pi_j = lim_{n\to\infty} p_{ij}^{(n)} \quad \forall \quad i \tag{A.11}$$

Equation A.11 states that repeated applications of an ergodic Markov process converges towards a single probability for state j that is independent of the initial state i. The vector $\pi = (\pi_j)$ is called the limiting distribution of the chain. Equation A.11 can also be formulated as

$$\pi = lim_{n\to\infty} \quad \rho^{(1)}\Pi^{(n)} \tag{A.12}$$

where $\rho^{(1)}$ is any arbitraty initial distribution and $\Pi^{(n)}$ is the n-th application of Π . Π^n can be calculated through the following relations (noting $\Pi = \Pi^{(1)}$).

$$\begin{aligned} \Pi^{(n)} &= \rho^{(1)} \prod_{i=1}^{n} \Pi^{(i)} \\ &= \rho^{(n-1)}\Pi \end{aligned} \tag{A.13}$$

Equation A.12 shows that the limiting distribution π is an eigenvector of the transition matrix Π with eigenvalue 1.

The ergodic theorem shows that states X_i drawn from Π obey the following relationship.

$$lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \int_{R_k} f(x)\pi(x)dx \tag{A.14}$$

This theorem is very important becauses it says that if we have specified a transition matrix such that it has one (and only one) limiting distribution π, then in the limit of a large number of samples we are drawing from the distribution π without the need of specifying π *a priori*. Thus Markov chains are a powerful method to sample from complex distribution function π. Returning to equation 1.8, we now see that our problem is to formulate a Markov Chain such that its limiting distribution is the Boltzmann distribution.

A.3.2 Detailed Balance

A Markov Chain obeys the principle of microscopic reversibility or detailed balance if for every pair of states i,j we have the following equality :

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{A.15}$$

where π_i is the probability of state i in the limiting distribution π and p_{ij} is the transition probability from the matrix Π. It is simple to show that if the Markov Chain obeys detailed balance, then π must satisfy the eigenvalue equation $\pi\Pi = \pi$.

$$\begin{aligned} \sum_i \pi_i p_{ij} &= \pi_j \\ \sum_i \pi_j p_{ji} &= \pi_j \\ \pi_j \sum_i p_{ji} &= \pi_j \\ \pi_j &= \pi_j \end{aligned} \tag{A.16}$$

The first line is the property of an eigenvector, the second line introduces detailed balance, and the summation is unity on the third line because the rows of the matrix Π sum to one.

The utility of detailed balance is that it provides an easy way to modify the elements p_{ij} of a transition matrix Π such that repeated applications of Π leads to the limiting distribution of interest π. It is important to remember however that many transition matrices Π who do not obey detailed balance have a limiting distribution π, thus the use of detailed balance in the construction of a transition matrix is merely a convenience.

A.3.3 Performance of a Markov chain

While equation A.14 tell us that a suitably chosen transition matrix Π converges a Markov Chain towards a unique limiting distribution π it does not say how quickly the chain converges. Knowing this information is useful because in any simulation, the number of samples n is necessarily finite and in this situation it is useful to discard the data collected during the first k iterations because they were preleved when the Markov Chain was not well converged and have an adverse effect on the remaining statistics.

Consider the ergodic transition matrix Π and the eigenvalue problem

$$\mu\Pi = \lambda\mu \tag{A.17}$$

The Perron-Frobenius theorem states that Π has one dominant eigenvalue λ_d which is positive and all the elements of its associated eigenvector μ_d are non negative. Furthermore, if the rows of the matrix Π sums to unity, which is the case

for a stochastic matrix, then $\lambda_d = 1$. μ_d is also proven to be the only non negative eigenvector ($\mu_d(i) \geq 0 \ \ \forall i$) and clearly, only this eigenvector can be interpreted as a probability distribution π. By virtue of the Perron-Frobenius theorem, it is also shown that every other eigenvalue λ_o must be of magnitude lower than unity.[192] Π is a square matrix and can be diagonalised.

$$\Pi = PDP^{-1} \tag{A.18}$$

And we have

$$\begin{aligned} \Pi^{(n)} &= (PDP^{-1})...(PDP^{-1}) \\ &= PD^{(n)}P^{-1} \end{aligned} \tag{A.19}$$

D has only his diagonal elements non null and they are equal to the eigenvalues. Since $\lambda_d = 1$ and any other $\lambda_o \leq 1$, and because the diagonal elements of $D^{(n)}$ are simply λ^n then, assuming that the eigenvalue have been ordered by magnitude it comes that

$$D^{(n)}_{n\to\infty} = \begin{pmatrix} 1 & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & ... & 0 \end{pmatrix} \tag{A.20}$$

The other eigenvalues λ_0 must vanish to 0 for a large number of trials. The property of A.20 gives a mean to rate the performance of a a transition matrix. Suppose that we have determined a set of transition matrices $\Pi_k, k = (0,..,m)$ that converges towards the same limiting distribution π. Each matrix is diagonalised and the diagonalised matrices whose non dominant eigenvalues have the smallest magnitude will converge more quickly toward A.20.

A.3.4 The Metropolis Monte Carlo algorithm

We restate the procedure that implements the Metropolis Monte Carlo method.

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 $\chi = (\pi_j/\pi_i)$

4. If the move is accepted set i = j, otherwise i = i

5. Accumulate any property of interest A(i)

6. Return to 1 or terminate after a number of iterations

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

$$\begin{aligned} \pi_i Q_{ij} &= \pi_j Q_{ji} \\ \pi_i p_{ij} \alpha_{ij} &= \pi_j p_{ji} \alpha_{ji} \\ \pi_i p_{ij} \frac{\pi_j}{\pi_i} &= \pi_j p_{ji} \\ p_{ij} &= p_{ji} \end{aligned} \tag{A.21}$$

And we see that detailed balance is respected if the unmodified transition matrix is symmetric i.e, the probability of moving from i to j, *prior* weighting by π_i and π_j is the same as the probability of moving from j to i.

A.4 The connection with molecular simulations

When running a simulation of a chemical system of interest, it is easy to forget that we are estimating an integral with the aid of Markov chains and Monte Carlo importance sampling. This is because, unlike the the examples discussed previously, the number of states the system can occupy is embarrassingly huge and the transition matrix Π that specify our Markov process is so large that no computer will ever be able to form this matrix (this unfortunately prevents the straigthforward application of the linear algebra techniques discussed in the previous section to assess the ability of this matrix to approximate the distribution π in the smallest amount of steps). When using the Metropolis algorithm, we are randomly selecting a trial state j given a state i and this means that we are interested in only one row of Π at a time. Furthermore, because a move is proposed solely on the basis of

the transition probabilities p_{ij}, the acceptance probability χ needs to be evaluated only for the pair of states (i,j) and the calculation of π_j/π_j can be performed "on the fly".

It is useful to reflect on the way the attempted Monte Carlo moves underlies the transition matrix. Consider, for simplicity the random displacement of a molecule in state i by a maximum amount of dr in a cubic box of side a. There is a finite number of states j in the sphere of volume $4\pi/dr^3$ that can be reached from state i using this move. This number is smaller than the total number of states. Matrix A.22 highlights that for only a fraction of the number of states, there exist a probability of trying a transition for which p_{ij} is positive.

$$\Pi = \begin{pmatrix} \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & p_{ij} & p_{ii} & p_{ik} & 0 & 0 & 0 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & p_{nn} & \ldots \end{pmatrix} \tag{A.22}$$

Suppose the transition i to j is accepted, the same situation would apply for state j at the next iteration and it might take several iterations before we can reach state n if it is sufficiently far. The situation can be complicated by the acceptance test. Suppose state i and n have a significant weight in the distribution π. It is desirable that these states are present in the generated Marvov chain. If states i and n can not be connected in a single step, and that the sequence of states (j,k) that can connect them is such that the ratio π_k/π_j is close to zero, then the overall probability of traveling from i to n is very low and by consequent, the convergence of the Markov chain will be very slow. Even though there is a finite probability of making this transition (otherwise the chain would not be ergodic and yield a limiting distribution), the chance of this happening can be so low that the number of samples to draw before the transition occurs can be astronomically large. In the language of chemistry, we say that states i and n are separated by an energy barrier and we picture the probability that a ball has enough kinetic energy to overcome this barrier. In this case, the simulation may appear to converge toward a limiting distribution ρ different from the desired distribution π and any property evaluated

will be wrong. Even worse is the fact that there are no rigorous mathematical way to detect whether or not this problem has occured. In this situation the chemist intuition can be invaluable in detecting pathological cases. In chemical problems the potential energy function U is usually sufficiently complex that there exist several regions of low energy separated by high energy barriers and, assuming neighbouring states in phase space have neighbouring index, the vector π_{NVT} is characterized by short sequences of high probability separated by long sequences of low probability. We see then that the convergence of the Markov chain can be greatly improved if we design 'intelligent' moves such that transitions between regions of high probability are attempted.

References

[1] Jorgensen, W. L. *Science* **2004,** *303,* 1813-1818.

[2] Marrone, T. J.; Briggs, J. M.; McCammon, J. A. *Annu Rev Pharmacol* **1997,** *37,* 71-90.

[3] Tari, L. W.; Rosenberg, M.; Schryvers, A. B. *Expert Review of Proteomics* **2005,** *2,* 511-519.

[4] von Dongen, M.; Weigelt, J.; Uppenberg, J.; Schultz, J.; Wikstrom, M. *Drug Discov Today* **2002,** *7,* 471-478.

[5] Dean, P. M.; Lloyd, D. G.; Todorov, N. P. *Current Opinion in Drug Discovery and Development* **2004,** *7,* 347-353.

[6] Lamzin, V. S.; Perrakis, A. *Nat Struct Biol* **2000,** *7,* 978-981.

[7] Perrakis, A.; Morris, R.; Lamzin, V. S. *Nat Struct Biol* **1999,** *6,* 458-463.

[8] Brooijmans, N.; Kuntz, I. *Annu Rev Biophys Biomol Struct* **2003,** *32,* 335-373.

[9] Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. *J Comput Aid Mol Des* **2002,** *16,* 151-166.

[10] Shoichet, B. K.; Mcgovern, S. L.; Wei, B. Q.; Irwin, J. J. *Curr Opin Chem Biol* **2002,** *6,* 439-446.

[11] Krovat, E. M.; Steindl, T.; Langer, T. *Current Computer-Aided Drug Design* **2005,** *1,* 93-102.

[12] Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003,** *52,* 609-623.

[13] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. *J Comput Chem* **1998,** *19,* 1639-1662.

[14] Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. *J Mol Biol* **1996,** *261,* 470-489.

[15] Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. *J Comput Aid Mol Des* **2001,** *15,* 411-428.

[16] Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. *J Med Chem* **2004,** *47,* 1739-1749.

[17] Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. *J Comput Chem* **2003,** *24,* 1637-1656.

[18] Gohlke, H.; Klebe, G. *Curr Opin Struc Biol* **2001,** *11,* 231-235.

[19] Murray, C. W.; Auton, T. R.; Eldridge, M. D. *J Comput Aid Mol Des* **1998,** *12,* 503-519.

[20] Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. *J Comput Aid Mol Des* **1997,** *11,* 425-445.

[21] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J Mol Biol* **1997,** *267,* 727-748.

[22] Mooij, W. T. M.; Verdonk, M. L. *Proteins-structure function and bioinformatics* **2005,** *61,* 272-287.

[23] Shoichet, B. K.; Leach, A. R.; Kuntz, I. D. *Proteins* **1999,** *34,* 4-16.

[24] Gohlke, H.; Hendlich, M.; Klebe, G. *J Mol Biol* **2000,** *295,* 337-356.

[25] Muegge, I.; Martin, Y. C. *J Med Chem* **1999,** *42,* 791-804.

[26] Schaffer, L.; Verkhivker, G. M. *Proteins* **1998,** *33,* 295-310.

[27] Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. *Proteins-structure function and bioinformatics* **2005,** *60,* 325-332.

[28] Pearlman, D. A.; Charifson, P. S. *J. Med. Chem.* **2001,** *44,* 3417-3423.

[29] Gibbs, J. W. *Elementary principles in statistical Mechanics;* Dover Publications inc.: New York, USA, 1902.

[30] McQuarrie, D. A. *Statistical Mechanics;* Harper and Row: New York, USA, 1976.

[31] Cramer, C. J. *Essentials of Computational Chemistry;* Wiley: West Sussex, UK, 2004.

[32] Kuo, I. F. W.; Mundy, C. J.; Mcgrath, M. J.; Siepmann, J. I.; Vandevondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J Phys Chem B* **2004,** *108,* 12990-12998.

[33] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996,** *118,* 11225-11236.

[34] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995,** *117,* 5179-5197.

[35] Lii, J. H.; Allinger, N. L. *J. Comput. Chem.* **1991,** *12,* 186-199.

[36] Mackerell, A. D. *et al. J. Phys. Chem. B* **1998,** *102,* 3586-3616.

[37] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids;* Oxford University Press: Oxford, UK, 2001.

[38] Leach, A. R. *Molecular Modelling, Principles and Applications;* Longman: Harlow, UK, 1996.

[39] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953,** *21,* 1087-1092.

[40] Owicki, J. C.; Scheraga, H. A. *Chem. Phys. Lett.* **1977,** *47,* 600-602.

[41] Resat, H.; Mezei, M. *J. Am. Chem. Soc.* **1994,** *116*, 7451-7452.

[42] Pangali, C.; Rao, M.; Berne, B. J. *Chem. Phys. Lett.* **1978,** *55*, 413-417.

[43] Leontidis, E.; Forrest, B. M.; Widmann, A. H.; Suter, U. W. *J. Chem. Soc. Faraday T.* **1995,** *91*, 2355-2368.

[44] Hansmann, U. H. E. *Chem. Phys. Lett.* **1997,** *281*, 140-150.

[45] Okabe, T.; Kawata, M.; Okamoto, Y.; Mikami, M. *Chem. Phys. Lett.* **2001,** *335*, 435-439.

[46] Bedrov, D.; Smith, G. D. *J. Chem. Phys.* **2001,** *115*, 1121-1124.

[47] Voter, A. F. *Phys.Rev. B* **1986,** *34*, 6819-6829.

[48] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* **1977,** *23*, 327-341.

[49] Zwanzig, R. W. *J. Chem. Phys.* **1954,** *22*, 1420-1426.

[50] Mezei, M. *J. Chem. Phys.* **1987,** *86*, 7084-7088.

[51] Essex, J. W.; Severance, D. L.; Tiradorives, J.; Jorgensen, W. L. *J Phys Chem B* **1997,** *101*, 9663-9669.

[52] Lamb, M. L.; Jorgensen, W. L. *J Med Chem* **1998,** *41*, 3928-3939.

[53] Udier-blagovic, M.; Tirado-rives, J.; Jorgensen, W. L. *J Med Chem* **2004,** *47*, 2389-2392.

[54] Guimaraes, C. R. W.; Boger, D. L.; Jorgensen, W. L. *J Am Chem Soc* **2005,** *127*, 17377-17384.

[55] McCarrick, M. A.; Kollman, P. A. *J Comput Aid Mol Des* **1999,** *13*, 109-121.

[56] Fox, T.; Scanlan, T. S.; Kollman, P. A. *J Am Chem Soc* **1997,** *119*, 11571-11577.

[57] Miyamoto, S.; Kollman, P. A. *Proteins-structure function and bioinformatics* **1993,** *16*, 226-245.

[58] Price, M. L. P.; Jorgensen, W. L. *J Am Chem Soc* **2000,** *122,* 9455-9466.

[59] Woods, C. J.; Essex, J. W.; King, M. A. *J Phys Chem B* **2003,** *107,* 13703-13710.

[60] Woods, C. J.; Essex, J. W.; King, M. A. *J Phys Chem B* **2003,** *107,* 13711-13718.

[61] Schafer, H.; van Gunsteren, W. F.; Mark, A. E. *J Comput Chem* **1999,** *20,* 1604-1617.

[62] Liu, H. Y.; Mark, A. E.; van Gunsteren, W. F. *J Phys Chem-us* **1996,** *100,* 9485-9494.

[63] Oostenbrink, C.; van Gunsteren, W. F. *Proteins* **2004,** *54,* 237-246.

[64] Oostenbrink, C.; van Gunsteren, W. F. *Proceedings of the national acedemy of sciences of the united states of america* **2005,** *102,* 6750-6754.

[65] Oostenbrink, C.; van Gunsteren, W. F. *J Comput Chem* **2003,** *24,* 1730-1739.

[66] Pitera, J. W.; van Gunsteren, W. F. *J Phys Chem B* **2001,** *105,* 11264-11274.

[67] Hu, H.; Yun, R. H.; Hermans, J. *Mol. Simulat.* **2002,** *28,* 67-80.

[68] Jarzynski, C. *Phys. Rev. E* **1997,** *56,* 5018-5035.

[69] Jarzynski, C. *Phys. Rev. Lett.* **1997,** *78,* 2690-2693.

[70] Hendrix, D. A.; Jarzynski, C. *J. Chem. Phys.* **2001,** *114,* 5974-5981.

[71] Hummer, G. *J. Chem. Phys.* **2001,** *114,* 7330-7337.

[72] Hummer, G. *Mol. Simulat.* **2002,** *28,* 81-90.

[73] Liphardt, J.; Dumont, S.; Smith, S. B.; Tinoco, I.; Bustamante, C.

[74] Zuckerman, D. M.; Woolf, T. B.

[75] Wu, D.; Kofke, D. A. *J Chem Phys* **2005,** *122,* art.

[76] Åqvist, J.; Medina, C.; Samuelsson, J. E. *Protein Eng.* **1994,** *7,* 385-391.

[77] Åqvist, J.; Hansson, T. *J. Phys. Chem.* **1996,** *100*, 9512-9521.

[78] Wall, I. D.; Leach, A. R.; Salt, D. W.; Ford, M. G.; Essex, J. W. *J. Med. Chem.* **1999,** *42*, 5142-5152.

[79] Rizzo, R. C.; Udier-Blagovic, M.; Wang, D.; Watkins, E. K.; Kroeger Smith, M. B.; Smith Jr., R. H.; Tirado-Rives, J.; Jorgensen, W. L. *J. Med. Chem.* **2002,** *45*, 2970-2987.

[80] Tominaga, Y.; Jorgensen, W. L. *J Med Chem* **2004,** *47*, 2534-2549.

[81] Wesolowski, S. S.; Jorgensen, W. L. *Bioorg Med Chem Lett* **2002,** *12*, 267-270.

[82] Ostrovsky, D.; Udier-blagovic, M.; Jorgensen, W. L. *J Med Chem* **2003,** *46*, 5691-5699.

[83] Zhou, R. H.; Friesner, R. A.; Ghosh, A.; Rizzo, R. C.; Jorgensen, W. L.; Levy, R. M. *J Phys Chem B* **2001,** *105*, 10388-10397.

[84] Massova, I.; Kollman, P. A. *Perspect Drug Discov* **2000,** *18*, 113-135.

[85] Chong, L. T.; Duan, Y.; Wang, L.; Massova, I.; Kollman, P. A. *Proc. Natl. Acad. Sci.* **1999,** *96*, 14330-14335.

[86] Reyes, C. M.; Kollman, P. A. *J. Mol. Biol.* **2000,** *295*, 1-6.

[87] Gouda, H.; Kuntz, I. D.; Case, D. A.; Kollman, P. A. *Biopolymers* **2003,** *68*, 16-34.

[88] Wang, J. M.; Morin, P.; Wang, W.; Kollman, P. A. *J. Am. Chem. Soc.* **2001,** *123*, 5221-5230.

[89] Kuhn, B.; Gerber, P.; Schulz-gasch, T.; Stahl, M. *J Med Chem* **2005,** *48*, 4040-4048.

[90] Fogolari, F.; Moroni, E.; Wojciechowski, M.; Baginski, M.; Ragona, L.; Molinari, H. *Proteins* **2005,** *59*, 91-103.

[91] Pearlman, D. A. *J Med Chem* **2005,** *48*, 7796-7807.

[92] Woo, H. J.; Roux, B. *Proceedings of the national academy of sciences of the united states of america* **2005,** *102,* 6825-6830.

[93] Tomasi, J.; Persico, M. *Chem Rev* **1994,** *94,* 2027-2094.

[94] Born, M. *Z. Phys.* **1920,** *1*, 45.

[95] Jackson, J. *Classical Electrodynamics;* Wiley: West Sussex, UK, 1998.

[96] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J Am Chem Soc* **1990,** *112,* 6127-6129.

[97] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem Phys Lett* **1995,** *246,* 122-129.

[98] Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. *J Phys Chem A* **1997,** *101,* 3005-3014.

[99] Bashford, D.; Case, D. A. *Annu Rev Phys Chem* **2000,** *51,* 129-152.

[100] Michel, J.; Taylor, R. D.; Essex, J. W. *J Comput Chem* **2004,** *25,* 1760-1770.

[101] Pitera, J. W.; van Gunsteren, W. F. *J Am Chem Soc* **2001,** *123,* 3163-3164.

[102] Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J Phys Chem B* **2000,** *104,* 6271-6285.

[103] Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J Comput Chem* **2002,** *23,* 517-529.

[104] Gallicchio, E.; Ley, R. M. *J Comput Chem* **2004,** *25,* 479-499.

[105] Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. *Proteins-structure function and bioinformatics* **2004,** *56,* 310-321.

[106] Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J Chem Phys* **2003,** *119,* 5740-5761.

[107] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A.

[108] Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J Comput Chem* **2000,** *21*, 132-146.

[109] Jakalian, A.; Jack, D. B.; Bayly, C. I. *J Comput Chem* **2002,** *23*, 1623-1641.

[110] Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J Am Chem Soc* **1985,** *107*, 3902-3909.

[111] Mulliken, R. S. *J. Chem. Phys.* **1955,** *23*, 1833-1840.

[112] Case, D. A. *et al.* "AMBER 8", University of California, 2004.

[113] Reddy, M. R.; Erion, M. D.; Agarwal, A.; Viswanadhan, V. N.; Mcdonald, D. Q.; Still, W. C. *J Comput Chem* **1998,** *19*, 769-780.

[114] Jayaram, B.; Sprous, D.; Beveridge, D. L. *J Phys Chem B* **1998,** *102*, 9571-9576.

[115] Zhu, T. H.; Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J Chem Phys* **1998,** *109*, 9117-9133.

[116] Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J Comput Chem* **2004,** *25*, 265-284.

[117] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J Phys Chem-us* **1996,** *100*, 19824-19839.

[118] Zhang, W.; Hou, T.; Qiao, X.; Xu, X. *J. Phys. Chem B* **2003,** *107*, 9071-9078.

[119] Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning;* Addison-Wesley: Boston, USA, 1989.

[120] Wall, M. "GALIB", http://lancet.mit.edu/ga/.

[121] Rankin, N. K.; Traian, S.; Enrico, O. P. *J Comput Chem* **2003,** *24*, 954-962.

[122] Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proceedings of the National Academy of Sciences of the United S states of America* **2001,** *98*, 10037-10041.

[123] Onufriev, A.; Case, D. A.; Bashford, D. *J Comput Chem* **2002,** *23*, 1297-1304.

[124] Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tiradorives, J. *J Chem Phys* **1988,** *89*, 3742-3746.

[125] Jorgensen, W. L.; Severance, D. L. *J Am Chem Soc* **1990,** *112*, 4768-4774.

[126] Jorgensen, W. L. *J Am Chem Soc* **1989,** *111*, 3770-3771.

[127] Masunov, A.; Lazaridis, T. *J Am Chem Soc* **2003,** *125*, 1722-1730.

[128] Soetens, J. C.; Millot, C.; Chipot, C.; Jansen, G.; Angyan, J. G.; Maigret, B. *J Phys Chem B* **1997,** *101*, 10910-10917.

[129] Rozanska, X.; Chipot, C. *J Chem Phys* **2000,** *112*, 9691-9694.

[130] Kollman, P. *Chem Rev* **1993,** *93*, 2395-2417.

[131] Jorgensen, W. L. "MCPRO 1.5", 1996.

[132] Fukunishi, Y.; Suzuki, M. *J Phys Chem-us* **1996,** *100*, 5634-5636.

[133] Dominy, B. N.; Brooks, C. L. *J Phys Chem B* **1999,** *103*, 3765-3773.

[134] Lazaridis, T.; Karplus, M. *Proteins* **1999,** *35*, 133-152.

[135] Beglov, D.; Roux, B. *J Chem Phys* **1994,** *100*, 9050-9063.

[136] Tucker, E. E.; Christian, S. D. *J Phys Chem* **1979,** *83*, 426-427.

[137] Graziano, G.; Lee, B. *J Phys Chem B* **2001,** *105*, 10367-10372.

[138] Jorgensen, W. L.; Madura, J. D. *Mol Phys* **1985,** *56*, 1381.

[139] Gilson, M. K.; Honig, B. *J Comput Aid Mol Des* **1991,** *5*, 5-20.

[140] Onufriev, A.; Bashford, D.; Case, D. A. *Proteins-structure function and bioinformatics* **2004,** *55*, 383-394.

[141] Onufriev, A.; Bashford, D.; Case, D. A. *J Phys Chem B* **2000,** *104*, 3712-3720.

[142] Sorin, E. J.; Engelhardt, M. A.; Herschlag, D.; Pande, V. S. *J Mol Biol* **2002,** *317,* 493-506.

[143] Sorin, E. J.; Rhee, Y. M.; Nakatani, B. J.; Pande, V. S. *Biophys J* **2003,** *85,* 790-803.

[144] Ulmschneider, J. P.; Jorgensen, W. L. *J Chem Phys* **2003,** *118,* 4261-4271.

[145] Ulmschneider, J. P.; Jorgensen, W. L. *J Am Chem Soc* **2004,** *126,* 1849-1857.

[146] Siepmann, J. I.; Frenkel, D. *Mol Phys* **1992,** *75,* 59-70.

[147] Woods, C. J.; Michel, J. "ProtoMS2.1", in house Monte Carlo Code, 2005.

[148] Shrake, A.; Rupley, J. A. *J Mol Biol* **1973,** *79,* 351-371.

[149] Taylor, N. R.; Cleasby, A.; Singh, O.; Skarzynski, T.; Wonacott, A. J.; Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Bethell, R.; Colman, P.; Varghese, J. *J Med Chem* **1998,** *41,* 798-807.

[150] Kurumbail, R. G.; Stevens, A. M.; Gierse, J. K.; Mcdonald, J. J.; Stegeman, R. A.; Pak, J. Y.; Gildehaus, D.; Miyashiro, J. M.; Penning, T. D.; Seibert, K.; Isakson, P. C.; Stallings, W. C. *Nature* **1996,** *384,* 644-648.

[151] Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J Mol Biol* **1999,** *285,* 1735-1747.

[152] Gelb, L. D. *J Chem Phys* **2003,** *118,* 7747-7750.

[153] Hetenyi, B.; Bernacki, K.; Berne, B. J. *J Chem Phys* **2002,** *117,* 8203-8207.

[154] Bernacki, K.; Hetenyi, B.; Berne, B. J. *J Chem Phys* **2004,** *121,* 44-50.

[155] Iftimie, R.; Salahub, D.; Wei, D. Q.; Schofield, J. *J Chem Phys* **2000,** *113,* 4852-4862.

[156] Zhu, J. A.; Shi, Y. Y.; Liu, H. Y. *J Phys Chem B* **2002,** *106,* 4844-4853.

[157] Weiser, J.; Shenkin, P. S.; Still, W. C. *J Comput Chem* **1999,** *20,* 217-230.

[158] Xie, W. L.; Chipman, J. G.; Robertson, D. L.; Erikson, R. L.; Simmons, D. L. *Proceedings of the national academy of sciencs of the united states of america* **1991,** *88,* 2692-2696.

[159] Penning, T. D. *et al. J Med Chem* **1997,** *40,* 1347-1365.

[160] Marnett, L. J.; Kalgutkar, A. S. *Curr Opin Chem Biol* **1998,** *2,* 482-490.

[161] Cheng, Y.; Prusoff, W. H. *Biochem Pharmacol* **1973,** *22,* 3099-3108.

[162] Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. *J Comput Aid Mol Des* **1995,** *9,* 87-110.

[163] Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J Chem Theory Comput* **2006,** *2,* 128-139.

[164] Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J Phys Chem B* **2005,** *109,* 14769-14772.

[165] Liu, H. Y.; Kuntz, I. D.; Zou, X. Q. *J Phys Chem B* **2004,** *108,* 5453-5462.

[166] Wikipedia, "Influenza — Wikipedia, The Free Encyclopedia", http://en.wikipedia.org/wiki/Influenza.

[167] Barillari, C. *The Role of Water in Protein-Ligand Interactions: Implications for Rational Drug Design;* PhD Thesis, University of Southampton: Southampton, United Kindgom, 2006.

[168] Zou, X. Q.; Sun, Y. X.; Kuntz, I. D. *J Am Chem Soc* **1999,** *121,* 8033-8043.

[169] Boresch, S.; Karplus, M. *J Phys Chem A* **1999,** *103,* 103-118.

[170] Boresch, S.; Karplus, M. *J Phys Chem A* **1999,** *103,* 119-136.

[171] Beutler, T. C.; Mark, A. E.; Vanschaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. *Chem Phys Lett* **1994,** *222,* 529-539.

[172] Zacharias, M.; Straatsma, T. P.; McCammon, J. A. *J Chem Phys* **1994,** *100,* 9025-9031.

[173] Straatsma, T. P.; Zacharias, M.; McCammon, J. A. *Chem Phys Lett* **1992,** *196*, 297-302.

[174] Pitera, J. W.; van Gunsteren, W. F. *Mol Simulat* **2002,** *28*, 45-65.

[175] Daura, X.; Hunenberger, P. H.; Mark, A. E.; Querol, E.; Aviles, F. X.; Vangunsteren, W. F. *J Am Chem Soc* **1996,** *118*, 6285-6294.

[176] Anwar, J.; Heyes, D. M. *J Chem Phys* **2005,** *122*, art.

[177] Jorgensen, W. L.; Ravimohan, C. *J Chem Phys* **1985,** *83*, 3050-3054.

[178] Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J Phys Chem B* **2003,** *107*, 9535-9551.

[179] Laneuville, O.; Breuer, D. K.; Dewitt, D. L.; Hla, T.; Funk, C. D.; Smith, W. L. *J Pharmacol Exp Ther* **1994,** *271*, 927-934.

[180] Futaki, N.; Takahashi, S.; Yokoyama, M.; Arai, I.; Higuchi, S.; Otomo, S. *Prostagl* **1994,** *47*, 55-59.

[181] Warner, T. D. “Determinations of COX-1/COX-2 Selectivity: Relationships to Clinical Effects”, http://www.toxforum.org/winter2001/.

[182] Rowlinson, S. W.; Kiefer, J. R.; Prusakiewicz, J. J.; Pawlitz, J. L.; Kozak, K. R.; Kalgutkar, A. S.; Stallings, W. C.; Kurumbail, R. G.; Marnett, L. J. *J Biol Chem* **2003,** *278*, 45763-45769.

[183] Michel, J.; Taylor, R. D.; Essex, J. W. *J Chem Theory Comput* in press.

[184] Khandelwal, A.; Lukacova, V.; Comez, D.; Kroll, D. M.; Raha, S.; Balaz, S. *J Med Chem* **2005,** *48*, 5437-5447.

[185] Grater, F.; Schwarzl, S. M.; Dejaegere, A.; Fischer, S.; Smith, J. C. *J Phys Chem B* **2005,** *109*, 10474-10483.

[186] Rodinger, T.; Howell, P. L.; Pomes, R. *J Chem Phys* **2005,** *123*,.

[187] Borjesson, U.; Hunenberger, P. H. *J Chem Phys* **2001,** *114*, 9706-9719.

[188] Burgi, P. A.; Kollman, W. *Proteins* **2002,** *47*, 469-480.

[189] Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997,** *27*, 523-544.

[190] Evans, A.; Swartz, T. *Approximating Integrals via Monte Carlo and Deterministic Methods;* Oxford University Press: Oxford, UK, 2000.

[191] Gammerman, D. *Markov Chain Monte Carlo;* Chapman and Hall: London, UK, 1997.

[192] Henry, M. *Nonnegative Matrices;* John Wiley and Sons: New York, USA, 1988.