UNIVERSITY OF SOUTHAMPTON

Development and Application of a QM/MM Method for Free Energy Calculations

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

Michael Keith Carter

School of Chemistry
University of Southampton

May 2014

Supervisor: Prof. Jonathan Essex

Industrial Supervisors: Anny Odile-Colson / Peter Hunt

Advisor: Chris Skylaris

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF NATURAL AND ENVIRONMENTAL SCIENCES

SCHOOL OF CHEMISTRY

Doctor of Philosophy

DEVELOPMENT AND APPLICATION OF A QM/MM METHOD FOR FREE ENERGY CALCULATIONS

MICHAEL CARTER

In this thesis, a simplified Quantum Mechanics/Molecular Mechanics (QM/MM) method has been developed, which incorporates Density Functional Theory (DFT) and Free Energy Perturbation (FEP) to calculate QM/MM corrections for classically obtained MM free energies. This method has been applied to the calculation of hydration free energies and protein-ligand binding free energies. The hydration free energy study showed that for small organic compounds the QM/MM method could perform as well as standard MM. Further analysis highlighted that implementing QM/MM appeared to over-polarise for compounds with hydrogen bonding moieties. This over-polarisation was caused by the embedding technique utilised. Hence, the embedding strategy was adapted to utilize a Gaussian blurring approach, which enabled the minimisation impact of this over-polarisation. For the protein-ligand studies three proteins were investigated; COX-2, neuraminidase and CDK2. The results from these studies showed that the QM/MM method obtain less accurate results than conventional MM. This has been attributed to several factors, including; too simplistic embedding techniques leading to over polarisation for extremely polar protein-ligand systems (neuraminidase) and poor agreement for protein-ligand systems with small pocket sizes. Also, GCMC simulations have identified that erroneous system setup for

CDK2 is the root cause of extremely poor correlation in both MM and QM/MM free energy studies.

DECLARATION OF AUTHORSHIP

I, MICHAEL KEITH CARTER, declare that this thesis

"DEVELOPMENT AND APPLICATION OF A QM/MM METHOD FOR FREE ENERGY CALCULATIONS"

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

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

Signed: ……..

Date: ……

Acknowledgements

Firstly I would like to thank my supervisor, Jon, for his support and guidance throughout my PhD. I will never forget some of the more humorous conversations had during my time in Southampton. In addition, I would like to thank Jon for helping me through a tough final year of my PhD, without which many of the ideas presented in this work would not have been possible.

I would also like to thank my industrial supervisors Anny Odile-Colson and Peter Hunt for their advice throughout my PhD, and to Novartis and the ESPRC for funding my research.

Many of the ideas and implementations in this thesis have come from discussions with various people, some of which I would like to thank here. Thanks to Dr Patrick Schöpf and Dr Frank Beierlein for their help and advice in implementing the QM/MM method, in particular for the hydration free energy study. I would also like to thank Dr Julien Michel for providing me with input data for the protein-ligand studies, which make up a significant proportion of this thesis. I would also like to thank Peter Guthrie for providing me with experimental data sets, which gave new insights into several results shown in this thesis. And finally Paul Sherwood, for helping me implement a 'Gaussian Blur' approach to charge embedding within QM/MM simulations.

I would also like to express my gratitude to several members of the Essex research group. In particular I would like to thank Michael Bodnarchuk and Samuel Genheden for taking the time to listen to my ideas and for giving me useful feedback when helping me analyse my results. I would also like to thank Zohra Ouaray, Barbara Sander and Donna Goreham for the endless conversations had during our smoking breaks.

I would also like to thank my housemates Alvaro Ruiz-Serrano, Samuel Golten and Yannis Haldoupis, who were not only great housemates, but amazing friends as well.

Finally I would like to thank my family, my mother Sharon, my father Keith and my sisters Michelle and Christine, who have supported me in everything I have ever done. Your love and encouragement throughout the years has led to this thesis, and I dedicate this thesis to you all.

Contents

1 Introduction

Understanding how drug molecules bind to target biomolecules is an integral part of modern drug design. The number of protein structures available via the Protein DataBank (PDB) has grown exponentially over the past 20 years with over 90,000 structures available today [1]. Allied to the growth in protein structure availability has been the growth in computing power. These changes have seen atomistic computer simulations of biomolecular systems become increasingly popular with pharmaceutical companies and academic associations [2].

1.1 Free Energy Calculations

Many methods and algorithms have been developed to simulate biomolecular systems and obtain relevant thermodynamic properties (i.e. binding free energies). The accurate calculation of protein-ligand binding is still a major challenge in modern computational drug design. The prediction of such binding free energies is not only important for the design of new therapeutic compounds, but is also a valuable tool in other areas of research such as the role of catalysts in reaction mechanisms [3, 4] and the action of molecules on inorganic surfaces [5]. More approximate methods like docking [6] allow for the prediction of how a drug molecule binds in a target biomolecules binding site, but such methods do not allow for the accurate prediction of binding free energies, this problem is commonly referred to as the 'docking problem' [7]. Thermodynamically rigorous methods, including Thermodynamic Integration (TI) [8] and Free Energy Perturbation (FEP) [9] offer much greater accuracy when calculating binding free energies, at a much higher computational cost. These

methods exploit free energy cycles to calculate the relative binding free energies of drug molecules via alchemical perturbations between different drug compounds. With such improved methods and computing resources, many examples of lead optimization guided by free energy calculations have been reported [10, 11, 12], suggesting that these methodologies are becoming commonplace within the drug design process.

1.2 Limitations of Free Energy Simulations in drug design

Most current free energy methods use classical MM forcefield Hamiltonians, such as AMBER99 [13], Generalised Amber ForceField (GAFF) [14], Optimised Parameters for Liquid Simulations (OPLS) [15] and GROMOS [16]. Although classical forcefields provide a relatively accurate description of protein-ligand interactions they often neglect the polarization of a ligand by its environment (protein, solvent, cofactors and ions). This can have a large effect when considering complex electronic environments, such as protein-ligand systems. One possible approach to correct this is to include polarization terms within the forcefields to create polarizable forcefields [17], however, work in this area is very limited and greater development is needed before this is can be considered a viable solution. Another common approach is to use exaggerated point charges to overestimate dipole moments within a ligand and thus incorporate solvation effects implicitly [18].

A more general approach is to use combined classical non-polarizable forcefields, and to correct the free energies obtained using QM or QM/MM approaches. Many studies have used a similar approach for enzymatic reaction mechanisms [19, 20] and free energy studies. [21, 22, 23]. In this thesis we shall

describe the use of a simplified MM→QM/MM method which incorporates Density Functional Theory (DFT) and FEP to obtain QM/MM corrections for classically (MM) obtained free energies of hydration and protein-ligand binding free energies.

2 Computational Methods

2.1 Introduction

With ever increasing computational power, it is now possible to simulate complex processes including protein folding [24], membrane formation [25] and cell membrane permeation [26], simulations which 10 years ago would not have been possible. The following sections outline the basic principles behind molecular modelling. Owing to the huge numbers of studies and methods available in the literature, it is impossible to fully, or even partially, explain all of the methods in this vast and ever expanding field. As such, there are excellent introductory sources which go further into the finer point of these methods, such as Molecular Modelling: Principles and Applications by Leach [27] and Computer Simulation of Liquids, by Allen and Tildesley [28]. In the following sections we shall focus on aspects of molecular modelling relevant to the work presented here.

2.2 Statistical Mechanics

Statistical mechanics links thermodynamic observables of large-scale systems to the microscopic properties of the constituent particles of the system. Statistical mechanics, using the laws derived by Boltzmann, allow for the precise calculation of entropy based upon the number and population of microstates in the system.

2.2.1 Boltzmann Distribution

The following derivation is taken from pages 510-511 of Physical Chemistry by Atkins [29]. The Boltzmann distribution is fundamental to statistical mechanics. It describes the population of energy states within a system. Consider a system containing *N* particles where each particle can exist in a given energy state, ε_0, ε_1, where ε_0 is the lowest energy state. The instantaneous configuration of the system fluctuates with time, yet some configurations are more likely than others. Critically, the total energy of each configuration must be the same. The likelihood of any one of these states can be expressed numerically as the configurational weight of the configuration (eq. 2.1):

$$W = {}^{N!}/_{n_0! n_1! n_2! \ldots} \tag{2.1}$$

Where n_0 is the population of energy state ε_0 etc…. A question which arises is whether there is a dominant configuration which in turn controls the observed system properties. To go about answering this question, two constraints must be applied:

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

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

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

In order to obtain the most likely configuration, we must differentiate W with respect to all of the populations in the system subject to the constraints. This achieved using Langrange multipliers. This gives the condition for the most probable configuration as:

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

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

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

Combining equation 2.5 with equation 2.1 gives:

$$\ln W = \ln {}^{N!}/_{n_0! \, n_1! \, n_2! \ldots} \tag{2.6}$$

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

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

We can now estimate a solution to equation 2.4:

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

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

Putting this result into equation 2.4, we obtain:

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

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

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

Using constraint 2.3, we obtain:

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

Rearranging equations 2.12 and 2.13, produces the Boltzmann distribution:

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

2.2.2 The Molecular Partition function

The probability distribution, π, which is the probability of a component of the system having and energy, ε_i, of the canonical ensemble is related to equation 2.14 and the Boltzmann equation:

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

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

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

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

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

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

In equations 2.17 and 2.18, k_b is the Boltzmann constant, equal to 1.38 x10^{-23} J.K^{-1}. T is the temperature of the system and V the volume of the system.

In the limit of a quantised energy states, equation 2.16 can be used to define the partition function. However, when dealing with a continuum of states, equation 2.16 is replaced by an integral which considers the 6 dimensional space, known as phase space.

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

In equation 2.19, $\boldsymbol{p}^n$ and $\boldsymbol{r}^n$ are the momenta and positions of each of the N particles. h is Planck's constant, equal to 6.63 x 10^{-34} J.s, and the *N*! term arises from the fact that all particles are indistinguishable.

The partition function can be adapted to introduce the notion of interacting particles through the concept of ensembles. An ensemble can be thought of as a large collection of replica systems, where each replica in the ensemble could be representative of the true state of the system. An example of an ensemble is the canonical ensemble. This ensemble can be thought of as a collection of identical systems which are in thermal contact with each other, allowing the exchange of energy between each system but keeping the temperature, volume, and number of particles in each system constant.

2.3 Molecular Mechanics - Empirical Forcefields

From analysis of the partition function, it can be seen that the energy of the system of the system must be calculated. This is commonly achieved from splitting the energy contributions into two parts; the kinetic and potential energy. The kinetic energy of the

system can be found analytically from the masses and velocities of the particles, using equation 2.20.

$$E_k = \frac{1}{2} m \boldsymbol{v}^2 \tag{2.20}$$

In equation 2.20, m is the mass of the particle and $\boldsymbol{v}$ the velocity of the particle. The potential energy of the system cannot be found this way. There are two commonly used techniques to calculate the potential energy of the system; classically via Molecular Mechanics (MM) and though the use of Quantum Mechanics (QM) (section 2.4). MM typically finds the potential energy of the system as a function of the co-ordinates of the system.

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

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

E_{bond} and E_{angle} are typically represented by harmonic potentials:

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

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

In equations 2.22 and 2.23, k is the force constant, l the bond length, θ is the angle and l_{eq} and θ_{eq} are the equilibrium bond length and angle respectively.

The dihedral energy is modelled as a cosine series:

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

In equation 2.24, n is the multiplicity of the function (the number of minima as the bond is rotated through 360°). δ is the phase angle, which determines the point where the dihedral energy is at its lowest value. ϕ is the dihedral angle, whilst k_n is the amplitude of the cosine function and represents the force constant.

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

$$E_{inter} = \sum_{i} \sum_{j>i} \left\{ \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \right\} \tag{2.25}$$

In equation 2.25, i and j represent intermolecular atom pairs, with q_i and q_j atomic partial charges on atoms i and j. ε_{ij} and σ_{ij} are the Lennard-Jones well depth and collision diameter for atoms i and j, with r_{ij} the inter-atomic distance.

The parameters used within forcefields have been derived differently, meaning that the combination of different forcefields is not advised. For example, the GAFF obtains it parameters through a combination of empirical and QM means [30]. Partial charges are typically obtained using high lever quantum theory such as Hartree-Fock calculations, whilst bond lengths are typically obtained by a combination of

experimental methods (X-ray crystallography) and high level *ab-intio* calculations. A similar procedure is used to obtain the parameters for the angle bending and torsional terms. In comparison, the GROMOS forcefield is purely empirical, with non-bonded terms parameterised to fit experimental properties such as the free enthalpy of hydration. Bonded terms are parameterised purely against crystallographic and spectroscopic data [31].

2.3.1 Polarizable Forcefields

Polarizable forcefields aim to capture the changes in polarisation that classical fixed-charge MM forcefields neglect. Due to the complex nature of polarisation there are several methods available to include it within a forcefield.

For example, one of the first polarizable forcefields ENZYMIX [32] was based upon an induced dipole model. Other efforts to include polarization have focused on point charge models, where classical MM forcefields, including CHARMM [33] and AMBER [13], use exaggerated charges to include polarisation effects. More recent advances have seen polarizable potentials based upon distributed multipoles become very popular. Several potentials, AMEOBA [34], SIBFA [35] and ORIENT [36], have been designed to utilise this multipole based approach. Despite these recent advances in polarisable MM forcefields, their application to drug design has been limited due to the success of classical MM forcefields at describing biomolecular systems. However, given the extensive development of polarizable forcefields in the past few years, it is believed that polarizable potentials will become routinely used within computational simulations, particularly on systems where standard MM forcefields fail [17].

2.4 Quantum Mechanics

Quantum Mechanics (QM) differs from MM as it uses wave functions (Ψ) to describe the behaviour of electrons. QM methods are generally much more accurate than MM methods as they explicitly consider electrons. This means that QM methods can be used to study very polar systems with a high degree of accuracy. The pitfalls of QM lie in its computational expense. The following sections shall briefly describe the background of QM for a more detailed description into QM the textbooks "Introduction to Advanced Electronic Structure Theory" by Szabo and Ostlund [37], "Molecular Electronic Theory" by Helgaker, Jorgensen and Olse [38], and "Molecular Quantum Mechanics" by Atkins and Friedman [39] are recommended.

2.4.1 Schrödinger Equation

The main aim of QM methods applied to simple chemical problems0 is to find an approximate solution to the non-relativistic time-independent Schrödinger Equation, equation 2.26.

$$\mathcal{H}|\Phi\rangle = \mathcal{E}|\Phi\rangle \tag{2.26}$$

In equation 2.26, $\mathcal{H}$ is the Hamiltonian operator for a system of nuclei and electrons described by position vectors $\boldsymbol{R}_A$ and $\boldsymbol{r}_i$, respectively. $\mathcal{E}$ is the energy of the system at state Φ. For a molecular coordinate system containing two nuclei (A and B) and two electrons (i and j), the distance between the ith electron and the Ath nucleus is $r_{iA} = |\boldsymbol{r}_{iA}| = |\boldsymbol{r}_i - \boldsymbol{R}_A|$; the distance between the ith electron and jth electron is $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$, and the distance between the Ath nucleus and the Bth nucleus is

$R_{AB} = |\boldsymbol{R}_A - \boldsymbol{R}_B|$. The Hamiltonian of a system containing *N* electrons and *M* Nuclei is given by:

$$\mathcal{H} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{1}{2M_A}\nabla_A^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_a}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{\boldsymbol{r}_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} \quad (2.27)$$

In equation 2.27, M_A is the ratio of the mass of nucleus A to the mass of an electron, and Z_A is the atomic number of nucleus A. The Laplacian operators ∇_i^2 and ∇_A^2 involve differentiation with respect to the coordinates of the ith electron and the Ath nucleus [37]. The first term in equation 2.27 is the operator for the kinetic energy of the electrons; the second term is the operator for the nuclear energy; the third term represents the coulomb attraction between the electrons and nuclei; the fourth and fifth terms describe the repulsion between electrons and between nuclei, respectively.

2.4.2 Born-Oppenheimer approximation

The Born-Oppenheimer (BO) approximation [40] is fundamental to quantum chemistry. The approximation focusses on distinguishing the nuclei and electrons. Since the nuclei in our system are much heavier than electrons, they move at a much slower rate. Therefore, to a good approximation, one may consider the electrons in a molecule to be moving in a field of fixed nuclei. Within this approximation, the second term in equation 2.27, the kinetic energy of the nuclei, can be neglected and the last term in equation 2.27, the repulsion between nuclei, can be considered a constant. The remaining terms in equation 2.27 are called the electronic Hamiltonian, equation 2.28:

$$\mathcal{H}_{elec} = -\sum_{i=1}^{N}\frac{1}{2}\nabla_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_a}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{\boldsymbol{r}_{ij}} \tag{2.28}$$

The following derivation is taken from Molecular Quantum Mechanics by from Atkins and Friedman [39]. The solution to a Schrödinger equation involving the electronic Hamiltonian is given by equation 2.29:

$$\mathcal{H}_{elec}\Phi_{elec} = \mathcal{E}_{elec}\Phi_{elec} \tag{2.29}$$

In equation 2.29, Φ_{elec} is the electronic wave function, which can be described as:

$$\Phi_{elec} = \Phi_{elec}(\{\boldsymbol{r}_i\};\ \{\boldsymbol{R}_A\}) \tag{2.30}$$

Equation 2.30 describes the motion of the electrons and explicitly depends on the electronic coordinates, $\boldsymbol{r}_i$, but depends parametrically on the nuclear coordinates, $\boldsymbol{R}_A$, as does the electronic energy $\mathcal{E}_{elec}$in equation 2.29:

$$\mathcal{E}_{elec} = \mathcal{E}_{elec}(\{\boldsymbol{r}_i\};\ \{\boldsymbol{R}_A\}) \tag{2.31}$$

A parametric dependence means that, for different arrangements of the nuclei; $\Phi_{elec}$is a different function of electronic coordinates. The nuclear coordinates do not appear explicitly in Φ_{elec}. The total energy of the system for fixed nuclei must also include the constant nuclear repulsion:

$$\mathcal{E}_{tot} = \mathcal{E}_{elec} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{R_{AB}} \tag{2.32}$$

Equations 2.28 and 2.32 constitute the electron problem; the solutions to this problem shall be described in the next sections.

2.4.3 Pauli Exclusion Principle

The electronic Hamiltonian, equation 2.28, depends only on the spatial coordinates of electrons. To fully describe an electron it is necessary, in addition, to specify its spin. This can be done in the context of non-relativistic theory by introducing two spin functions $\alpha(\omega)$ and $\beta(\omega)$, corresponding to spin up and down respectively. In this new formalisation an electron is described not only by the three spatial coordinates $\boldsymbol{r}$ but also by a spin coordinate ω. These collective four coordinates are described as follows:

$$\boldsymbol{x} = \{\boldsymbol{r}, \omega\} \tag{2.33}$$

The wave function for an *N*-electron system is then a function of $\boldsymbol{x}_1$, $\boldsymbol{x}_2$,…, $\boldsymbol{x}_n$. Which is typically written as $\Phi(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)$ [39].

A more satisfactory theory can be obtained, if the following additional requirement is placed upon the wave function : *"A many-electron wave function must be antisymmetric with respect to the interchange of the coordinate* $\boldsymbol{x}$ *(both space and spin) of any two electrons"* [37].

$$\Phi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_N) = -\Phi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_N) \quad (2.34)$$

The requirement stated by equation 2.34 is sometimes called the *anitsymmetry principle,* is a very general statement of the more familiar Pauli exclusion principle. It is an independent fact of QM. Thus the exact form of the wave function not only has to satisfy the Schrödinger equation, it must also be antisymmetric in the sense of equation 2.34.

2.4.4 Slater Determinants

For a two electron case in which both spin orbitals, χ_i and χ_j are occupied with one electron, with electron one in χ_i and electron two in χ_j, the following is obtained:

$$\Psi_{1\,2}^{HP}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \chi_i(\boldsymbol{x}_1)\chi_j(\boldsymbol{x}_2) \quad (2.35)$$

In equation 2.35, $\Psi_{1\,2}^{HP}$ is the many-electron wave function termed the Hartree product, with electron one being described by the spin orbital χ_i, and electron two by the spin orbital χ_j [37]. On the other hand, if electron one occupies χ_j and electron two occupies χ_i, the following is obtained:

$$\Psi_{2\,1}^{HP}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \chi_i(\boldsymbol{x}_2)\chi_j(\boldsymbol{x}_1) \quad (2.36)$$

Each of these Hartree products (equations 2.35 and 2.36) clearly distinguishes between electrons; however, the wave function obtained does not, and which satisfies

the requirement of the anitsymmetry principle by taking into account the appropriate linear combination of these two Hartree products:

$$\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = 2^{-1/2} \left(\chi_i(\boldsymbol{x}_1)\chi_j(\boldsymbol{x}_2) - \chi_i(\boldsymbol{x}_2)\chi_j(\boldsymbol{x}_1)\right) \tag{2.37}$$

The factor $2^{-1/2}$ is a normalisation factor. The minus sign ensures $\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is antisymmetric with respect to the interchange of the coordinates of electrons one and two [37].

The antisymmetric wave function of equation 2.37 can be written as a determinant:

$$\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = 2^{-1/2} \begin{vmatrix} \chi_i(\boldsymbol{x}_1) & \chi_j(\boldsymbol{x}_1) \\ \chi_i(\boldsymbol{x}_2) & \chi_j(\boldsymbol{x}_2) \end{vmatrix} \tag{2.38}$$

Equation 2.38 is called a Slater determinant. For a system of *N*-electrons the generalization of equation 2.38 is:

$$\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N) = (N!)^{-\frac{1}{2}} \begin{vmatrix} \chi_i(\boldsymbol{x}_1) & \chi_j(\boldsymbol{x}_1) & \dots & \chi_k(\boldsymbol{x}_1) \\ \chi_i(\boldsymbol{x}_2) & \chi_j(\boldsymbol{x}_2) & \dots & \chi_k(\boldsymbol{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_i(\boldsymbol{x}_N) & \chi_j(\boldsymbol{x}_N) & \dots & \chi_k(\boldsymbol{x}_N) \end{vmatrix} \tag{2.39}$$

In equation 2.39, $(N!)^{-\frac{1}{2}}$ is the normalisation factor. This Slater determinant has *N* electrons occupying *N* spin orbitals ($\chi_i, \chi_j, \dots, \chi_k$) without specifying which electron is in which orbital. The rows of an *N*-electron Slater determinant are labeled by

electrons: first row $\boldsymbol{x}_1$, second row $\boldsymbol{x}_2$, etc., and the columns are labeled by spin orbitals: first column χ_i, second column χ_j, etc. Interchanging the coordinates of two electrons corresponds to interchanging two rows of the Slater determinant, which changes the sign of the determinan [37]. Hence Slater determinants meet the requirement of the anitsymmetry principle. It is convenient to introduce a short-hand notation for a normalised Slater determinant, which includes the normalisation constant and only shows the diagonal elements of the determinant:

$$\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N) = |\chi_i(\boldsymbol{x}_1)\chi_j(\boldsymbol{x}_2)\cdots\chi_k(\boldsymbol{x}_N)\rangle \tag{2.40}$$

If the electron labels are ordered $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N$, in equation 2.40, then this can be further shortened to:

$$\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N) = |\chi_i\chi_j\cdots\chi_k\rangle \tag{2.41}$$

2.4.5 Hartree-Fock Approximation

The following derivation is taken from by Introduction to Advanced Electronic Structure Theory by Szabo and Ostlund. [38] The simplest antisymmetric wave function, which can be used to describe the ground state of an *N*-electron system, is a single slater determinant:

$$|\Psi_0\rangle = |\chi_i\chi_j\cdots\chi_N\rangle \tag{2.42}$$

The variation principle states that the best wave function of this functional form is the one which gives the lowest possible energy:

$$E_0 = \langle \Psi_0 | \mathcal{H} | \Psi_0 \rangle \tag{2.43}$$

In equation 2.43, $\mathcal{H}$ is the full electronic Hamiltonian, E_0 is the energy of the wave function, Ψ_0. By minimizing E_0 with respect to the choice of spin orbitals, the Hartree-Fock (HF) equation can be derived (equation 2.45), which determines the optimal energy for the spin orbitals ($\varepsilon\chi$).

$$f(i)\chi(\boldsymbol{x}_i) = \varepsilon\chi(\boldsymbol{x}_i) \tag{2.44}$$

In equation 2.44, $f(i)$ is an effective one-electron operator, called the Fock operator, which has the following form:

$$f(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \upsilon^{HF}(i) \tag{2.45}$$

In equation 2.45, υ^{HF} is the average potential experienced by the ith electron due to the presence of the other electrons. The essence of the HF approximation is to replace the complex many-electron problem by a one-electron problem in which electron-electron repulsion is treated in an average way.

The HF potential $\upsilon^{HF}(i)$, or alternatively the 'field' seen by the ith electron, depends on the spin orbitals of the other electrons. Thus the HF equation (equation

2.45) is non-linear and must be solved iteratively. The procedure for solving the HF equation is called the self-consistent-field (SCF) method.

The basic premise behind the SCF method is simple. By taking an initial guess at the spin orbitals, it is possible to calculate the average field (υ^{HF}) seen by each electron and then solve the eigenvalue equation (equation 2.44) for a new set of spin orbitals. Using these new spin orbitals, new fields are calculated, this procedure is repeated until self-consistency is reached (i.e., until the fields no longer change and the spin orbitals used to construct the Fock operator are the same as its eigenfunctions).

2.4.6 Density Functional Theory

Density functional theory (DFT) is a theory of correlated many-body systems. It has become one of the most important tools for calculation of electronic structure in condensed matter, and is increasingly used for quantitative studies of molecules and other finite systems.

The huge success of the approximate local density (LDA) [41] and generalised-gradient approximation (GGA) [42] functionals within the Kohn-Sham approach has led to widespread interest in DFT as one of the most promising approaches for accurate, practical methods in theory of materials.

2.4.6.1 Hohenberg-Kohn Theorems

The modern formulation of DFT originated from Hohenberg and Kohn (HK). The approach of HK was to formulate DFT as an exact theory of many body systems [42].

The formulation applies to any system of interacting particles in an external potential $V_{ext}\ (\boldsymbol{r})$, including any problem of electrons and fixed nuclei.

DFT is based upon two theorems first proved by HK [43], stated below.

- **Theorem I:** For any system of interacting particles in an external potential $V_{ext}\ (\boldsymbol{r})$, the potential $V_{ext}\ (\boldsymbol{r})$ is determined uniquely, except for a constant, by the ground state density $n_0(\boldsymbol{r})$.
- **Corollary I**: Since the Hamiltonian is fully determined, except for a constant shift of the energy, it follows that the many-body wave functions for all states (ground and excited) are determined. Therefore all properties of a system are completely determined given only the ground state density $n_0(\boldsymbol{r})$.
- **Theorem II**: A universal functional $E[n]$ in terms of density $n(\boldsymbol{r})$ can be defined, valid for any external potential $V_{ext}\ (\boldsymbol{r})$. For any particular $V_{ext}\ (\boldsymbol{r})$, the exact ground state energy of the system is the global minimum value of this functional, and the density $n(\boldsymbol{r})$ that minimises the functional is the exact ground state density $n_0(\boldsymbol{r})$.
- **Corollary II**: The functional $E[n]$ alone is sufficient to determine the exact ground state energy and density. In general, excited states of the electrons must be determined by other means.

2.4.6.2 Kohn-Sham Density Functional Theory

The Kohn-Sham (KS) approach is to replace the difficult many-body system obeying the Hamiltonian (equation 2.27) with a different auxiliary system that can be solved more easily [44]. Since there is no unique option for choosing the simpler auxiliary system,

this is an approach that rephrases the problem. The approach of KS construction of the auxiliary system rests upon two assumptions:

1. The exact ground state density can be represented by the ground state density of an auxiliary system of non-interacting particles. This is termed "*non-interacting-V-representability;*" although there are no rigorous proofs for real systems of interest, we will proceed assuming its validity.
2. The auxiliary Hamiltonian is chosen to have the usual kinetic operator and an effective local potential (V_{eff}^{σ}) acting on an electron of spin σ at point $\boldsymbol{r}$. The local form is not essential, but is an extremely useful simplification that is often taken as the defining characteristic of the KS approach. The external potential $V_{ext}\ (\boldsymbol{r})$ is assumed to be spin independent.

The actual calculations are performed on the auxiliary independent-particle system defined by the auxiliary Hamiltonian:

$$\mathcal{H}_{aux}^{\sigma} = -\frac{1}{2}\nabla^2 + V^{\sigma}(\boldsymbol{r}) \tag{2.46}$$

In equation 2.46, $\mathcal{H}_{aux}^{\sigma}$ is the auxiliary Hamiltonian and $V^{\sigma}(\boldsymbol{r})$ is the auxiliary potential. At this point the form of $V^{\sigma}(\boldsymbol{r})$ is not specified and the expressions must apply for all $V^{\sigma}(\boldsymbol{r})$ in some range. For a system of $N = N\uparrow + N\downarrow$ independent electrons obeying the Hamiltonian, the ground state has one electron in each of the N^{σ} orbitals $\varphi_i^{\sigma}(\boldsymbol{r})$ with the lowest eigenvalues ϵ_i^{σ} of the Hamiltonian (equation 2.46). The density of the auxiliary system is given by sums of squares of the orbitals for each spin state:

$$n(\boldsymbol{r}) = \sum_{\sigma} n(\boldsymbol{r}, \sigma) = \sum_{\sigma} \sum_{i=1}^{N^{\sigma}} |\varphi_i^{\sigma}(\boldsymbol{r})|^2 \tag{2.47}$$

The independent-particle kinetic energy T_s is defined as:

$$T_s = -\frac{1}{2} \sum_{\sigma} \sum_{i=1}^{N^{\sigma}} \langle \varphi_i^{\sigma} | \nabla^2 | \varphi_i^{\sigma} \rangle = \frac{1}{2} \sum_{\sigma} \sum_{i=1}^{N^{\sigma}} \int d^3 r |\nabla \varphi_i^{\sigma}(\boldsymbol{r})|^2 \tag{2.48}$$

The classical Coulomb interaction energy of the electron density $n(\boldsymbol{r})$ interacting with itself is given by:

$$E_{Hartree}[n] = \frac{1}{2} \int d^3 r d^3 r' \frac{n(\boldsymbol{r}) n(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} \tag{2.49}$$

The KS approach to full interacting many-body problem is to rewrite the HK expression for the ground state energy functional as:

$$E_{KS} = T_s[n] + \int d\boldsymbol{r} V_{ext}(\boldsymbol{r}) n(\boldsymbol{r}) + E_{Hartree}[n] + E_{II} + E_{xc}[n] \tag{2.50}$$

In equation 2.50, $V_{ext}(\boldsymbol{r})$ is the external potential due to the nuclei and any other external fields, E_{II} is the interaction between the nuclei and $E_{xc}[n]$ the exchange-correlation functional. Hence the sum of these terms V_{ext}, $E_{Hartree}$, and E_{II} forms a neutral grouping that is well defined. The independent-particle kinetic energy T_s is given explicitly as a functional of the orbitals; however, T_s must be a unique functional

for each spin state σ of the density $n(\boldsymbol{r}, \sigma)$ by application to the independent-particle Hamiltonian (equation 2.47).

All of the many-body effects of exchange and correlation are grouped into the exchange-correlation energy E_{xc}. Therefore the KS approach explicitly separates out the independent-particle kinetic energy and the long range Hartree terms, the remaining exchange-correlation functional $E_{xc}[n]$ can reasonably be approximated as a local or nearly local functional of the density. This means that the energy E_{xc} can be expressed as:

$$E_{xc}[n] = \int d\boldsymbol{r}\, n(\boldsymbol{r}) \epsilon_{xc}([n], \boldsymbol{r}) \tag{2.51}$$

In equation 2.51, ϵ_{xc} $([n], \boldsymbol{r})$ is the energy per electron at point $\boldsymbol{r}$ that depends only upon the density $n(\boldsymbol{r}, \sigma)$ in some vicinity of point $\boldsymbol{r}$.

2.4.7 Basis Sets

To consider performing a closed shell calculation we must give some consideration to basis sets. If each molecular orbital, $\emptyset_{i\lambda\alpha}$, is represented as a linear combination, it is possible to obtain the following:

$$\emptyset_{i\lambda\alpha} = \sum_{p} c_{i\lambda p} \chi_{p\lambda\alpha} \tag{2.52}$$

As the wave function should be a single valued, finite, continuous, and quadratically integrable, the basis functions $\chi_{p\lambda\alpha}$ selected should also have these properties. For a

basis set of size m, there are $\frac{1}{2}m(m+1)$ one-electron integrals and $\frac{1}{8}(m^4+2m^3+3m^2+2m)$ two-electron integrals [37]. Even if m is small, it can easily seen that the number of integrals is rather large; this leads to a need for large amounts of computational resources. Therefore, it is clear that the basis set needs to have a reasonable size (smaller is better). However, the basis set must not be so small as to affect the accuracy adversely; it must be sufficiently flexible to describe the distortion resulting from molecule formation.

2.4.7.1 Slater Type Orbitals – Minimal Basis Sets

The solution of the Schrodinger equation for hydrogen-like atoms suggests the use of atomic orbitals with the functional form:

$$\chi_{nlm} = r^{n-1}e^{-\xi r}Y_{lm}(\theta,\phi) \tag{2.53}$$

In equation 2.53, the radial variation consists of a power r multiplied by an exponent. A 1s function depends on $e^{-\xi r}$, a 2p function are $re^{-\xi r}$, a 3d function on $r^2e^{-\xi r}$, and so on. The angular part in θ and ϕ is as spherical harmonic $Y_{lm}(\theta,\phi)$. The exponent ξ is an adjustable parameter.

This type of function is usually normalised and is then called a Slater-type orbital (STO). It is possible to construct a basis set for a many-electron atom by taking one or more STOs of the correct symmetry for each occupied orbital. STOs are used as basis functions for most accurate calculations on atoms and small molecules. The problem is that the many-center two electron integrals are extremely expensive to

compute. This is a major problem for larger molecules and was the major reason for the introduction of alternative types of basis functions.

2.4.7.2 Gaussian Type Orbitals (GTO) – Pople Basis Sets & Correlation Consistent Basis Sets

The Cartesian Gaussian orbitals of the form $x^1 y^m z^n e^{-\alpha r^2}$, first suggested by Boys [46], have proved extremely useful in *ab initio* calculations of polyatomic molecules. The product of two GTOs is another GTO so that many-center two-electron integrals reduce to a much simpler form. With GTOs all s-functions are taken to behave as $e^{-\alpha r^2}$. Similarly, all p_z GTOs behave as $ze^{-\alpha r^2}$ and all d_{xy} GTOs behave as $xye^{-\alpha r^2}$ [37].

The main disadvantage of the Gaussian function is that it does not resemble the form of real atomic orbital wave functions. This defect may be overcome through the use of multiple Gaussian functions, however this introduces in the calculation as it becomes very difficult to get an iterative process to converge with a very large number of basis functions.

However, a method has been found to reduce the number of variables in the SCF calculation with very little loss of accuracy. Instead of allowing all the coefficients of the basis function to expand freely, certain coefficients are fixed relative to one another, forming groups of Gaussian functions, each know as a 'contracted Gaussian' or CGTO. The molecular orbital can be expressed as:

$$\emptyset_{i\lambda\alpha} = \sum_p c_{i\lambda p} \gamma_{p\lambda\alpha} \tag{2.54}$$

In equation 2.54, $\Upsilon_{p\lambda\alpha}$ is a small contraction of Gaussians of the same type on the same center. In this way, a large basis set may be broken up into a much smaller number of groups.

2.5 Sampling Phase Space

MM and QM allow us to calculate the interaction energies between atoms in a system, but they do not allow us to sample the phase space of the system. There are two major methods which are used to sample the phase space of systems; Monte Carlo (MC) and Molecular Dynamics (MD). These methods are explained in the following sections.

2.5.1 Monte Carlo

MC simulations are an example of a stochastic technique which is used to sample properties of a system. Attempting to calculate the properties of a system by averaging over every configuration is impossible, so a method for sampling the states which make the largest contribution to the partition function is required. This can be achieved by generating a Markov chain of configurations, whereby each new configuration is generated by a random change in the preceding configuration. Such changes are typically made through making a change to the Cartesian coordinates of one or more particles in the system. A problem with this method is that a huge portion of the phase space does not make an important contribution to the partition function. Hence, a method is needed to sample the states which contribute most to the partition function.

A method for sampling the most relevant states to the partition function was developed by Metropolis [46]. The Metropolis algorithm is detailed below:

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

It is crucial that in the Metropolis MC algorithm that detailed balance is preserved. That is, the probability of moving from *i* to *j*, before weighting by π_i and π_j is the same as the probability of moving from *j* to *I* [47]. When this is obeyed, the acceptance test for the move from state *i* to *j* is:

$$\frac{\pi_j}{\pi_i} = \frac{\exp\left(-\beta U_j\right)/Z_N}{\exp\left(-\beta U_i\right)/Z_N} \tag{2.55}$$

In equation 2.55, Z is the configurational integral of the system, and is proportional to the potential energy part of the partition function. Fortunately it can be seen that the two configurational integrals in equation 2.19 cancel, as this parameter cannot be determined for large systems since it is a 6 dimensional integral.

This leaves the acceptance test as equation 2.56:

$$\frac{\pi_j}{\pi_i} = \frac{\exp\left(-\beta U_j\right)}{\exp(-\beta U_i)} = \exp\left(-\beta\left(U_j - U_i\right)\right) \tag{2.56}$$

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

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

A procedure such as this ensures that only configurations which make the largest contribution to the partition function are included in the running average. Once the simulation has finished, the properties of interest are found by averaging over all of the accepted configurations using equation 2.57:

$$\langle A \rangle = \frac{1}{M}\sum_{i=1}^{M} A(\boldsymbol{r}^N) \tag{2.57}$$

In equation 2.57, M is the number of configurations and $\boldsymbol{r}^N$ the Cartesian coordinates of that particular configuration. MC simulations can be performed in at least three different ensembles, each with their own acceptance tests:

- Canonical ensemble (NVT): constant temperature, number of particles and volume. Equation 2.57 shows the acceptance test for this move.
- Isothermal-Isobaric ensemble (NPT): constant temperature, pressure and number of particles. Equation 2.58 shows the acceptance test for this move.

$$acc(A \rightarrow B) = min\left[1, \exp\left(\frac{-\Delta E + P\,(V^n - V^0)}{k_B T} + N \ln \frac{V^n}{V^0}\right)\right] \quad (2.58)$$

In equation 2.58, P is the pressure of the system, with V^n and V^0 denoting the new and original volumes of the system respectively. N is the number of molecules in the system.

- Grand-Canonical ensemble (μVT): constant chemical potential, volume and temperature. The acceptance test for the Grand-Canonical ensemble shall be discussed in a later section (section 6.3.2).

2.5.2 Molecular Dynamics

In contrast to the stochastic MC technique, MD is deterministic, meaning that the preceding configurations can be found from the current configuration. In MD, the n+1th configuration is found by integrating Newton's laws of motion. A "trajectory" is

found during a MD simulation, which tracks the positions and velocities of the particles as a function of time. This trajectory is based upon Newton's second law of motion, $F = m\boldsymbol{a}$ and the resultant differential equations:

$$\frac{d^2\boldsymbol{x}_i}{dt^2} = \frac{\boldsymbol{F}_{xi}}{m_i} \tag{2.59}$$

Given an initial starting configuration and velocities, all details relating to the trajectory can be found at any point in space. Since the movement of one atom in the system will affect the velocity and position of other atoms in the system, integrators are used in MD to help calculate the new positions and velocities. Although these integrators are extremely efficient they require a large number of computations to be performed in parallel, meaning that running an MD simulation requires significantly more computational power than a MC simulation due to the large number of forces which must be calculated. Although, this can be seen as a potential drawback to using MD, it does allow efficient sampling of extremely large systems to be performed, something which can be very difficult for MC simulations to achieve.

2.6 Free Energy Simulations

For the canonical ensemble, the free energy of the system is expressed as the Helmholtz function, **A**, as in equation 2.17. Free energy is integral to chemistry, since it is the driving force behind most, if not all, chemical processes. Thus it can be seen as a highly desired quantity to obtain, although this can be extremely difficult for systems with a large number of particles. As shown in equation 2.19 the partition function *Q* is

a 6 dimensional integral, and evaluating this integral for large systems becomes an impossible task. Methods such as MC and MD can only sample the low energy regions of space, and any attempts to sample the large number of degrees of freedom which contribute to *Q* becomes computational intractable. As such, methods have been devised which allow the calculation of relative free energies between two systems; a calculation which is much easier to perform [48]. Such approaches are generally classified into two distinct categories; rigorous methods and approximate methods.

2.6.1 Rigorous Methods

Free Energy Perturbation

According to Zwanzig [49] the free energy difference between two states, A and B, can be expressed as:

$$\Delta G_{A\to B} = -k_B T \, ln \langle \exp\left(-\frac{\Delta E}{k_B T}\right)\rangle_A \tag{2.60}$$

In equation 2.60, $\langle \dots \rangle_A$ represents the ensemble average over system A and ΔE represents the energy change between states B and A. A is termed the reference state, whilst B is known as the perturbed state. One potential problem with this method is caused if state A and B do not overlap well in phase space. If this is the case, a simulation run at a potential, U_A, would not sample enough relevant configurations of U_B which would result in a large value of ΔE. Equation 2.60 highlights that this results in the exponent becomes very small – resulting in small numbers counting to the overall system average. The rare occasions where ΔE is small has a disproportionate

effect on the convergence of the average term, making the overall calculated free energy unreliable without performing very long simulations. A solution is to link states A and B via the use of a coupling parameter, λ, which introduces intermediate states. For the example above, state A would correspond to λ=0, whereas state B represents λ=1.

In the free energy perturbation approach (FEP), the simulation is broken into multiple λ windows between the two end states. Each window is defined by a specific λ value, and the free energy of that reference state is calculated. The energy between each λ window is found, and the utilized by equation 2.61:

$$\Delta G = \sum_{\lambda=0}^{1} -k_B T \; ln \langle \exp\left(-\frac{\Delta E'}{k_B T}\right)\rangle_\lambda \tag{2.61}$$

In equation 2.61, $\Delta E'$ refers to the free energy difference between the states λ + Δλ and λ, where Δλ is the interval between two successive λ windows.

Thermodynamic Integration

While FEP directly uses equation 2.61 to calculate the difference in free energy along the λ coordinate, TI takes a different approach. Instead of calculating the difference in energy between neighboring λ values, it calculates the rate of change of free energy, with respect to each λ, across a λ trajectory. The free energy gradients are then integrated to give the relative free energy change as in equation 2.62:

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \langle \frac{\delta G}{\delta \lambda} \rangle_{\lambda} \, d\lambda \tag{2.62}$$

Finite Difference Thermodynamic Integration

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

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \langle \frac{\Delta G}{\Delta \lambda} \rangle_{\lambda} \, d\lambda \tag{2.63}$$

The simulation is broken up into multiple λ windows and, for each value of λ, the free energy is computed over a small interval, typically as low as λ+0.001 [50], using FEP. The total free energy change is then obtained via integrating over all the computed values. Therefore, FDTI is very similar to FEP. In FEP, the perturbed states are the neighbouring λ windows, whilst in FDTI the perturbed states are Δλ above and below each λ window.

Replica Exchange Thermodynamic Integration

The replica exchange thermodynamic integration (RETI) method can be considered as a combination of TI with the Hamiltonian replica exchange method [51, 52]. The λ co-ordinate scales the forcefield terms linearly, leading to a system which has a different

Hamiltonian at each λ value. Within RETI, the co-ordinates between neighbouring λ values are periodically swapped according to the following Metropolis test:

$$rand(0,1) \leq \exp\left[\frac{1}{k_B T}\left(E_B(j) - E_B(i) - E_A(j) + E_A(i)\right)\right] \quad (2.64)$$

In equation 2.64, A and B are two neighbouring λ windows and i and j are replicas of the system at those λ values. RETI has been applied to the calculation of relative binding free energies of small "drug-like" molecules to proteins, and has been shown to enhance the sampling of phase space. Since configurations of different λ windows are passed across the λ-coordinate, it has been observed that better free energy convergence can also be obtained [48].

2.6.2 Approximate Free Energy Methods

Whereas rigorous free energy methods take into account the intermediate λ states between our endpoints, λ=0 and λ=1, approximate methods typically only take into account the end points of the simulation. As a result of this approximate free energy methods are typically much faster than more rigorous methods, despite the increase in computational efficiency; the accuracy of such methods is typically significantly poorer. One of the most widely used approximate methods is the Molecular Mechanics / Generalised Born (Poisson Boltzmann) Surface Area (MM/GB (PB)SA) methods. In MM/GB(PB)SA the two end points are simulated using MD or MC and the free energy is calculated using equation 2.65:

$$\Delta G_{bind} = \langle \Delta E_{mm} \rangle + \Delta G_{solv} - T\Delta S \quad (2.65)$$

In equation 2.65, $\langle \Delta E_{mm} \rangle$ is the difference in the MM energy between the complex and the isolated protein and ligand. ΔG_{solv} is found as the difference in the solvation energy between the complex and its individual components, although it usually extremely challenging to calculate the non-polar contribution to the solvation free energy [53]. The methods also require the calculation of an entropic term to calculate the change in binding free energy, although this is often difficult to estimate. This term is commonly ignored when considering structurally similar ligands, since it is assumed that the change in entropy between structurally similar ligands binding to the same receptor is sufficiently small enough to ignore [54]. The results obtained using these methods are typically much less accurate than those obtained via more rigorous methods. Despite this, MM/GB(PBSA) methods have been utilised in the re-scoring of docked protein-ligand complexes [53].

Another end-point approach is the GCMC method [55, 56], which shall be discussed at length in section 6.3.2.

2.7 Conclusions

In this chapter, a brief overview of statistical mechanics was presented. As all of the key quantities which we would like to calculate from a molecular simulation, such as free energy, can be derived from the molecular partition function, the target for many computational methods is to try and calculate this property. Owing to the extremely complex nature of the partition function, direct calculation is impossible, and thus various methods have been devised to calculate the relative free energy difference

between two systems. Sampling methods such as MD allow us to look at large systems as it is easily parallelised, although it is often limited in its ability to perform novel sampling schemes, unlike MC. Using TI to investigate the relative binding free energy between two inhibitors is a rigorous approach to produce accurate estimates. Although it has known that purely MM representation of protein-ligand systems can lead to large errors in these free energy estimates, due to the lack of polarisation terms within conventional MM forcefields. This drawback means it is unlikely that pharmaceutical companies would screen lead compounds using such methods. The inclusion of QM or QM/MM corrections in free energy simulations could lead to improved accuracy within such calculations, and is a matter which will be addressed in this study.

3 QM/MM – Background and Review

3.1 Introduction

Hybrid QM/MM approaches were first introduced to the study of enzymatic processes by Warshel and Levitt in 1976 when they applied their methodology to calculate the dielectric, electrostatic and steric stabilization via the carbenium ion within the lysozyme enzyme [57]. This seminal research was recently recognized and awarded the Nobel Prize for Chemistry in 2013, the first Nobel Prize for a computational paper in the awards distinguished history. Since this, various QM/MM models have been developed, with multiple QM being combined with many classical MM forcefields. The development of specialized QM/MM softwares and codes, including ONIOM [58] and PUPIL [59] demonstrate the importance of QM/MM, and its ever growing popularity as a tool for describing large biomolecular systems. The following sections shall discuss the methodological issues with QM/MM, its historical development, and its use in computational chemistry; with a particular focus of QM/MM free energy studies.

3.1.1 The QM/MM method

This section will focus on the QM/MM method, describing how a QM/MM hybrid system is constructed and the methodologies developed to couple the two different systems.

3.1.1.1 QM – MM Partition

The basic premise of all QM/MM methods is to treat the part of the system which undergoes the most important electronic changes using QM, whilst the rest of the system is described using MM (Figure 3.1).

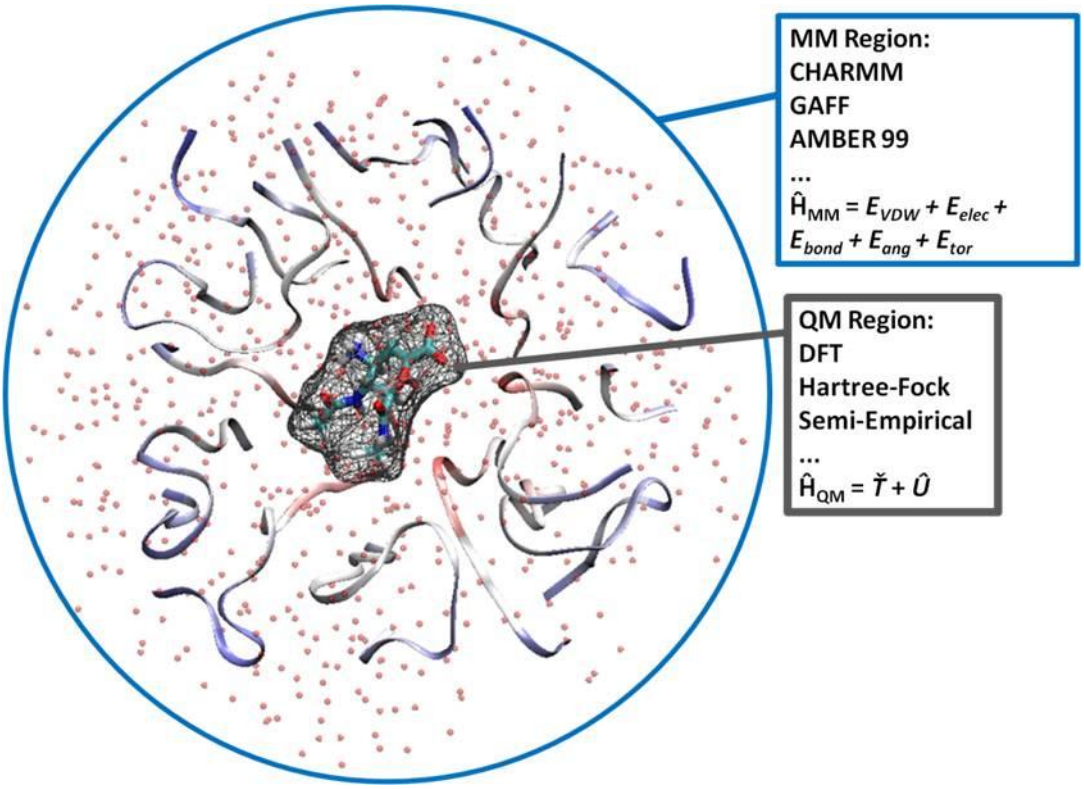


Figure 3.1: A depiction of a hybrid QM/MM system for Tamiflu bound to N9-Neuraminidase. The grey caged region around Tamiflu represents our QM region, with the protein (ribbons) and water (red dots) making up the MM part of the system. This figure was generated using VMD v1.8.6.

The MM region is described using a classical MM forcefield, while within the QM region, which is typically polarized by the MM environment, electronic properties and reaction mechanisms can be studied. QM methods, such as *ab-initio* Hartree-Fock (HF), Density Functional Theory (DFT), and semi-empirical methods have all been combined

with classical forcefields for QM/MM simulations [60, 61]. The overall QM/MM Hamiltonian ($\mathcal{H}_{QM/MM}$) for such a system can be described as a combination of the MM Hamiltonian of choice ($\mathcal{H}_{MM}$) and QM Hamiltonian of choice ($\mathcal{H}_{QM}$) equation 3.1:

$$\mathcal{H}_{QM/MM} = \mathcal{H}_{MM} + \mathcal{H}_{QM} \tag{3.1}$$

3.1.1.2 QM/MM Energy Expressions

To obtain a QM/MM description of the system the MM and QM Hamiltonians must be combined. Two main approaches exist to do this; these are known as additive and subtractive QM/MM schemes. An additive scheme, equation 3.2, requires an MM calculation of the outer MM region ($\mathcal{H}_{MM}$), a QM calculation of the inner QM region ($\mathcal{H}_{QM}$), and explicit treatment of QM/MM coupling terms ($\mathcal{H}_{QM/MM\ -\ coupling\ terms}$).

$$\mathcal{H}_{QM/MM} = \mathcal{H}_{MM} + \mathcal{H}_{QM} + \mathcal{H}_{QM/MM\ -\ coupling\ terms} \tag{3.2}$$

The QM/MM coupling terms normally include bonded terms across the QM – MM boundary, non-bonded, van der Waals (VDW) and electrostatic terms. In contrast, subtractive schemes, equation 3.3, require an MM calculation of the entire system ($\mathcal{H}_{MM-entire\ system}$), a QM calculation of the inner QM region ($\mathcal{H}_{QM}$), and an MM calculation of the inner QM region ($\mathcal{H}_{MM-\ inner\ (QM)\ region}$). The QM/MM energy ($\mathcal{H}_{QM/MM}$) is then obtained by simply summing and subtracting to avoid double counting.

$$\mathcal{H}_{QM/MM} = \mathcal{H}_{MM-entire\ system} + \mathcal{H}_{QM} - \mathcal{H}_{MM-\ inner\ (QM)\ region} \quad (3.3)$$

In such a scheme, the QM/MM interactions are handled at the MM level of theory. This can lead to complications in regard to electrostatic interactions which will typically involve atomic point charges interacting in both the MM and QM regions. Such drawbacks make subtractive schemes less attractive than additive schemes in QM/MM applications [59]. Despite this, it is important to note the implementation and generalization of subtractive schemes is often easier and faster due to the lack complex QM/MM coupling terms.

3.1.2 QM/MM Electrostatic Interactions

As mentioned above, QM/MM coupling terms include terms to couple the electrostatics, bonded and VDW components of our MM and QM systems. Of these terms, it is often the electrostatic coupling which is considered the most important. The electrostatics can be coupled through the use of three different embedding techniques; mechanical embedding, electrostatic embedding, and polarized embedding.

3.1.2.1 Mechanical Embedding

Mechanical embedding (ME), much like a subtractive scheme, considers the QM/MM electrostatics at the MM level of theory and both the QM and MM environments remain un-polarized. To apply ME an MM calculation is performed on the entire

system followed by a vacuum QM calculation on the QM region (Figure 3.2). The resultant energies can either be added or subtracted dependent upon the QM/MM energy scheme employed.

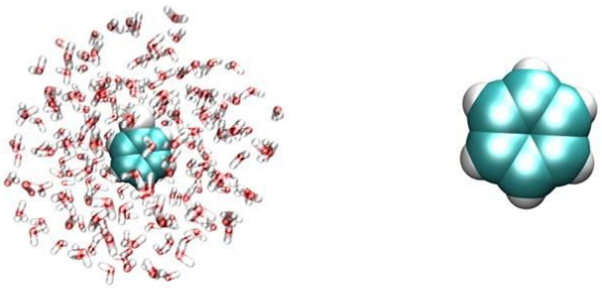


Figure 3.2: A ME type QM/MM system with a full MM simulation (A) combined with a QM vacuum (B) calculation. This figure was generated using VMD v1.8.6.

The original integrated molecular orbital molecular mechanics (IMOMM) method developed by Morokuma *et al.* [62, 63, 64], which is also referred to as the two layer ONIOM(MO:MM) method is an example of a method employing ME.

Such a simple treatment for the QM/MM electrostatic coupling has drawbacks. First, it requires accurate MM parameter sets for both the MM and QM regions. This can be very difficult for the QM region as this region typically undergoes significant electronic changes. Second, such a scheme neglects the potential polarisation of the QM region via the atomic charges of the MM region. It is possible to consider deriving the atomic charges for the QM region dynamically as the reaction progresses. However, this would increase the computational effort needed considerably, making such methods rather unattractive.

3.1.2.2 Electrostatic Embedding

In contrast to an ME scheme, Electrostatic Embedding (EE) allows the QM charge density to be polarized via the atomic charges of the surrounding MM environment. To implement EE an MM calculation of the entire system is followed by a QM calculation on the QM region with the atomic charges of the MM region embedded in this calculation (Figure 3.3).

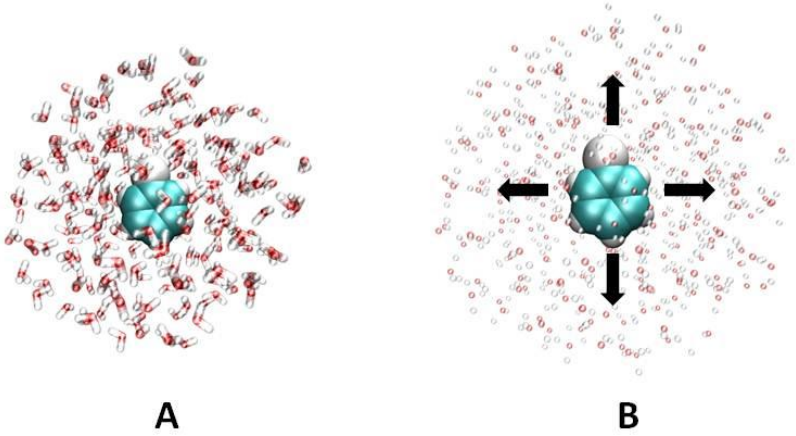


Figure 3.3: An EE approach with a full MM simulation (A) combined with a QM/MM calculation (B) with embedded point charges (red and white spheres). This figure was generated using VMD v1.8.6.

This is a much better approximation than ME as the QM charge density can adjust based on the MM environment. The cost of this improved QM/MM representation is a more complicated implementation and increased computational cost.

This problem of improving the embedding scheme within a QM/MM system was tackled by Morokuma *et al.* with their three layered ONIOM method [65]. This method attempts to overcome the drawbacks of a ME two-layer ONIOM(MO:MM) [62] by introducing a buffer (middle) layer, which is treated by an appropriate lower-level QM theory (e.g., semi-empirical molecular orbital theory), which is computationally

less expensive than the method used for the innermost primary subsystem. One can label such a treatment as QM1:QM2:MM or QM1/QM2/MM. The second QM layer is designed to allow a consistent treatment of the polarization of the active center by the environment. The new treatment does improve the description of the QM/MM system, but, as with ME, it does not solve the problem completely, since the QM calculation for the first layer is still performed in the absence of the rest of the atoms from the system.

Another issue is in ascertaining the best EE strategy to polarize the QM region. In principle, the MM and QM regions will polarize each other until their charge distributions are self-consistent; this is usually computed using an iterative scheme [66] or by an extended Lagrangian scheme [67]. Ideally, an EE scheme should include this self-consistently, but usually the charge distribution of the MM region is considered frozen for a given set of MM nuclear co-ordinates. Schemes that relax this constraint are known as polarizable embedding (PE) schemes (Figure 3.4).

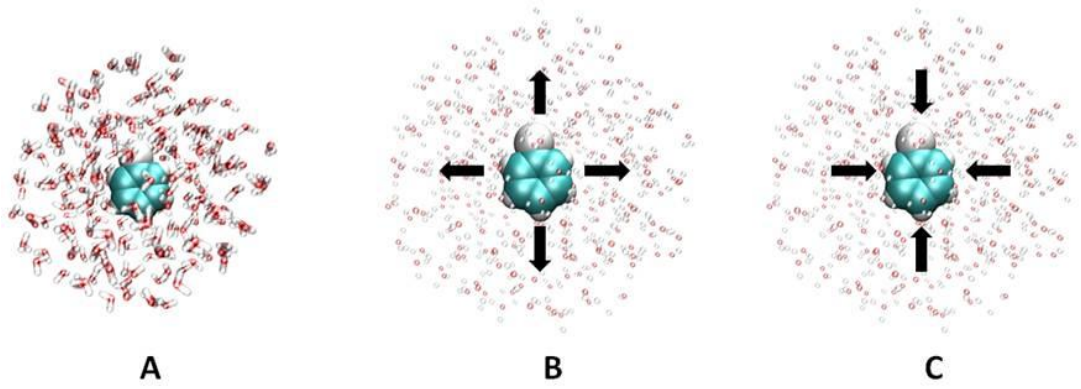


Figure 3.4: A PE scheme with a full MM simulation (A), combined with a QM/MM simulation (B) with embedded point charges (red and white spheres), combined with a further QM/MM simulation (C) to calculate the back polarisation of our MM point charges (red and white spheres) via the now polarised QM ligand. This figure was generated using VMD v1.8.6.

However, self-consistency is difficult to achieve, because it requires a polarizable MM force field [68, 69], which has the flexibility to respond to perturbation by an external electric field. Such flexibility is not available in today's most popular MM force fields, although research to develop a polarizable force field has received much attention [70, 71]. Moreover, the use of a self-consistent embedding scheme also brings additional complications to the treatment of the boundary between the QM and MM regions. It also increases the computational effort, since iterations are required to achieve self-consistent polarization of the MM and QM regions. Thus, in most EE implementations, the QM region is polarized by the MM, but the MM is not back polarized by the now polarized QM region. Early examinations on the self-consistent embedding scheme were carried out by Thompson and Schenter [72] and Bakowies and Thiel [73]. Their treatments are based on models that describe the mutual polarization of QM and MM fragments via the use of the reaction field [74, 75] theory, with the difference that the response is generated by a discrete reaction field (atomic polarizabilities) rather than a

continuum. Their results suggest that the polarization of the MM region by the polarized QM region can be crucial in applications involving a charged QM region that generates large electric fields.

3.1.3 Boundary Treatment

If a substrate is covalently bonded to the enzyme, or if the inclusion of parts of the enzyme environment (waters, ions, and cofactors) are desirable for other reasons, special techniques are needed for the treatment of the QM/MM border region as it is essential to avoid half-filled orbitals within the QM region, which would arise if the bonds were simply truncated (Figure 3.5).

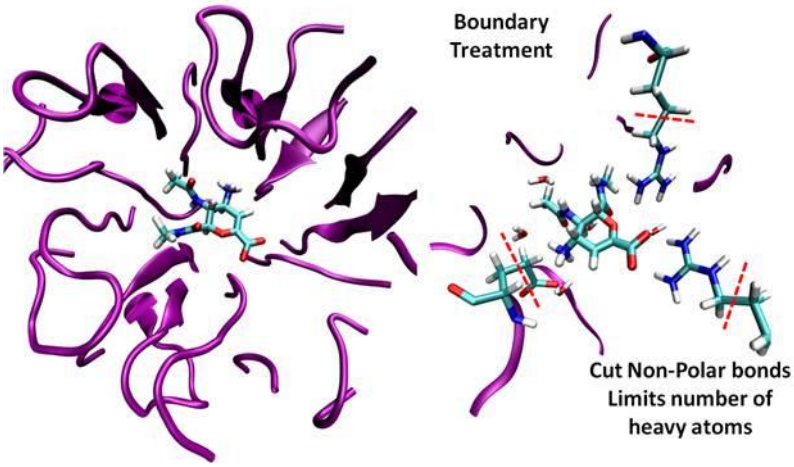


Figure 3.5: An example of boundary treatment in QM/MM. On the left hand side we have a classical MM description of protein (ribbons) and ligand (licorice). On the right hand side we show how in QM/MM description the cutting of non-polar bonds of key binding site residues may be necessary to improve the QM/MM description of our system. This figure was generated using VMD v1.8.6.

It is often the case that non-polar bonds are cut, to limit the complexity of capping the cut bond, as any cap would severely affect polarity of a polar bond [76]. There are two

general approaches which have been developed to deal with this problem; link atoms and local orbitals.

3.1.3.1 Link Atoms

The link atom (LA) approach use 'link atoms' to cap the dangling bond at the 'frontier atom' of the QM region (Figure 3.6). The link atom is usually a hydrogen atom [77], or a parameterized atom, for example, a one free valence atom in the 'connection atom', [78], 'psuedobond' [79], and 'quantum capping potential' [80] schemes. These schemes involve a parameterized semiempirical Hamiltonian [78] or a parameterized effective core potential [79, 80], adjusted to mimic the properties of the cut bond.

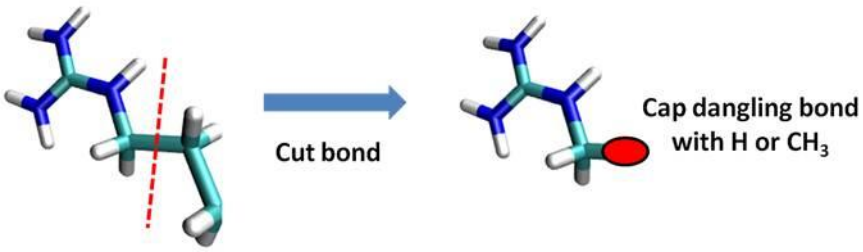


Figure 3.6: The link atom approach where a cut bond is capped by a H or CH_3 (red sphere) [78].

This figure was generated using VMD v1.8.6.

3.1.3.2 Local Orbitals

The second type of boundary treatment consists of methods that use local orbitals (LOs). In such a scheme a hybrid sp^3 / sp^2 orbital containing one electron is placed along the QM – MM partition (Figure 3.7).

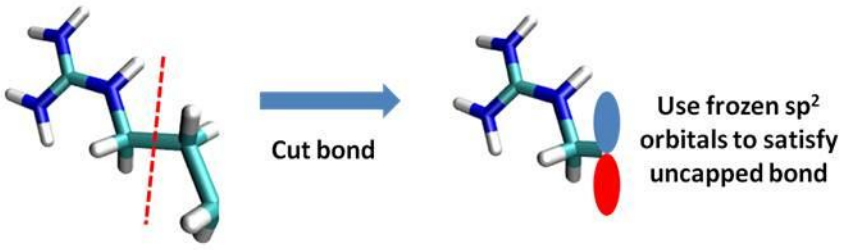


Figure 3.7: A local orbital approach [82, 83] to capping cut bonds in QM/MM where the bond is capped by a sp^3/sp^2 orbital (blue and red ovals). This figure was generated using VMD v1.8.6.

This is theoretically satisfying as the cut bond is dealt with at the QM level of theory and there is no need for the addition of another atom to the system as in the LA approach. Examples of such methods include the Local Self Consistent Field (LSCF) approach [81, 82], where the bonds connecting the QM and MM regions are represented by a set of strictly localized bond orbitals (SLBOs) that are generated by calculations on small model compounds. The SLBOs are excluded from the SCF optimization to prevent them interfering with other QM basis functions. Another approach is the Generalized Hybrid Orbital (GHO) approach [83, 84]. In this approach a set of four sp^3 hybrid orbitals are assigned to each of the MM boundary atoms. The hybridization is determined from the local geometry of the three MM atoms to which the boundary atom is bonded. The hybrid orbital that is directed towards the QM 'frontier atom' is known as the active orbital, while the other three are termed auxiliary orbitals. All four hybrid orbitals are included in the QM calculations, but only the active orbital participates in the SCF optimizations.

3.1.3.3 Link Atoms or Local Orbitals?

Both boundary treatment methods have their strengths and weaknesses. The LA approach is straightforward, easy to implement and is widely used. However, it does introduce artificial atoms that are not present in the original biomolecular system. This makes defining the QM/MM energy of the system difficult, whilst also presenting complications within the SCF optimizations [83]. In addition to this it is easy to generate unphysical polarization between the QM 'frontier atom' and the LA due to the nearby point charge on the MM 'boundary atom'. For example, when cutting a C···C bond the MM 'boundary atom' charge is around 0.5Å from the LA. At such short distances special treatments are needed to ensure MM charges near the boundary avoid this unphysical polarization [85, 86].

3.2 QM/MM – Applications

QM/MM methods are used to simulate a variety of chemical processes including reaction paths, [87] reaction mechanisms [88] and free energy studies [89]. The following sections shall discuss these different uses for QM/MM simulations, with particular focus on QM/MM free energy studies.

3.2.1 Reaction Paths, Stationary Points, and Reaction Mechanisms

In chemical reactions it is common to consider the path a reaction takes from reactants to products. Generally this reaction path (also known as a reaction co-ordinate) can be studied via computational methods. The application of QM/MM to such reactions can give insights into reaction mechanisms, the nature of transition state structures and the accurate assessment of the energies involved in such processes.

The first QM/MM study of an enzyme from Warshel and Levitt [57] demonstrated how QM/MM can be applied to understand enzymatic reactions, their transition states and the energies associated with these processes. They applied their QM/MM method to study the stability of the oxacarbenium ion intermediate formed in the cleavage of a glycosidic bond by lysozyme. It was found that electrostatic stabilization was an important factor in formation of the carbenium ion intermediate and that steric factors such as the strain of the substrate binding to lysozyme did not contribute significantly.

This seminal study of reaction mechanisms gave a platform for many more studies into enzymatic reactions, such as in the work of Mulholland *et al.* who have used QM/MM methods to study the reaction mechanisms in many enzymes; including cytochrome P450 [90, 91] where many studies have been performed to understand not only the activation barriers of enzyme catalysis, [92] but also the reaction mechanism for benzene hydroxylation via cytochrome P450 [91]. In other work they have also used QM/MM methods to understand inhibitor binding in fatty acid amide hydrolase [93] and the formation of a covalent intermediate in hen egg white lysozyme [94].

3.2.2 QM/MM – Free Energy Studies

There are many different approaches proposed and utilised within the literature to solve the problem of computing accurate QM/MM free energies for chemical reactions in solutions including enzymatic reactions [95]. A general approach within these methodologies is to use a fast but less accurate method to sample phase space and use

this sampling to estimate high level QM/MM free energies with a modest number of QM/MM calculations.

For example, in the quantum mechanical free energy (QM-FE) approach developed by Jorgensen *et al.* [96], a reaction pathway for atoms in the QM region is calculated in a vacuum. Free energies for the interaction between the QM and MM atoms are then calculated along the reaction pathway via performing MM FEP or TI calculations where electrostatic interactions between the QM and MM atoms are defined via point charge interactions. In the implementation by Jorgensen *et al.*; and later by Kollman *et al.* [97], the point charges used to represent QM atoms were derived in a vacuum. Jorgensen *et al.* used this method to study organic reactions in solution; Kollman *et al.* then extended the method to that of enzymatic reactions, including amide hydrolysis in trypsin [98] and methyl transfer by catechol O-methyltransferase [99].

In other work Yang *et al.* applied their QM/MM free energy method (QM/MM – FE) to the enzymes triosephosphate isomerase, [87] enolase [100] and 4-oxalocrotonate tautomerase [101]. The QM/MM – FE approach is an improvement upon the QM – FE approach as the QM/MM optimized reaction pathway and the QM derived energies and point charges are obtained via QM/MM calculations. In this approach the QM region is polarized via the surrounding MM environment. This approach was also adopted by Ishida and Kato to study acylation by serine proteases [102].

An alternative is the *ab initio* QM/MM approach (QM(ai)/MM) developed by Warshel *et al.* [103]. In this method MD simulations are used to sample phase space with a reference potential given by the empirical valence bond (EVB) method [104].

The use of umbrella sampling ensured that the entire reaction pathway was sampled and a potential of mean force (PMF) could be calculated from it. The changes in free energy between the system described by the reference potential and by DFT were calculated with FEP; in this way a high level QM/MM PMF can be obtained. In principle this methodology is exact with respect to how free energy changes are calculated. However, in practice the free energy for the protonated and deprotonated states of an aspartic acid surface residue of the bovine pancreatic trypsin inhibitor and a buried lysine residue in a hydrophobic pocket of the T4-lysozyme mutant did not converge due to large fluctuations of the difference between the reference potential and the high level QM/MM potential, although the electrostatic interactions did converge well [103]. Hence, Warshel *et al.* used more approximate methods, such as semi-empirical QM calculations, to calculate the free energy difference between the systems described using EVB and by high level QM/MM method [103].

In work similar to Warshel *et al.*, Roux *et al.* use FEP in their *ab initio*/classical free energy perturbation (ABC – FEP) approach, which they used to calculate the hydration energies of water and Na^+ and Cl^- ions at different physical conditions [105]. In this approach, only solute-solvent interaction energies are perturbed to the QM level.

Schofield and Bandyopadhyay developed a similar approach termed the molecular mechanics-based importance function (MMBIF) method [106]. They also used a MM reference potential to sample the phase space and to calculate high level QM/MM energies for a set of configurations. Based on two sets of energies they utilised a Metropolis-Hastings algorithm to generate a QM/MM canonical ensemble from which QM/MM free energies could be calculated.

More recently, *Senn et al.* developed a method which can be considered a combination of the QM/MM – FE method and the approach taken by Warshel *et al.* [107]. In this method the reaction pathway was optimized using QM/MM and a selected number of configurations for the QM region along the reaction pathway. Based on calculated point charges for the QM region, classical MM – QM interaction energies were calculated between subsequent fixed QM configurations along the reaction pathway. The QM/MM free energy change for each QM configuration was then calculated, and a high level QM/MM PMF was obtained. With this approach, Senn *et al.* obtained a converged PMF for the methyl transfer reaction in catechol O-methyltransferase. Consistent with results by Warshel and coworkers, they showed that the electrostatic interaction energies between the QM region and the MM region can be converged to high accuracy.

In addition, Reddy *et al.* implemented a QM/MM-FEP based approach to calculate free energies of hydration for a small set of organic molecule and to calculate the binding free energies of a set of five Cyclin Dependent Kinase 2 (CDK2) inhibitors [108, 109]. This work utilises the ME based approach to embed their MM system into their QM/MM representation. The results obtained from these studies showed little to no improvement on results obtained from purely classical simulations. This is not a significant surprise as the embedding technique neglects polarization of the QM ligand via the MM point charges. However, these studies highlighted the need for extended simulation time to achieve appropriate convergence of QM/MM free energies when sampling the full QM/MM Hamiltonian.

In contrast to some of the previously described methods, which sample the full QM/MM Hamiltonian, there are several methods which use QM/MM to provide

corrections for classical (MM) free energies. Woods *et al.* first published work in this area, where upon performing a conventional MM FEP between two systems, they then utilised QM/MM to calculate MM→QM/MM free energies, which were used as corrections for the MM free energies [110]. In this work they also implemented a Metropolis-Hastings acceptance criterion to build the QM/MM ensemble. This works by implementing a Metropolis-Hastings acceptance criterion to select snapshots from their endpoints of their MM ensemble that ensure the intramolecular energy of the QM region within MM snapshot is suitable for the QM/MM ensemble. The QM/MM system was represented as with a QM region (solute) with the rest of the system described via MM point charges, hence they employed EE within their QM/MM representation. This methodology was tested on a set of water (TIP3P and TIP4P) to methane. The MM results from this study showed excellent agreement between calculated and experimental value for this perturbation. The QM/MM results from this study showed poorer agreement with experiment, with an error ≈ 2kcal.mol^{-1}, for each combination of force field type (OPLS all-atom / OPLS united-atom) with different water models (TIP3P and TIP4P).

In a similar fashion, Beierlein *et al.* [111] implemented an MM→QM/MM FEP approach. Unlike Woods *et al.*, this method does not employ a Metropolis-Hastings acceptance criterion to build the QM/MM ensemble. Instead, all snapshots generated from each endpoint of a classical MM free energy calculation are used to build the QM/MM ensemble. This method was used to calculate the relative hydration free energy of methane → TIP4P – water and to calculate the relative binding free energy for a cyclooxygenase 2 (COX-2) ligand. For the hydration free energy study, the MM and QM/MM results followed a very similar pattern to Woods *et al.*, with QM/MM

calculated free energies giving slightly poorer agreement with experiment when compared to classical MM calculated free energies. The protein-ligand binding free energy of the COX-2 inhibitor produced QM/MM free energies which showed increased accuracy with experiment when compared to classically obtained MM binding free energies.

3.3 Conclusions

This section has highlighted the importance of QM/MM for understanding chemical and biological phenomena. The extensive methodological development of QM/MM has led to a wealth of methods combining different QM and MM approaches to understand a wide range of problems. The application of QM/MM has extended from biological systems to combat issues with inorganic surfaces to excited state spectroscopy. In particular, the application of QM/MM to calculate free energies has been scrutinised here, with QM/MM methods ranging from highly accurate *ab-initio* based calculations to more approximate QM/MM models were discussed. The focus of this thesis shall now turn to the application of our QM/MM method to calculate free energies.

4 Methodology: The MM→QM/MM – Free Energy Perturbation Approach

The MM→QM/MM method described by Beierlein *et al.* [111] was used for the free energy simulations reported in this thesis. This method employs a modified thermodynamic cycle perturbation approach, which first uses the MC/RETI [51, 52] technique to sample the MM free energy landscape and provide free energies for alchemical perturbations. Configurations from the endpoints of the sampled λ-coordinate are then used as input for DFT based QM/MM single point energy calculations. The desired energies are extracted from the MM and QM/MM data and placed into the Zwanzig equation to extract MM→QM/MM-FEP corrected free energies. The pathway independence of the calculated MM→QM/MM-FEP corrected free energies is tested through the use of charge perturbations.

This theory has been applied extensively to the calculation of hydration free energies (section 4.1) and protein-ligand free energies (section 4.2).

4.1 Hydration Free Energies

Owing to its simplicity, the calculation of hydration free energies has been used to test and validate free energy methodologies [112, 113]. Today, the prediction of hydration free energies remains of critical importance for testing and developing force fields and new methods. The free energy of hydration corresponds to the free energy of transferring a compound from well-defined reference state (gas) to another (aqueous) [114]. In addition, as the interaction of a solute with its environment in the gaseous

state is effectively zero, only interactions between the solute with the aqueous environment need to be considered.

Therefore, to calculate the hydration free energy between two end point states (λ=0 and λ=1) we must construct a free energy cycle (Figure 4.1) where both endpoints are simulated in both aqueous and gaseous states:

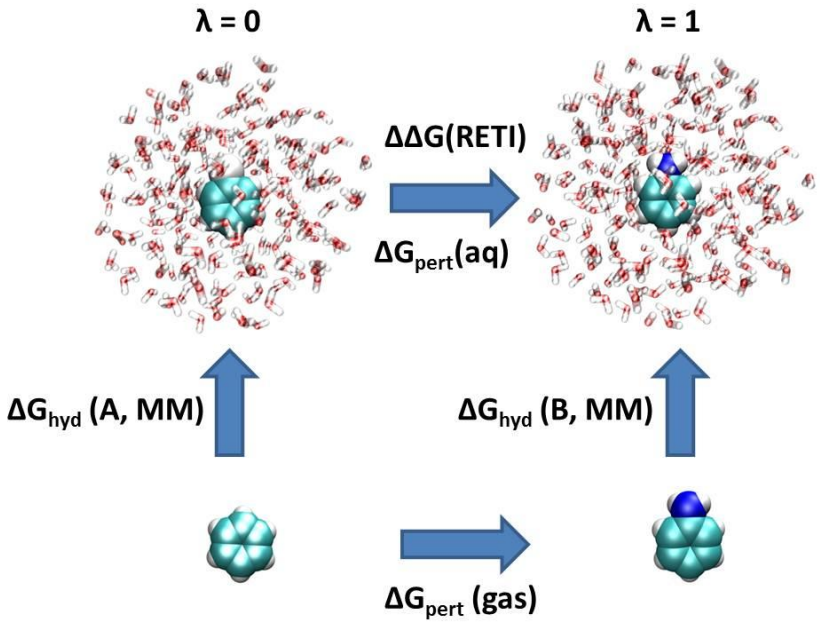


Figure 4.1: A MM hydration free energy cycle for perturbing benzene to aniline. This figure was generated using VMD v1.8.6.

As free energy is a state function, the cycle shown in Figure 4.1 must close, no matter how we travel around the cycle (equation 4.1). Crucially this enables us to extract free energies of interest. For this particular example, we want to know the free energy difference of perturbing benzene into aniline. If we close our free energy cycle we obtain:

$$0 = \Delta G_{pert}(aq) - \Delta G_{hyd}(B, MM) - \Delta G_{pert}(gas) + \Delta G_{hyd}(A, MM) \quad (4.1)$$

In equation 4.1, $\Delta G_{hyd}(A, MM)$ corresponds to the free energy of hydration for the λ=0 state (benzene), $\Delta G_{hyd}(B, MM)$ is the free energy of hydration for the λ=1 state (aniline) and $\Delta G_{pert}(aq)$ and $\Delta G_{pert}(gas)$ denote the free energy differences between the two endpoints. Rearranging equation 4.1 yields:

$$\Delta G_{hyd}(B, MM) - \Delta G_{hyd}(A, MM) = \Delta G_{pert}(aq) - \Delta G_{pert}(gas) \quad (4.2)$$

$$\Delta\Delta G(RETI) = \Delta G_{hyd}(B, MM) - \Delta G_{hyd}(A, MM) = \Delta G_{pert}(aq) - \Delta G_{pert}(gas) \quad (4.3)$$

where $\Delta\Delta G(RETI)$is the free energy difference between our two end states (A and B) using the RETI method, equations 2.62 and 2.64. From equations 4.2 and 4.3, it is clear that via the alchemical perturbation between λ=0 and λ=1 in both explicit solvent and in vacuum we can obtain $\Delta G_{pert}(aq)$ and $\Delta G_{pert}(gas)$ respectively, which will enable the prediction of the relative MM hydration free energy change between λ=0 and λ=1.

Subsequently, configurations are taken from each endpoint of the classical MM simulation (λ = 0 and λ = 1) and use these as inputs for DFT – QM/MM single point energy calculations. In the QM/MM representation of our system, any waters within the MM cut-off distance are represented as point charges around the DFT defined ligand. Hence, EE is used to allow our QM region (our ligand) to become polarized by our MM charges (Figure 4.2).

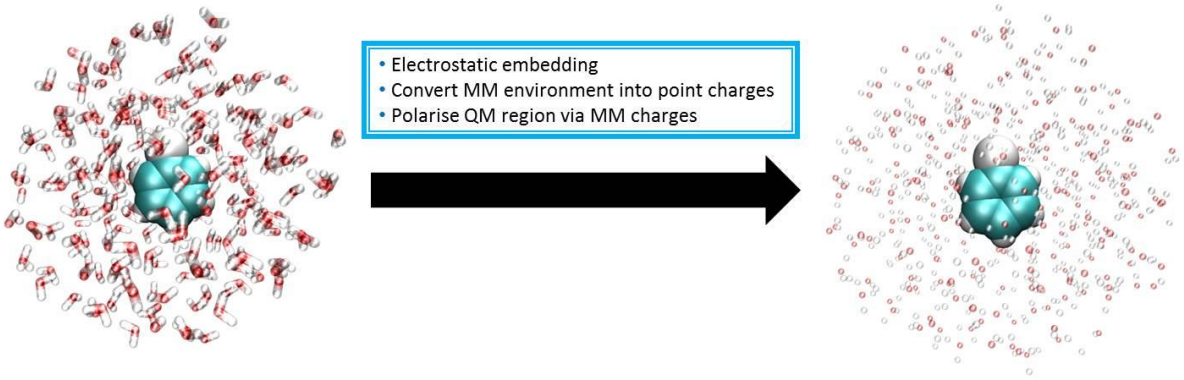


Figure 4.2: The EE approach used in this QM/MM method, where MM point charges (red and white spheres) are embedded around the QM ligand (benzene). This figure was generated using VMD v1.8.6.

Performing DFT – QM/MM single point energy calculations enables the addition of QM/MM legs to the thermodynamic cycle which reflect the free energy difference between a classical MM and a QM/MM representation of the system (Figure 4.3).

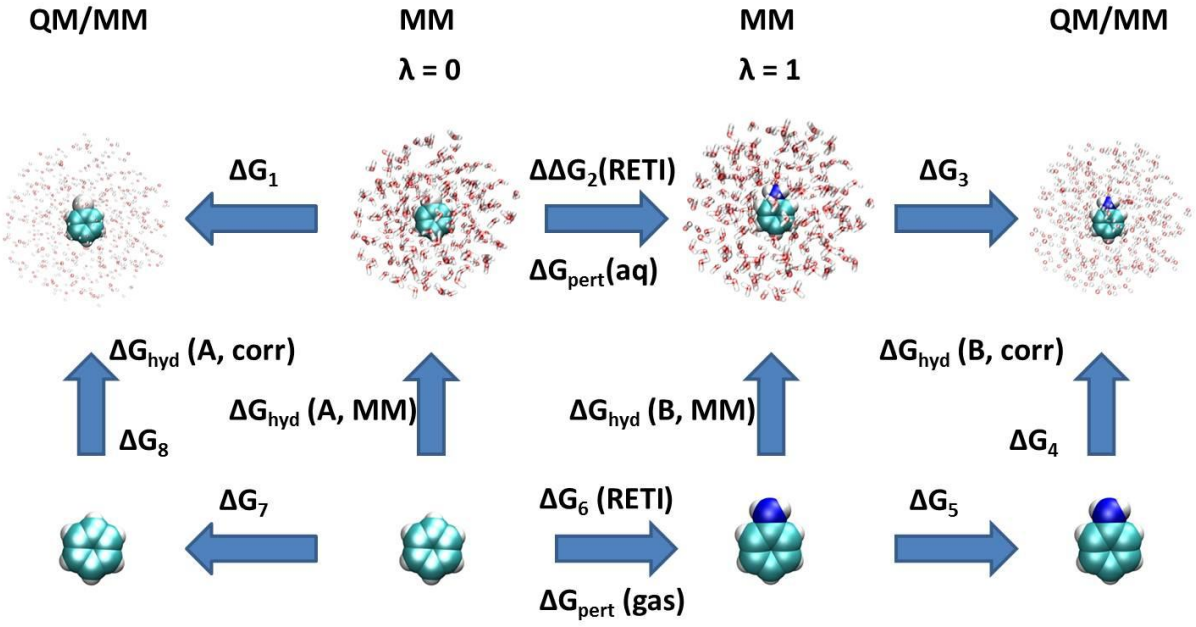


Figure 4.3: A MM→QM/MM hydration free energy cycle for perturbing benzene to aniline. This figure was generated using VMD v1.8.6.

The MM perturbation free energy ΔG_2 is calculated by RETI in the classical MM part of the protocol. The corresponding gas phase free energy change between the two endpoints of our perturbation is ΔG_6. ΔG_1 and ΔG_3 describe the difference in the solute-solvent interaction between a QM/MM and a pure MM representation of the system. These two correction terms are calculated using the formulation of the Zwanzig equation, equation 4.7. These free energies represent FEP energies for the transition MM→QM/MM, which is done in a single step (i.e. without intermediate λ states). From Figure 4.3 we can deduce the MM→QM/MM correction terms $\Delta G_{hyd}(A, corr)$and $\Delta G_{hyd}(B, corr)$:

$$0 = -\Delta G_1 + \Delta G_2 + \Delta G_3 - \Delta G_4 - \Delta G_5 - \Delta G_6 + \Delta G_7 + \Delta G_8 \quad (4.4)$$

$$\Delta G_4 - \Delta G_8 = -\Delta G_1 + \Delta G_2 + \Delta G_3 - \Delta G_5 - \Delta G_6 + \Delta G_7 \quad (4.5)$$

$$\begin{aligned} \Delta G_{hyd}(B, corr) - \Delta G_{hyd}(A, corr) &= \Delta G_4 - \Delta G_8 \\ &= -\Delta G_1 + \Delta G_2 + \Delta G_3 - \Delta G_5 - \Delta G_6 + \Delta G_7 \end{aligned} \quad (4.6)$$

The MM→QM/MM correction terms ($\Delta G_1$and ΔG_3) are calculated for each endpoint of our perturbations (λ = 0 and λ = 1) using the Zwanzig equation:

$$\Delta G_{MM\rightarrow QM/MM} = -RT \ln exp\langle \frac{-U_{QM/MM} - U_{QM,vac} - U_{MM,charges} - U_{MM,coul,sol-solv}}{RT} \rangle_{MM} \quad (4.7)$$

In equation 4.7, $U_{QM/MM}$ is the total energy as obtained from a QM/MM calculation with background charges that polarise the QM wavefunction (the solute serves as the QM part, the background charges as MM part of the QM/MM system). $U_{QM,vac}$ is the single point energy of the solute (the QM part) in vacuum, $U_{MM,charges}$ is the sum of all Coulomb interactions of the background charges. $U_{MM,coul,sol-solv}$ is the Coulomb solute-solvent interaction energy as obtained from the MM part of the protocol. Therefore, the solute-solvent energy for an MM treatment of the system is subtracted from the corresponding solute-solvent interaction energy as obtained from QM/MM.

Unlike standard FEP implementations no intermediate λ states between MM and QM/MM representations of the system, and the phase space is only sampled using the MM Hamiltonian. Normally, the phase space would also be sampled using the QM/MM Hamiltonian, and would calculate $\Delta G_{QM/MM\rightarrow MM}$, which should be similar in value as $\Delta G_{MM\rightarrow QM/MM}$. Therefore, the approach taken here is an approximation which uses a simple post-processing of classically obtained MM ensembles by selecting a certain percentage of configurations for QM/MM single point calculations.

As our approach neglects any QM/MM sampling, we subsequently check the pathway-independence of the free energies obtained. This is achieved through the use of charge perturbation pathways (Figure 4.4) where the original MM charges are perturbed using an arbitrary scaling factor (SF).

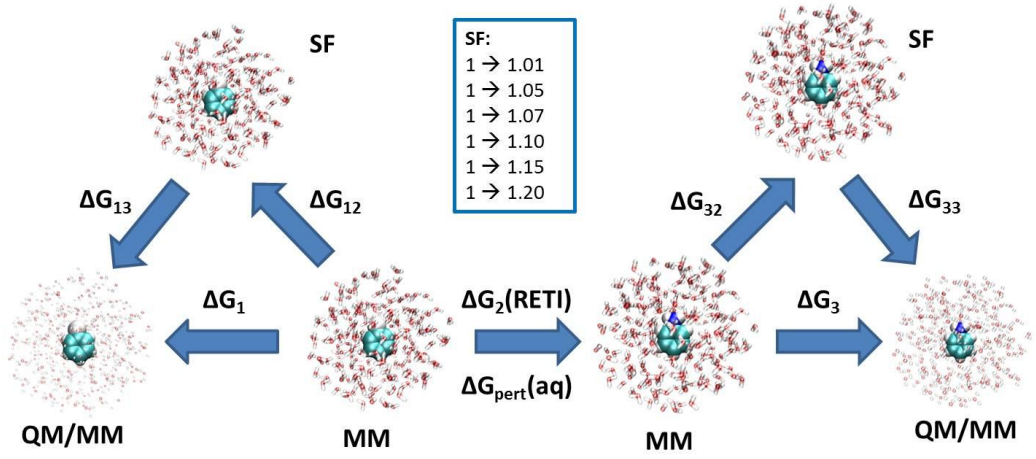


Figure 4.4: A charge perturbation free energy cycle for benzene to aniline, where the charges on each solute are scaled by a SF and the response measured in both MM and QM/MM. This figure was generated using VMD v1.8.6.

The free energy for the perturbation from original to perturbed charges within the MM ensemble (ΔG_{12} and ΔG_{32}) should be cancelled out via the free energy change in the QM/MM ensemble (ΔG_{13} and ΔG_{33}):

$$\Delta G_1 = \Delta G_{12} + \Delta G_{13} \tag{4.8}$$

$$\Delta G_3 = \Delta G_{32} + \Delta G_{33} \tag{4.9}$$

If the resultant free energies from our charge perturbed MM ($\Delta G_{12}$and ΔG_{32}) and QM/MM ($\Delta G_{13}$and ΔG_{33}) simulations equal the original QM/MM corrections ($\Delta G_1$and ΔG_3) in equations 4.8 and 4.9, then our calculations have met the minimum condition to ensure our calculations are working correctly. Furthermore, this validation will also

reveal if our QM/MM ensemble gives internally consistent results, which again can be considered as a minimum requirement for our simulations.

4.2 Protein-Ligand Binding Free Energies

Of particular interest in the field of drug design is the ability to predict the strength and specificity with which a molecule binds to a target macromolecule. Many drug molecules function by binding to the active site of a particular enzyme so strongly that the natural substrate for the enzyme cannot bind and as a result some particular biological pathway is stalled. Multiple algorithms exist [115, 116] which concern the placement of drug molecules into the active site of enzymes. This process is commonly termed 'docking'. These algorithms can reproduce experimentally known binding modes with very good efficiency [117], unfortunately they tend to obtain a poor relationship between predicted and experimental binding affinities [118]. Hence, more rigorous free energy techniques are needed to obtain accurate protein-ligand binding affinities. In recent years, multiple studies have shown that with the use of rigorous free energy methods, one can obtain protein-ligand binding affinities that are within ±1 kcal.mol^{-1} of experimentally observed affinities [11, 12, 13].

To calculate the protein-ligand binding free energy between two end point states (λ=0 and λ=1) we must construct a free energy cycle (Figure 4.5) where both endpoints are simulated in both bound and unbound (aqueous) states:

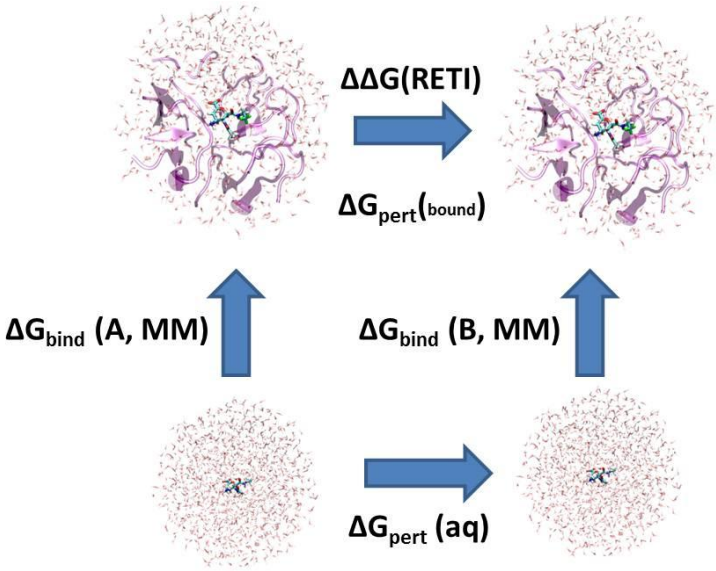


Figure 4.5: A MM protein-ligand binding free energy cycle. This figure was generated using VMD v1.8.6.

As free energy is a state function, the cycle shown in Figure 4.5 must close, no matter how we travel around the cycle. Crucially this enables us to extract free energies of interest. If we close our protein-ligand binding free energy cycle we obtain:

$$0 = \Delta G_{pert}(bound) - \Delta G_{bind}(B, MM) - \Delta G_{pert}(aq) + \Delta G_{bind}(A, MM) \quad (4.10)$$

In equation 4.10, $\Delta G_{bind}(A, MM)$ corresponds to the binding free energy for the λ=0 state, $\Delta G_{bind}(B, MM)$ is the binding free energy for the λ=1 state and $\Delta G_{bind}(bound)$ and $\Delta G_{bind}(aq)$ denote the free energy changes between the two endpoints. Rearranging equation 4.10 yields:

$$\Delta G_{bind}(B, MM) - \Delta G_{bind}(A, MM) = \Delta G_{pert}(bound) - \Delta G_{pert}(aq) \quad (4.11)$$

$$\Delta\Delta G(RETI) = \Delta G_{bind}(B, MM) - \Delta G_{bind}(A, MM) \quad (4.12)$$
$$= \Delta G_{pert}(bound) - \Delta G_{pert}(aq)$$

where $\Delta\Delta G(RETI)$is the free energy difference between our two end states (A and B) using the RETI method, equations 2.62 and 2.64. From equations 4.11 and 4.12, it is clear that via the alchemical perturbation between λ=0 and λ=1 in both bound and free states enables $\Delta G_{bind}(bound)$ and $\Delta G_{bind}(aq)$ to be obtained respectively, which will enable the prediction of the relative MM protein-ligand binding free energy change between λ=0 and λ=1, $\Delta\Delta G(RETI)$.

As with the hydration free energies, configurations from each endpoint of the classical MM simulation (λ = 0 and λ = 1) are used as inputs for DFT – QM/MM single point energy calculations. In the QM/MM representation of the system any protein residues and waters within the MM cut-off distance are represented as point charges around the DFT/QM defined ligand. Hence, this method uses EE electrostatic to allow the QM region (our ligand) to become polarized via the MM charges (Figure 4.6).

Figure 4.6: The EE approach used in our protein-ligand binding free energy cycle. Where the MM environment is embedded using point charges around our QM solute. This figure was generated using VMD v1.8.6.

Performing DFT – QM/MM single point energy calculations enables us to add additional legs to our thermodynamic cycle which reflect the free energy difference between a classical MM and a QM/MM representation of the system (Figure 4.7).

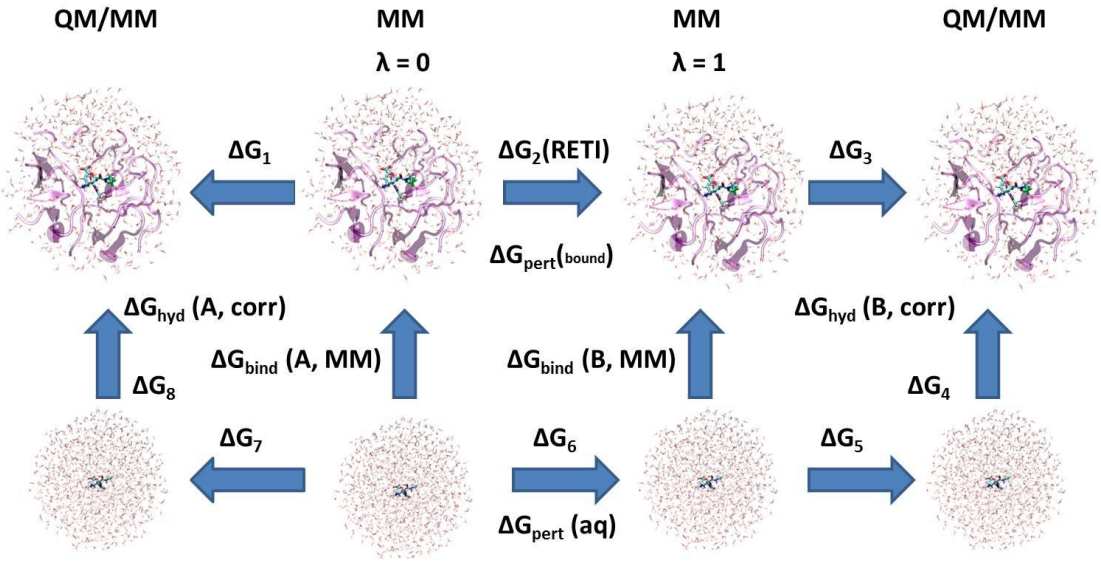


Figure 4.7: A MM→QM/MM free energy cycle for a protein-ligand system. This figure was generated using VMD v1.8.6.

In figure 4.7, ΔG_2 denotes the free energy change for bound leg of the MM free energy simulation, ΔG_6corresponds to the free energy change for free (aqueous) leg of the MM free energy simulation. ΔG_1 and ΔG_3 are the QM/MM correction free energies for the λ=0 and λ1 bound states respectively. ΔG_7 and ΔG_5 are the QM/MM correction free energies for the λ=0 and λ1 free states respectively. From Figure 4.7 we can deduce the MM→QM/MM correction terms $\Delta G_{bind}(A, corr)$and $\Delta G_{bind}(B, corr)$:

$$0 = -\Delta G_1 + \Delta G_2 + \Delta G_3 - \Delta G_4 - \Delta G_5 - \Delta G_6 + \Delta G_7 + \Delta G_8 \quad (4.13)$$

$$\Delta G_4 - \Delta G_8 = -\Delta G_1 + \Delta G_2 + \Delta G_3 - \Delta G_5 - \Delta G_6 + \Delta G_7 \quad (4.14)$$

$$\Delta G_{bind}(B, corr) - \Delta G_{bind}(A, corr) = \Delta G_4 - \Delta G_8 \quad (4.15)$$
$$= -\Delta G_1 + \Delta G_2 + \Delta G_3 - \Delta G_5 - \Delta G_6 + \Delta G_7$$

The MM→QM/MM bound leg correction terms ($\Delta G_1$and ΔG_3) are calculated for each endpoint of the perturbations (λ = 0 and λ = 1) using the Zwanzig equation described below:

$$\Delta G_{MM \to QM/MM} \quad (4.16)$$
$$= -RT \ln exp \left\langle \frac{-U_{QM/MM} - U_{QM,vac} - U_{MM,charges} - U_{MM,coul,solute-solv/protein,MM}}{RT} \right\rangle_{MM}$$

In equation 4.16, $U_{QM/MM}$ is the total energy as obtained from a QM/MM calculation with background charges that polarizes the QM wave function (the solute serves as the QM part, the background charges as MM part of the QM/MM system). $U_{QM,vac}$ is the single point energy of the solute (the QM part) in vacuum, $U_{MM,charges}$ is the sum of all Coulomb interactions of the background charges. $U_{MM,coul,solute-solv/protein,MM}$ is the Coulomb solute-solvent and solute-protein interaction energies as obtained from the MM part of the protocol. Therefore, the solute-solvent/solute protein energy for an MM treatment of the system is subtracted from the corresponding solute-solvent/solute protein interaction energy as obtained from QM/MM in the exponential

term of the Zwanzig equation. The MM→QM/MM free leg correction terms ($\Delta G_7$and ΔG_5) are calculated for each endpoint of the perturbations (λ = 0 and λ = 1) using the Zwanzig equation described previously for the hydration free energies (equation 4.7).

As this approach neglects any QM/MM sampling, we subsequently check the pathway-independence of the free energies obtained. This is achieved through the use of charge perturbation pathways (Figure 4.8) where the original MM charges are perturbed using an arbitrary SF.

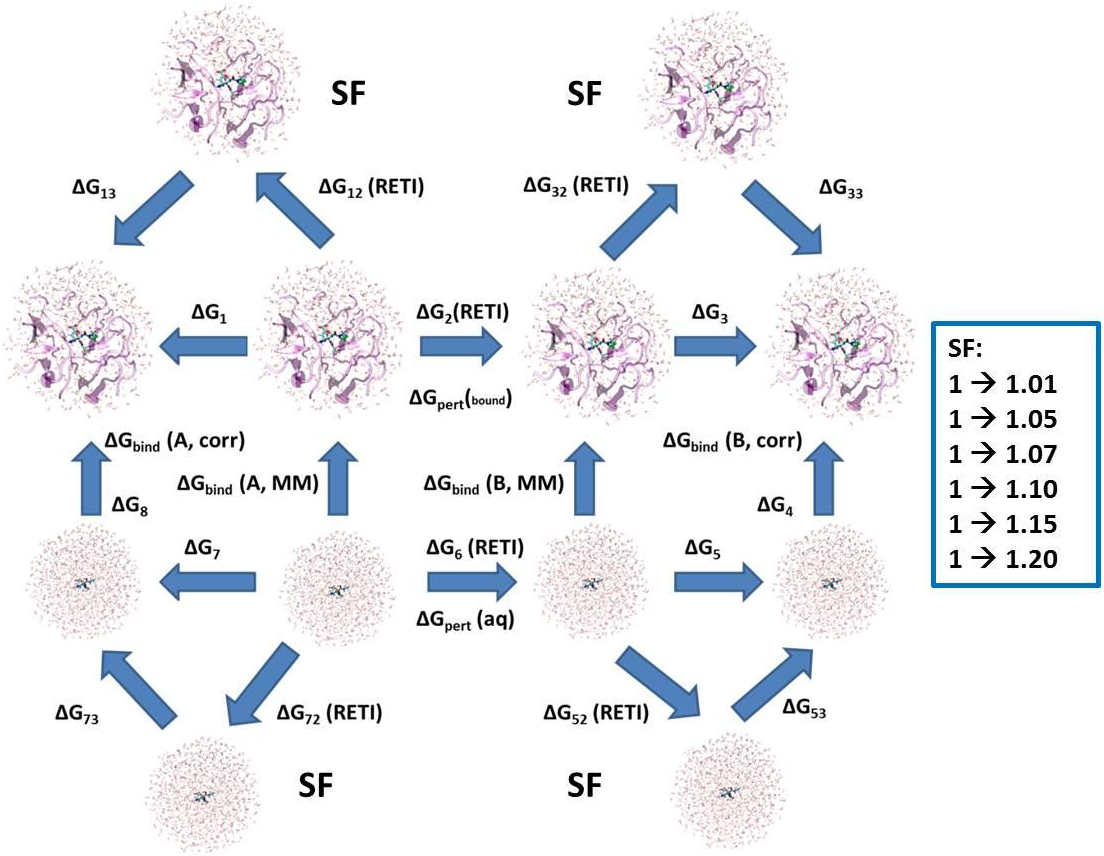


Figure 4.8: A charge perturbation free energy cycle for a protein-ligand system. This figure was generated using VMD v1.8.6.

The free energy for the perturbation from original to perturbed charges within the MM ensemble for the bound (ΔG_{12} and ΔG_{32}) and free (ΔG_{52} and ΔG_{72}) legs should be

cancelled out via the free energy change to the QM/MM ensemble for the bound (ΔG_{13} *and* ΔG_{33}) and free (ΔG_{53} and ΔG_{73}) legs repectively:

$$\Delta G_1 = \Delta G_{12} + \Delta G_{13} \tag{4.17}$$

$$\Delta G_3 = \Delta G_{32} + \Delta G_{33} \tag{4.18}$$

$$\Delta G_5 = \Delta G_{52} + \Delta G_{53} \tag{4.19}$$

$$\Delta G_7 = \Delta G_{72} + \Delta G_{73} \tag{4.20}$$

4.3 Conclusions

In this chapter, an outline of the thermodynamic process which allows the calculation of both MM→QM/MM hydration free energies and protein ligand binding free energies has been described. The calculation of QM/MM correction terms will allow the correction of classically obtained MM free energies and is hoped to produce more accurate free energy estimates. The methodology presented here is, however, very approximate as phase space is only sampled using our MM Hamiltonian. This approximation allows the method to be very fast, but it does mean that the reverse transformations between states cannot be performed, which conventional methods that do sample phase space with the QM/MM Hamiltonian can. As this method lacks

QM/MM sampling, charge perturbations are performed to ensure the pathway-independence of the free energies obtained.

5 Hydration Free Energy Study - Small Neutral Organic Molecules

5.1 Introduction

The calculation of hydration free energies has become a standard test for MM forcefields. A number of studies [119, 120, 121] have investigated the calculation of hydration free energies based on alchemical free energy calculations. Here we present a study into the calculation of MM→QM/MM relative hydration free energies for a dataset of 110 small organic molecules. This dataset was taken from a study by Mobley and co-workers, where a molecular dynamics based approach was used to calculate absolute hydration free energies for 504 small neutral organic molecules [122]. This study was performed using the MD package AMBER with the GAFF [14]/AM1-BCC [125] methods used for parameterisation of the solute within a box of TIP4P water molecules. In this original study the authors reported an R^2 = 0.94 between calculated and experimental absolute hydration free energies. Furthermore, they were also able to identify poorly performing functional groups and propose adjusted MM parameter sets to correct for these systematic errors. Hence, we chose to use a subset of structurally similar molecules from these data to understand any limitations in our QM/MM methodology at predicting hydration free energies, by comparing the results of our QM/MM hydration free energy study with our MM results and those reported within the literature we aim to understand any limitations in our QM/MM method compared to readily available MM methods, for example GAFF, which are well

parameterized for this particular problem. Furthermore, this study will also serve as an essential yardstick in this methods applicability to drug design.

5.1.1 Methods

Dataset

The 114 compounds chosen for this study exhibit a wide range of functional groups including, alcohols, alkyl-halides, amines, carboxylic acids, aldehydes, amides, ethers, esters, thioethers, thiols, cyanos, ketones, alkanes and alkenes for both aliphatic and aromatic molecules. To calculate hydration free energies for these compounds the MC simulation package ProtoMS v2.2 was used [123]. Within ProtoMS it is possible to calculate the relative hydration free energies for alchemical perturbations between structurally related compounds. Therefore, it was necessary to first constructed perturbation webs (Figure 5.1) of structurally related compounds to calculate the MM→QM/MM relative hydration free energies for each perturbation.

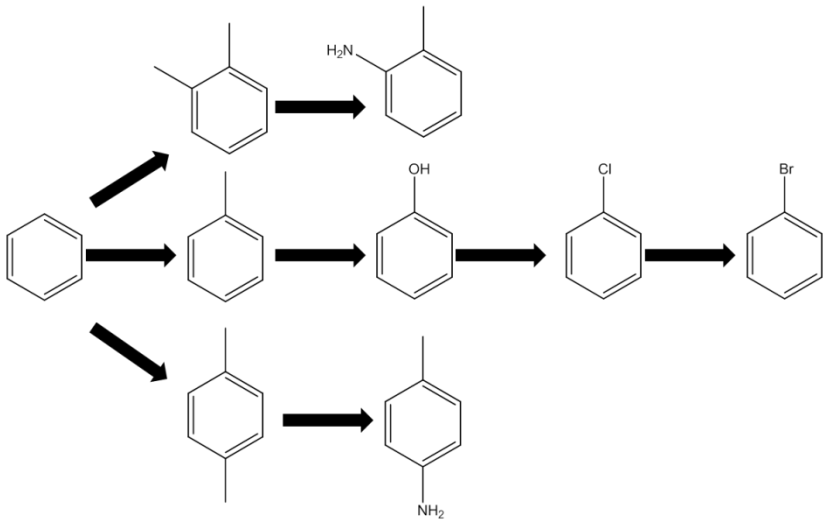


Figure 5.1: An example of perturbation web from the dataset of 114 compounds chosen for this hydration free energy study. In this example benzene is being perturbed to multiple toluene based molecules.

A total of 110 MM→QM/MM relative hydration free energies were calculated utilising 7 different perturbation webs. Each web is directed back to one of four reference molecules; methane, cyclopentane, cyclohexane and benzene. The absolute MM→QM/MM hydration free energies are calculated by annihilating these reference compounds (section 5.1.3) then using the free energies obtained from these annihilations to move back through our perturbation webs. A full list of each branch of our perturbation web can be found in Supporting Information 1 Figures 1.1 – 1.11.

Ligand setup

The mol2 structures for all 114 compounds were downloaded from http://pubs.acs.org/doi/suppl/10.1021/ct800409d. The hydrogen atoms are already assigned to these molecules. The ligands were parameterised using the GAFF forcefield [14] and partial atomic charges were derived from the AM1-BCC method [124] as implemented in the AMBER 10 suite. To relax the geometry of each ligand we minimised each structure in the SANDER module of the AMBER 10 suite with a Generalised Born model. The minimised ligand structures were solvated in a 40 x 40 x 40$Å^3$ TIP4P [125] water box using the LEaP module in AMBER.

Monte Carlo Simulation Protocol

The bond angles and torsions for the ligands were sampled during the simulation, with aromatic ring structures being the only exception. The bond lengths of the ligand were constrained. A 12 Å based cut-off from the ligands centre of mass was employed in all simulations.

For simulation in the aqueous state, solvent moves were attempted with a probability of 85.28%, and solute moves with a probability of 15.72.%. Replica exchange moves between adjacent values of λ were attempted every 200000 moves. The solvent was equilibrated for 20 million moves to remove any bad contacts with the solute. The system was then equilibrated at one state (the end state with the larger solute) for 10 million further moves where solute and solvent moves were attempted. The resulting configuration was distributed over the 16 values for the coupling parameter λ (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82, 0.88, 0.94, and 1.00) to allow smooth transition between the two end states. The system was then equilibrated for 10 million moves before collecting statistics for 40 million moves. All simulations were performed within the NPT ensemble.

QM Single Point Energy Protocol

Configurations from the endpoints (λ=0 and λ=1) of the classical free energy simulations were selected and used as input for DFT-QM/MM single point energy calculations with Gaussian 09 [126]. One QM single point energy calculation with background charges representing the solvent within the cut-off (Gaussian keyword 'CHARGE') were performed every 100000th MM MC moves, with symmetry operations disabled (Gaussian keyword 'NoSymm'). This gave a total of 400 QM/MM single points for each solute perturbation. Gaussian calculations with embedded background charges allow a polarisation of the QM wave function via the MM charges; however no back polarisation of the MM part via the polarised QM wave function was considered.

The QM energies were computed using the B3LYP and BLYP hybrid density functionals calculations with a range of basis sets; 6-31G*, 6-31G(d,p), 6-31G++, AUG-cc-pVDZ and AUG-cc-pVTZ, as implemented in Gaussian 09.

As the ligands were flexible it was necessary to compute the QM vacuum single point energies for each snapshot used. This was performed in Gaussian 09, but without the use of the 'CHARGE' and 'NoSymm' keywords, which are only necessary if embedding MM point charges in our calculation.

The QM vacuum energies were computed using the B3LYP and BLYP hybrid density functionals calculations with a range of basis sets; 6-31G*, 6-31G(d,p), 6-31G++, AUG-cc-pVDZ and AUG-cc-pVTZ, as implemented in Gaussian 09.

5.1.2 Results and Discussion – MM – RETI Results

The initial MM hydration free energy calculations were performed and analysed to compare our results to those from experiment and to results published in the previous MM study on this system. When comparing the results to those obtained in previous work (Figure 5.2) we found an excellent agreement (R^2 = 0.98) between our calculated MM relative hydration free energies and data generated from the Mobley *et al.* study [122]. The relevant free energies are summarised in Table 1.1 of Supporting Information 1.

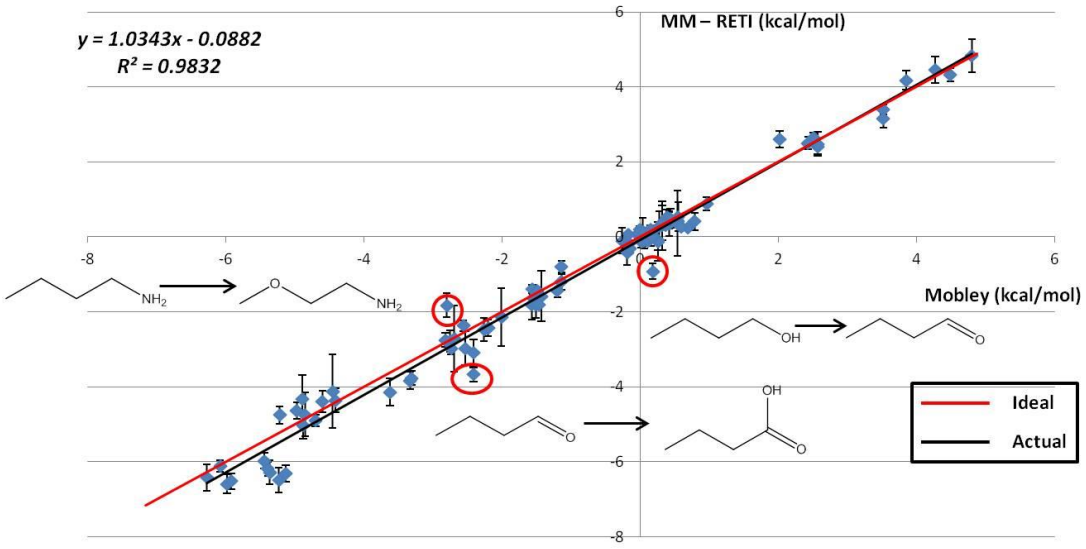


Figure 5.2: MM-RETI versus data generated by Mobley *et al.* [122] relative hydration free energies. The ideal correlation (1 to 1) is shown by the red line, whilst the actual correlation is shown via the black line. The error bars shown are calculated from four independent MM-RETI simulations using standard error.

The coefficient of determination (R^2) for our computed free energies compared to data from Mobley *et al.* is 0.98. This agreement is very good, however there are three small outliers between the two datasets. These are for long chained compounds which can form intra-molecular hydrogen bonds. It is suspected that our choice of MC is cauing these small discrepancies as these compounds do become locked into conformations which are highly energetically favourable when using MC rather than MD as Mobley *et al.* used. It is common to also compute the Mean Unsigned Error (MUE) for our calculated data versus experiment. This enables us to analyse the predictive error across the entire dataset. The MUE is calculated as follows:

$$MUE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (5.1)$$

In equation 5.1, n is the number of observations with a predicted value, f_i, and an experimental value y_i. For this dataset a MUE of 0.14 kcal.mol^{-1} is obtained. This excellent agreement is not unexpected as the two datasets use identical forcefields, charge models and water models for each system studied.

When comparing the results to experimental data [122] (Figure 5.3) an excellent agreement (R^2 = 0.92) between our calculated MM relative hydration free energies and experiment was found. The relevant free energies are summarised in Table 1.1 of Supporting Information 1.

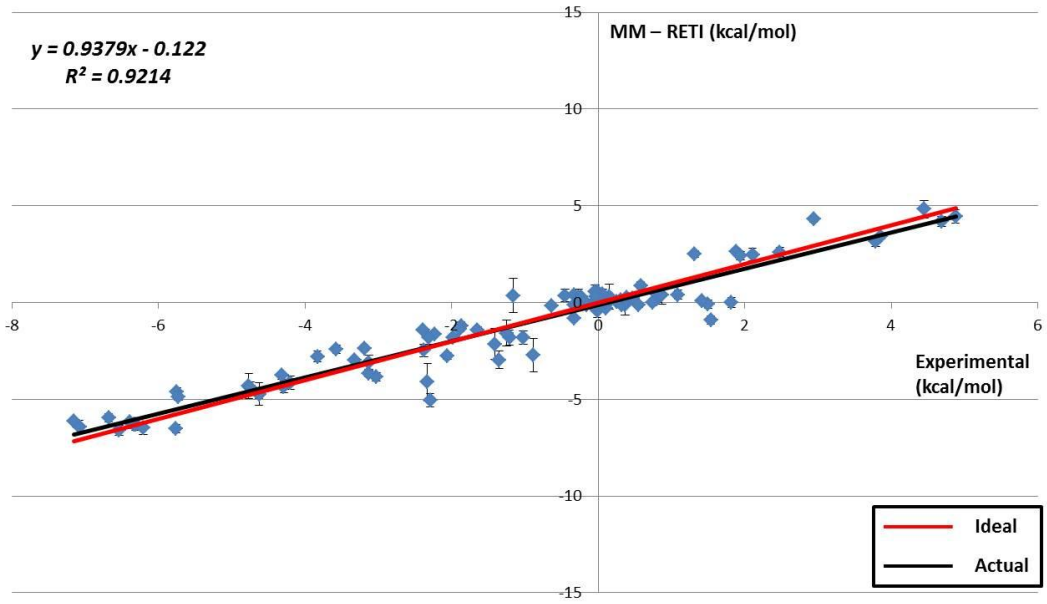


Figure 5.3: MM-RETI versus experimental [122] relative hydration free energies. The ideal correlation (1 to 1) is shown by the red line, whilst the actual correlation is shown via the black line. The error bars shown are calculated from four independent MM-RETI simulations using standard error.

The coefficient of determination (R^2) for our computed free energies compared to experiment is 0.92. For this dataset a MUE of 0.64 kcal.mol^{-1} is obtained. This good agreement is not unexpected as MM forcefields, like GAFF, are parameterised very accurately for these types of small organic molecules.

The free energy gradients for each perturbation were analysed (Figure 5.4) to ensure a smooth transition across the reaction co-ordinate (λ).

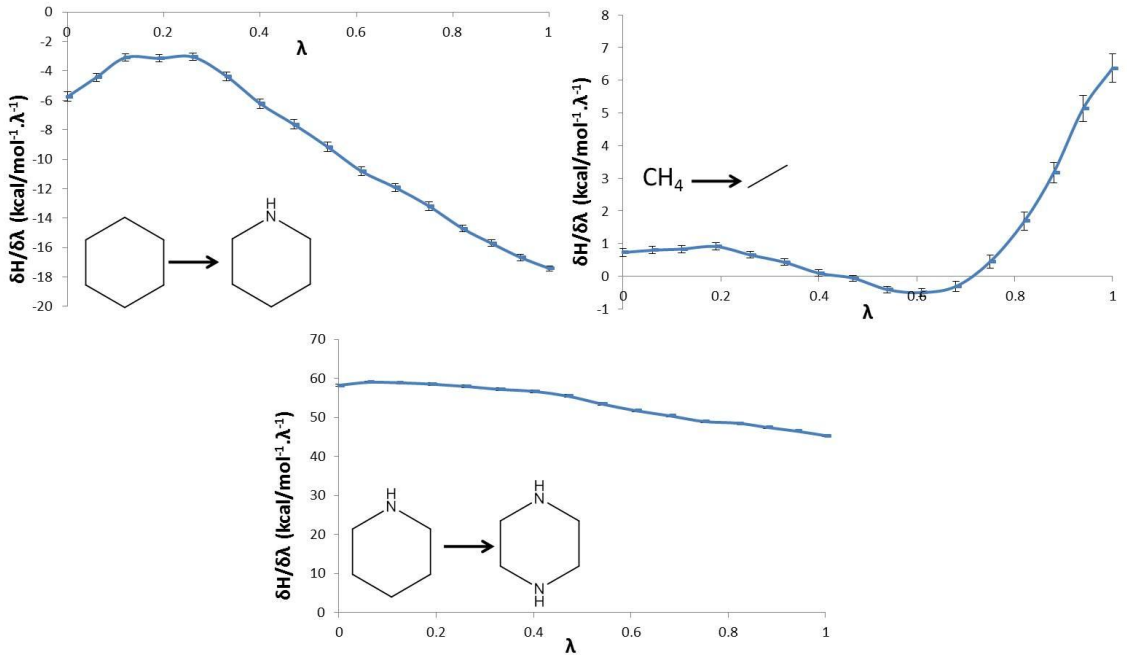


Figure 5.4: A set of free energy gradients for several perturbations, including; cyclohexane → piperidine, methane → ethane, and piperidine → piperazine. The error bars shown are calculated from four independent MM-RETI simulations using standard error.

Analysis of these free energy gradients shows that we do obtain smooth transition between the two end states of each system studied, which gives us confidence that the free energies obtained during this study are precise.

To compare the MM-RETI results with experiment is a standard method to draw conclusions regarding the MM method employed here. However, the experimental

data used by Mobley *et al.* [122] was experimental data collected prior to 2009. A more comprehensive collection of experimental small molecule hydration free energies was obtained from Prof. Peter Guthrie, the curator of multiple hydration free energy test sets [127]. A comparison between these two sets of experimental data for the 110 relative hydration free energies that were computed is shown in Figure 5.5.

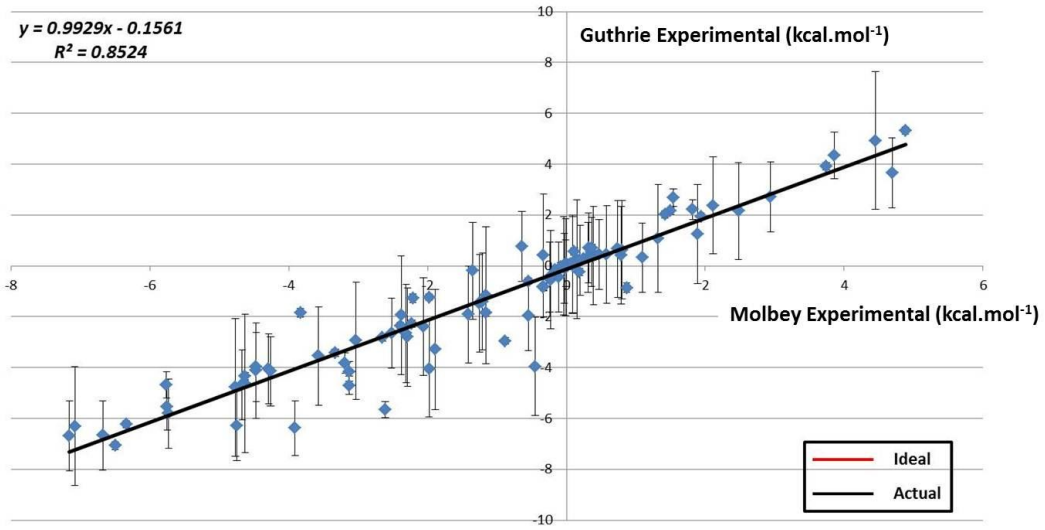


Figure 5.5: Comparison of Guthrie's experimental [127] versus Mobley's experimental data [122]. The error bars shown are taken from Guthrie's dataset, no error bars are shown for Mobley's dataset as no errors are provided with this experimental dataset.

This comparison shows that the two sets of experimental data show a coefficient of determination equal to 0.85 with a slope of nearly 1. This analysis suggests that a coefficient of determination between the MM-RETI calculated relative hydration free energies and experiment of 0.92 for Mobley *et al.* [122] and 0.85 for Guthrie [127] (Table 5.1) is as accurate as two independent experimental datasets and hence the results from the MM-RETI free energy study cannot be expected to perform any better than this. The comparison of our free energies to both experimental datasets is shown below (Table 5.1).

	R^2	Slope	MUE (kcal.mol^{-1})
MM vs. Guthrie	0.85	0.85	0.81
MM vs. Mobley	0.92	0.94	0.64

Table 5.1: Comparison of coefficient of determination (R^2), slopes and MUE for MM-RETI results versus Guthrie [127] and Mobley [122] experimental datasets.

Despite the excellent agreement between our computed MM free energies and experiment, there are several noticeable outliers. If a value of ≥1 kcal.mol^{-1} is taken between the computed hydration free energies and Guthrie's experimental free energies to be the definition of an outlier; then 30 of the 110 perturbations fall into this category (Figure 5.6).

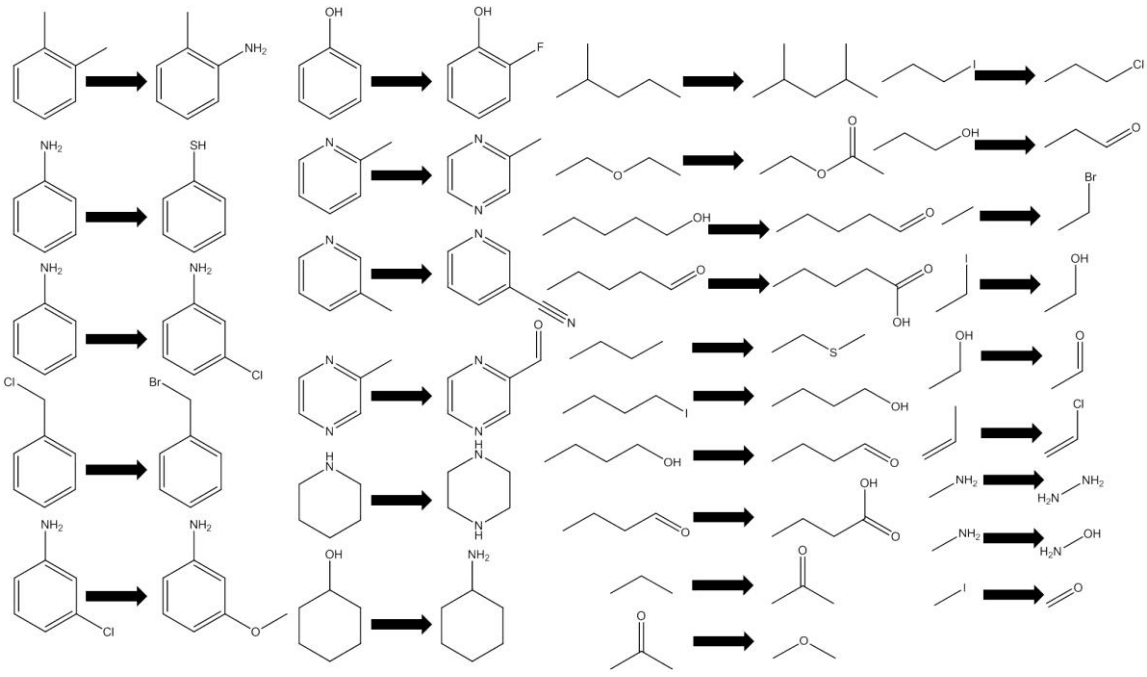


Figure 5.6: Outliers from the MM-RETI hydration free energy study when compared to Guthrie's experimental data [127].

Investigation of the outliers highlights the poor performance of MM forcefields for polar compounds. This is indicative of the lack of accuracy due to the neglect of polarisation terms within standard MM forcefields and hence it is hoped that the MM→QM/MM free energies can improve the predicted hydration free energies for such compounds.

5.1.3 MM→QM/MM Results

Configurations from the endpoints of the classical simulations were then used as input for DFT-QM/MM single point energy calculations in Gaussian 09. The resulting QM/MM correction free energies were calculated and the overall QM/MM relative hydration free energies were determined. Comparing these results to Guthrie's experimental data (Figure 5.7) shows a similar agreement to the classically obtained results; however, there are a far greater number of outliers. The relevant free energies are summarised in Table 1.2 of Supporting Information 1.

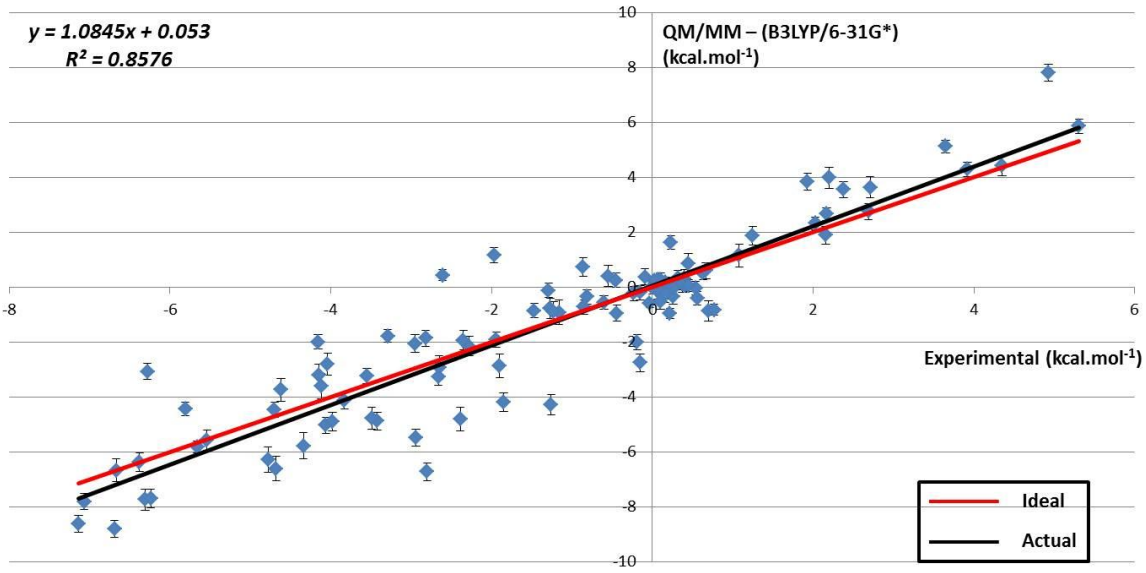


Figure 5.7: MM→QM/MM free energies versus Guthrie experimental data [127]. The red line represents ideal correlation (1 to 1), whilst the black line represents the actual correlation. The error bars shown are calculated from four independent simulations using standard error.

The coefficient of determination (R^2) is equal to 0.85 with a MUE of 0.86 kcal.mol^{-1}. A summary of the results compared to both experimental datasets is shown below (Table 5.2).

	R^2	**Slope**	**MUE (**kcalmol^{-1}**)**
QM/MM vs. Guthrie	0.85	1.06	0.86
QM/MM vs. Mobley	0.85	1.13	0.83

Table 5.2: Comparison of coefficient of determination (R^2), slopes and MUE for MM→QM/MM results versus Guthrie [127] and Mobley [122] experimental datasets.

This pattern of QM/MM relative hydration free energies producing good correlation, but with a larger number of outliers was consistent across the multiple number of DFT

functional / basis set combinations that were attempted to reproduce Guthrie's experimental data (Table 5.3).

QM Theory	R^2	Slope	MUE (kcalmol^{-1})
B3LYP/6-31G(d,p)	0.89	1.10	0.87
B3LYP/6-311G++	0.89	1.11	0.86
B3LYP/AUG-cc-pVDZ	0.88	1.09	0.88
B3LYP/AUG-cc-pVTZ	0.87	1.07	0.89
BLYP/6-31G*	0.85	1.07	0.85
BLYP/6-31G(d,p)	0.83	1.13	0.88
BLYP/6-311G++	0.84	1.17	0.86
BLYP/AUG-cc-pVDZ	0.89	1.06	0.84
BLYP/AUG-cc-pVTZ	0.90	1.07	0.83

Table 5.3: Comparison of coefficient of determination (R^2), slopes, and MUE for several combinations of DFT functional and basis sets for the 110 calculated relative QM/MM hydration free energy results compared to Guthrie experimental data [127].

Identification of the outlier perturbations showed that they all had similar properties; first they are mostly involved polar ligands, and secondly most have the ability to form hydrogen bonds with the solvent. Analysis into the QM/MM correction free energies was performed for several perturbation types: non-polar → non-polar, non-polar → polar, and polar → polar (Figure 5.7).

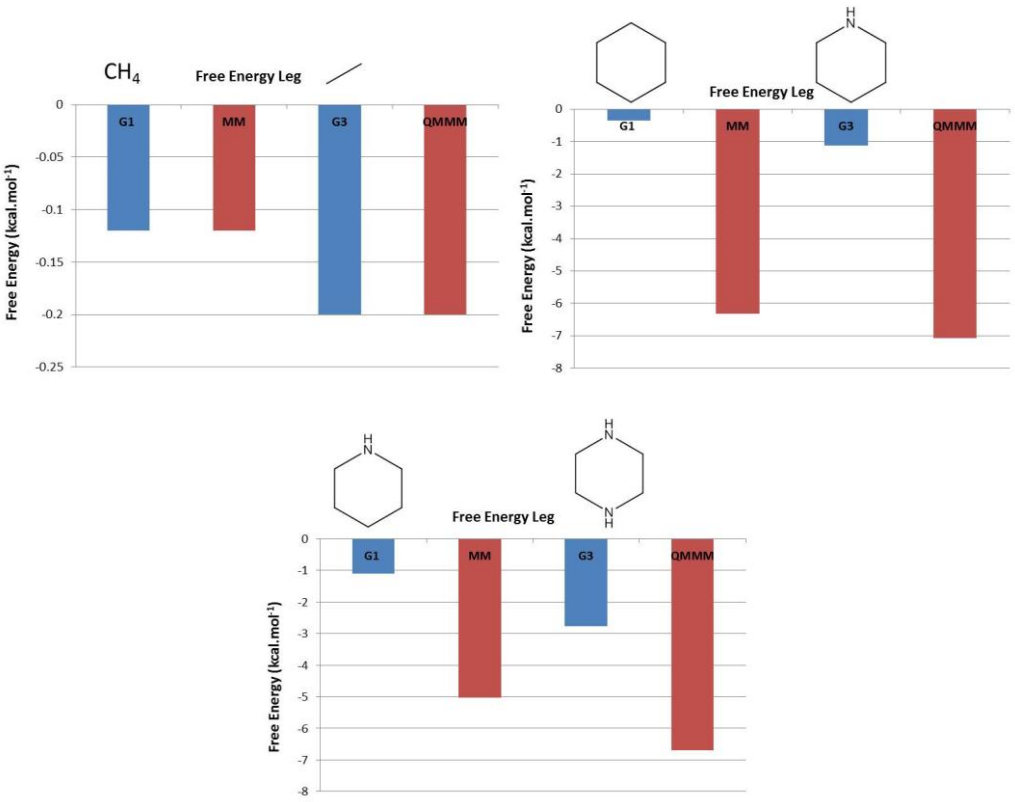


Figure 5.7: QM/MM energy breakdown for the hydration free energies of several perturbations, including; methane → ethane, cyclohexane → piperidine and piperidine → piperazine. G1 (blue bar) is the QM/MM correction free energy for λ=0, MM is the MM-RETI free energy (red bar), G3 (blue bar) is the QM/MM correction free energy for λ=1 and QM/MM is the MM→QM/MM free energy (red bar).

It is clear that for a perturbation involving two non-polar species (methane → ethane) there are only very small QM/MM correction free energies for both endpoints. This leads to a negligible difference in free energy between MM and QM/MM representations of the system. For a non-polar → polar case (cyclohexane → piperidine) there is a small QM/MM correction free energy equal to -0.35 kcal.mol^{-1} for the non-polar cyclohexane, but a significant QM/MM correction free energy of -1.12 kcal.mol^{-1} was obtained for the polar piperidine. This leads to difference of -0.77 kcal.mol^{-1} between MM and QM/MM. For the polar → polar example (piperidine →

piperazine) there are large and significant QM/MM correction free energies for both piperidine (-1.11 kcal.mol^{-1}) and piperazine (-2.77 kcal.mol^{-1}). This appears to indicate that this QM/MM approach is leading to large negative corrections for polar solutes. Investigation of the snapshots from each of these perturbations showed that TIP4P waters could be in very close proximity (< 2 Å) to ligands with H-bonding capabilities. Although this is fine in the MM ensemble, when we convert these snapshots into QM/MM we simply electrostatically embed the TIP4P waters as point charges. These MM charges very close to our ligand lead QM wavefunction to become overpolarized via the embedded MM charges. This in turn leads to the snapshots producing large MM → QM/MM corrections, which is the main driving force behind the outliers in our MM → QM/MM free energies. This is what our method intends to do, polarise the QM wavefunction via the MM environment, however in such cases it appears that this polarisation is too strong. This phenomenon has been identified in other QM/MM free energy studies [128, 129]. In studies by Beierlein *et al.* [111] and Woods *et al.* [110], this problem of overpolarization of the QM region via embedded MM charges was identified for the very simple perturbation of methane → TIP4P – water. These results were also corroborated calculations using the QM/MM method presented in this thesis (Table 5.4).

	MM (kcal.mol^{-1})	QM/MM-CH4 (kcal.mol^{-1})	QM/MM-TIP4P (kcal.mol^{-1})	QM/MM (kcal.mol^{-1})
Woods *et al.*	-8.90 (0.1)	-0.28 (0.01)	-1.38 (0.03)	-10.0 (0.1)
Beierlein *et al.*	-8.88 (0.1)	-0.17 (0.01)	-2.03 (0.04)	-10.74 (0.08)
This Thesis	-8.88 (0.1)	-0.02 (0.01)	-1.39 (0.07)	-10.25 (0.09)

Table 5.4: Comparison of MM and QM/MM free energies for the methane → TIP4P water perturbation from several studies

Having identified the embedding of the MM region within the QM/MM calculations as the main issue with the overall QM/MM corrections obtained, this leaves one remaining question: Can this overpolarisation of the QM/MM system via embedded charges be quantified? The simple answer is yes, it is possible to obtain the volume of the excess charge density that is generated by having an embedded charge close to a polar ligand compared to the ligand *in vacuo*. To do this the density of the ligand with and without the embedded MM point charges had to be calculated. Once this was performed, this density with embedded MM point charges was simply subtracted the density obtained *in vacuo*. The results showed that for polar ligands had a considerably larger density difference compared to non-polar ligands.

QM/MM Density Calculation Method

A set of QM/MM snapshots were taken and used for the QM/MM density calculations, which were performed using the cubegen utility in Gaussian 09. To use cubegen a formatted checkpoint file must be generated. This formatted checkpoint file is then converted into a .cube file using cubegen where density=scf was used to generate our

density plots. The two cube files (solvated and vacuum) were then subtracted from each other using a Python script.

QM/MM Density Calculation Results

In Figure 5.8 the density excess for two ligand snapshots is compared; one polar (piperazine – from piperidine → piperazine perturbation), and one non-polar (ethane – from methane → ethane perturbation). For the piperazine snapshot a volume density excess of 7.17$Å^3$ compared to just 0.017$Å^3$ for ethane. From this analysis it is clear that polar ligands have their density affected far greater than that for non-polar ligands; the large amount of electron density excess identified for the polar ligands leads to greater polarization effect felt by the QM ligand via the MM charges. This generally gives highly favourable QM/MM snapshots, which show a large free energy difference between the MM and QM/MM representations of each snapshot. This leads to large QM/MM free energy corrections for polar ligands and hence leading to poor agreement with experiment.

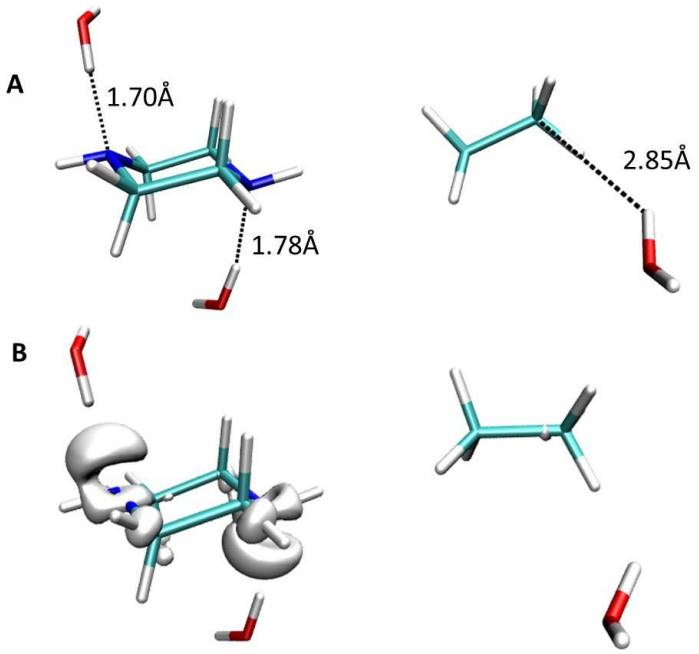


Figure 5.8: Density excess plots. Where A represents the MM system where the waters are interacting with the ligands. B represents the QM/MM density excess, which shows the density being sucked from the QM ligand to the nearby point charges. This plot was generated using a contour level of 0.5.

Previous studies [130, 131] have also shown this density excess for polar solutes, leading to poor QM/MM correction free energies. This suggests that a more elegant embedding strategy is needed for such polar ligands in order to avoid this problem; such a technique shall be described in section 5.2.

5.1.4 Validation - Charge Perturbations

As this QM/MM approach neglects any QM/MM sampling, the pathway-independence of the free energies obtained is a subsequently checked. This is achieved through the use of charge perturbation pathways. These calculations were performed on several examples and two are shown below (Figures 5.9 – 5.10)

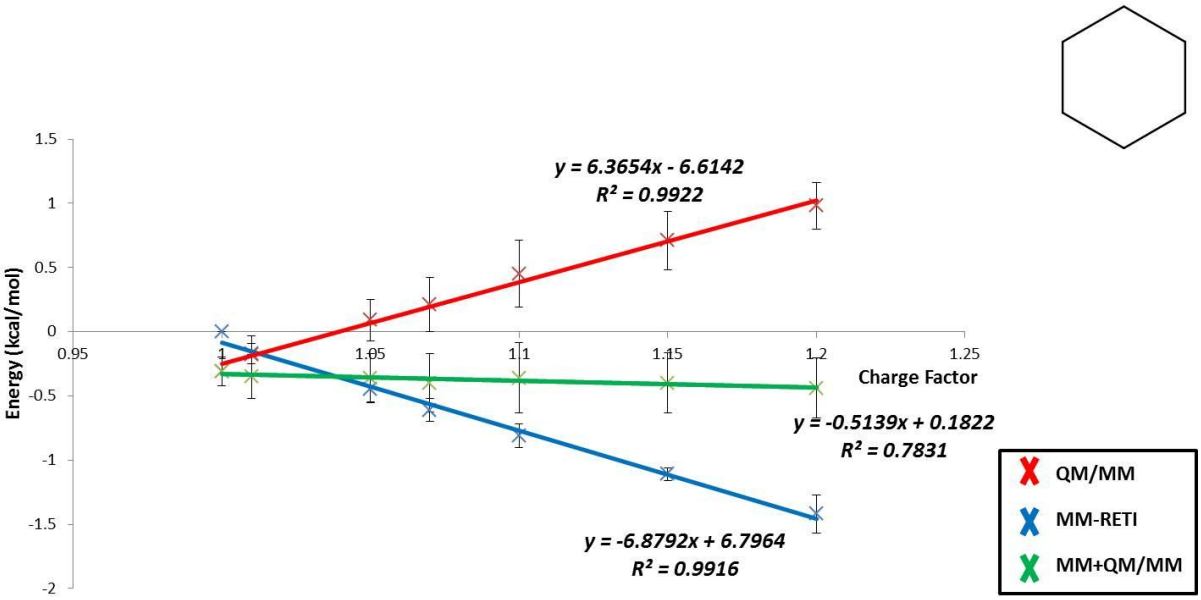


Figure 5.9: Charge perturbation free energies for cyclohexane at several scaling factors. The red line represents the QM/MM free energy, the blue line the MM-RETI free energy and the green line is the MM→QM/MM free energy. The error bars shown are calculated from four independent simulations using standard error.

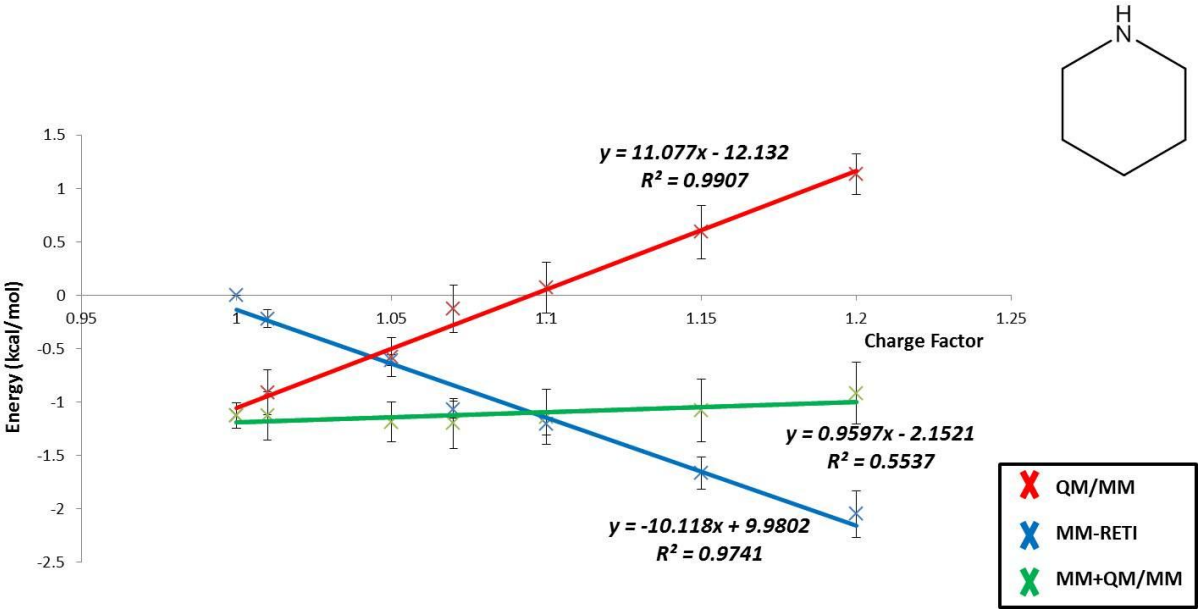


Figure 5.10: Charge perturbation free energies for piperidine at several scaling factors. The red line represents the QM/MM free energy, the blue line the MM-RETI free energy and the green line is the MM→QM/MM free energy. The error bars shown are calculated from four independent simulations using standard error.

In Figures 5.9 – 5.10 the sums of the charge perturbed free energy cycles are generally small (small slopes of the green fitted lines). However, if the free energies are pathway independent, the mean values of these sums calculated over all scale factors must be equal to that of the non-charge perturbation MM→QM/MM free energies. For cyclohexane (Figure 5.9) the average MM→QM/MM free energy is -0.38 (0.16) kcal.mol^{-1} which is very similar to the non-charge perturbed value of -0.32 (0.11) kcal.mol^{-1}. This is repeated for piperidine (Figure 5.10), where the average MM→QM/MM free energy is -1.12 (0.21) kcal.mol^{-1}, which again is highly similar to the non-charge perturbed value of -1.05 (0.23) kcal.mol^{-1}. This pattern was observed for all the charge perturbations studied here, suggesting that our free energies are pathway independent and internally consistent. This suggest that our QM/MM method produces reliable QM/MM free energies for the hydration free energy dataset studied.

5.1.5 Conversion of Relative Hydration Free Energies to Absolute Hydration Free Energies

To convert the relative hydration free energies into absolute free energies it was necessary to annihilate the starting points of the perturbation webs (methane, benzene, cyclohexane and cyclopentane). This was performed through the use of soft-core parameters [132, 133] which enable us to soften the Lennard-Jones and coulombic interactions of a solute with its surrounding environment.

This is performed using equations 5.2 [132] and 5.3 [133], where soft-core parameters δ and α are added to the Lennard-Jones and Coulomb equations.

$$V_{(LJ)}r = \left[\left(\frac{\sigma_{ij}^{12}}{\left(\delta\sigma_{ij}^{6} - r_{ij}^{6}\right)^{2}}\right)\left(\frac{\sigma_{ij}^{6}}{\delta\sigma_{ij}^{6} - r_{ij}^{6}}\right)\right] \quad (5.2)$$

$$V_{coul}(r) = \frac{(1 - \alpha)^{n} q_i q_j}{\sqrt{4\pi\varepsilon_0(\alpha + r_{ij}^{2})}} \quad (5.3)$$

If the suitable values are chosen (an excellent guide on how to choose soft-core parameters was published by Simmerling *et al.* [133]) for these parameters (δ and α), the solute will be annihilated and the absolute hydration free energy computed. This free energy was then used to go back through our perturbation webs to obtain the absolute hydration free energies for each compound.

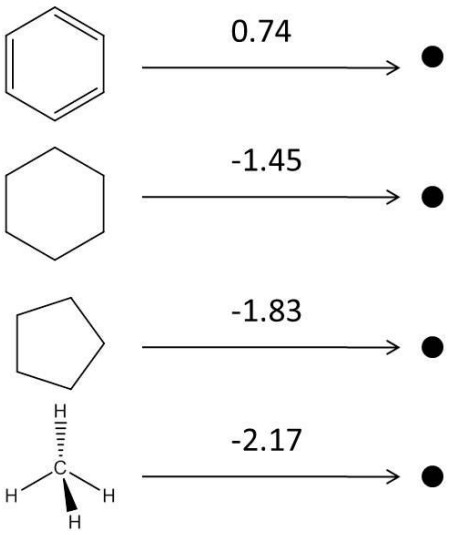
Soft-core Annihilation protocol

The bond angles and torsions for the ligands were sampled during the simulation, with aromatic ring structures being the only exception. The bond lengths of the ligand were constrained. A 12 Å based cut-off from the ligands centre of mass was employed in all simulations. A range of soft-core parameters were tested in order to assess which were best for each individual ligand, we choose to simulate with an *n* value of 6 (equations 5.2 and 5.3). Replica exchange moves were attempted every 200000 moves. The solvent was equilibrated for 20 million moves to remove any bad contacts with the solute. The system was then equilibrated at one state (the end state with the larger solute) for 10 million further moves where solute, and solvent moves were attempted. The resulting configuration was distributed over the 16 values for the coupling parameter λ (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82,

0.88, 0.94, and 1.00) and equilibrated for 10 million moves before collecting statistics for 40 million moves.

Soft-core Annihilation – Results

The results of the soft-core annihilations are shown below (Figure 5.11) along with the combination of soft-core parameters used for each ligand. The results are very accurate when compared with Guthrie's experimental data. The relevant free energies are summarised in Table 1.3 of Supporting Information 1.



Compound	δ	α
Benzene	1.5	1.0
Cyclohexane	1.5	1.0
Cyclopentane	1.5	1.5
Methane	1.0	2.0

Compound	Hydration free energy (kcal/mol)	Experimental (kcal/mol)
Benzene	-0.74 ± 0.25	-0.86
Cyclohexane	1.45 ± 0.29	1.23
Cyclopentane	1.83 ± 0.21	1.20
Methane	2.17 ± 0.32	1.99

Figure 5.11: MM-RETI free energies for annihilation of methane, benzene, cyclohexane and cyclopentane compared to experimental data [127]. The soft-core parameters utilised are also shown. The error bars shown were calculated from four independent simulations using standard.

Using these data it was possible to retrieve the QM/MM absolute binding free energies for the entire dataset (Figure 5.12) via moving back through the perturbation webs to account for these annihilations.

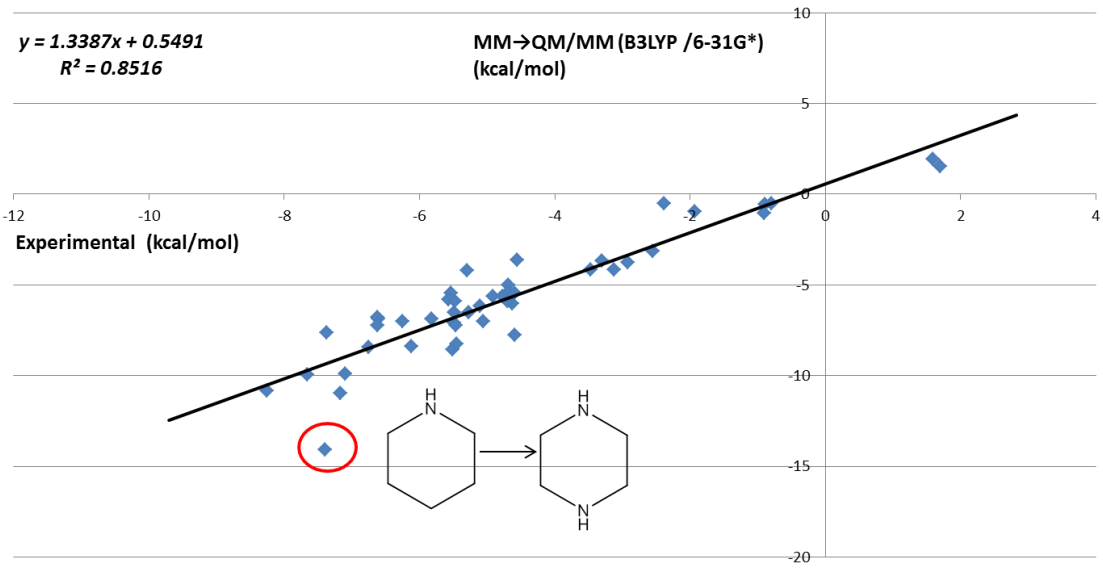


Figure 5.12: QM/MM absolute hydration free energies versus Guthrie experimental [127] absolute hydration free energies. The error bars shown were calculated from four independent simulations using standard error.

The results from this show an encouraging coefficient of determination of 0.85, if we compare this to the 0.84 achieved by Mobley *et al.* [122] we can see that our MM→QM/MM method performs comparably to standard MM techniques. This is an excellent result, which is made better as standard MM is parameterised for this exact problem, whereas this QM/MM method is not. There is one large outlier, piperazine from the piperidine → piperazine perturbation, however, this has been identified in MM in this study and in the study by Mobley *et al.* as an outlier and in the previous QM/MM study into relative hydration free energies this perturbation was identified as a large outlier due to the highly polar nature of both end states.

5.2 MM→QM/MM - Adapted Electrostatic Embedding

As mentioned previously, a more elegant charge embedding approach is needed to accurately describe polar ligands in this QM/MM method. Investigation of the literature found several approaches to deal with this issue [134, 135, 136, 137], however most of these involved costly QM calculations which are unfeasible for our system size. However, work from Brooks *et al.* in which they developed a Gaussian 'blurring' method where a point charge is delocalised using a Gaussian shape function [138] appeared to show promise.

The idea behind this is that by smoothing the charge it will lessen the impact of bare charges which are close to the QM region, and hence causing overpolarisation of the QM region. The interaction between the delocalised MM charge (i) with charge q_i and the QM nucleus (j) with charge Z_j at the distance r_{ij} is given by equation 5.4:

$$E_{ij} = \frac{q_i Z_j}{r_{ij}} \operatorname{erf}\left(\frac{r_{ij}}{\sigma}\right) \tag{5.4}$$

where σ is the Gaussian blur width and erf is an error function. In this method a blur width is applied to any MM point charge, a low blur value (< 1.0 Å) enables the electrostatic coupling to disappear, whereas a high blur value (>1000 Å) recovers point charge behaviour. This methodology was tested by Brooks *et al.* [138] on several small organic molecules, where the MM and QM regions were partitioned across the molecule. Energetics of this system, including rotational barriers, proton affinities and deprotonation energies were then computed using CHARMM (MM) and GAMESS (QM) using RHF/6-31G* and the results were compared to two link atom approaches; the

single link atom approach and the double link atom approach [138]. The results of this study showed that the Gaussian blur technique could produce more accurate results than either standard (unscaled) link atom approaches.

5.2.1 Gaussian Blurring – Test Case

This method was initially tested against the simple transformation of methane → TIP4P water. As mentioned earlier, this perturbation is over-polarised at the TIP4P water end state using point charge embedded QM/MM. Therefore, this is an ideal test case to understand if the Gaussian blur technique can be applied here.

Gaussian Blur Protocol

Gaussian input files were converted into GAMESS-UK format via a simple bash script. The background charges were embedded into the QM/MM calculation through the use of the keyword bq, and these charges were blurred though the use of the keyword blur, where the blur factor was also given. We tested at a range of blur values, including; 0.5 Å, 0.75 Å, 0.8 Å, 0.9 Å and 1.0 Å.

Gaussian Blur – Results

The results from the Gaussian blur test case are shown below in Figures 5.13 – 5.14.

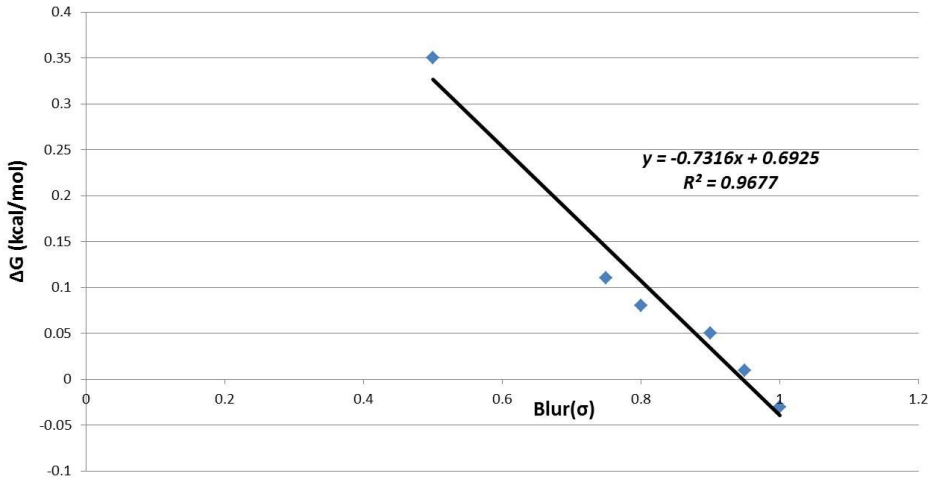


Figure 5.13: QM/MM correction energy versus Gaussian blur width for methane end point of methane → TIP4P water perturbation. All Gaussian blur values shown here are in Å.

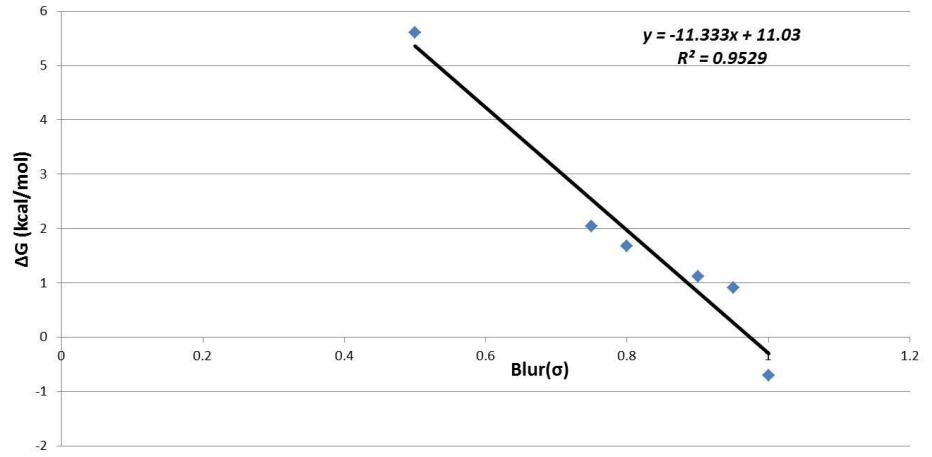


Figure 5.14: QM/MM correction energy versus Gaussian blur width for TIP4P water end point of methane → TIP4P water perturbation. All Gaussian blur values shown here are in Å..

These individual blur studies show we can control the overall QM/MM correction with Gaussian blur width. However, which blur factor gives the best results? Figure 5.15 shows that a Gaussian blur width between 0.8 – 1.0 Å gives rise to a QM/MM free energy which is within experimental error for this perturbation.

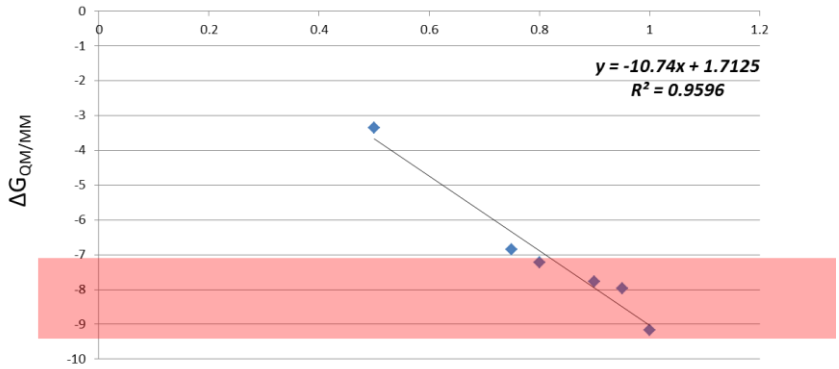


Figure 5.15: MM→QM/MM versus Gaussian blur width for methane→ TIP4P water. The red box represents the experimental error reported by Guthrie [127] for this perturbation.

In addition, when applying a Gaussian blur of 1000 Å the same result was obtained as previously produced (Table 5.4) using Gaussian 09. Therefore, because of this encouraging result, it was decided to extend this test set to analyse how this technique would perform on a range of functional groups.

5.2.2 Gaussian Blur – Extended Dataset

The extended dataset contained an additional 59 perturbations from the Mobley dataset. For a full list of the perturbations chosen please refer to Supporting Information 1, Figure 1.12. Each of the endpoints from these perturbations was subjected to the Gaussian blur technique described above, at two smearing values of 0.95 Å and 1.0 Å. The original MM-RETI and QM/MM free energies are shown in Figures 5.16 – 5.17.

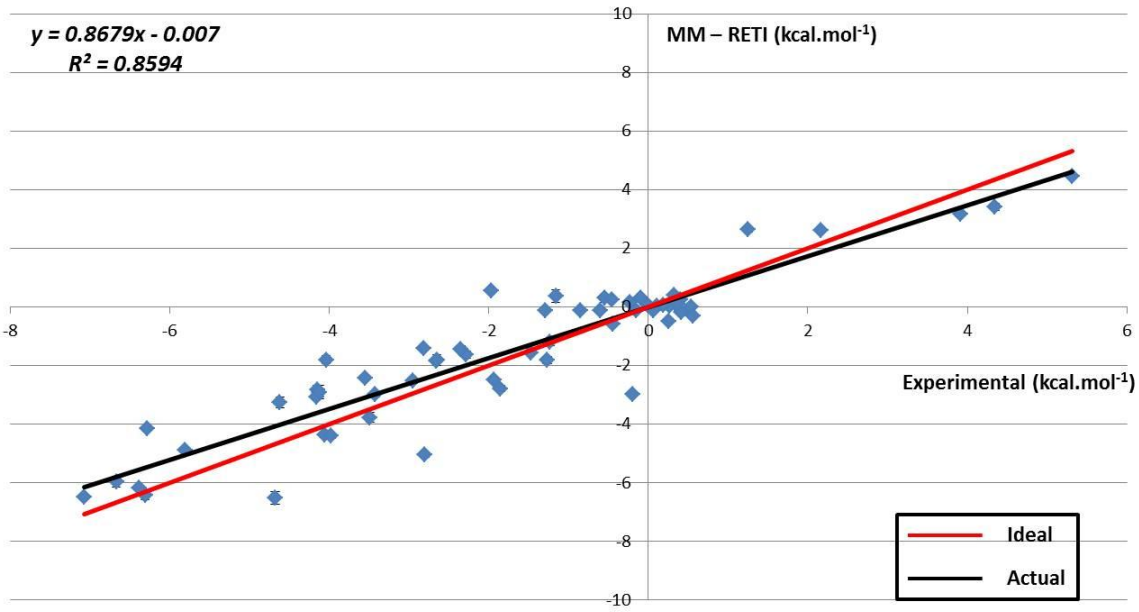


Figure 5.16: MM-RETI relative free energies versus Guthrie's experimental [127] relative hydration free energies for extended dataset. The red line represents ideal correlation (1 to 1) and the black line represents the actual correlation. The error bars shown are calculated from four independent MM-RETI simulations using standard error.

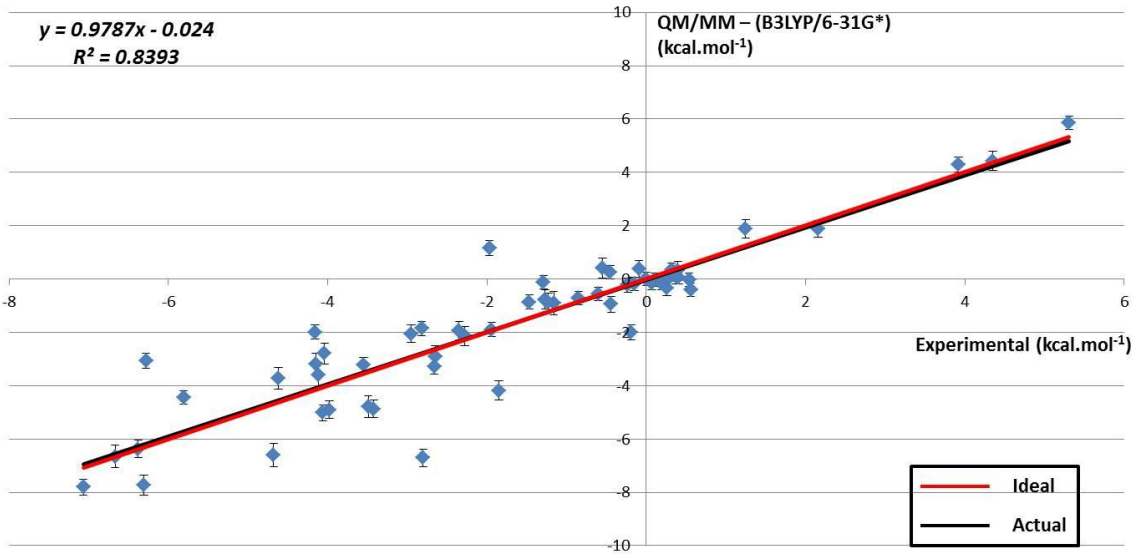


Figure 5.17: MM→QM/MM relative hydration free energies versus Guthrie's experimental [127] relative hydration free energies for extended dataset. The red line represents ideal correlation (1 to 1) and the black line represents the actual correlation. The error bars shown are calculated from four independent simulations using standard error.

Gaussian Blur Extended Dataset Results

The results from both Gaussian blur widths of 0.95 Å and 1.0 Å are shown below (Figures 5.18 – 5.19). The relevant free energies are summarised in Tables 1.4 and 1.5 of Supporting Information 1.

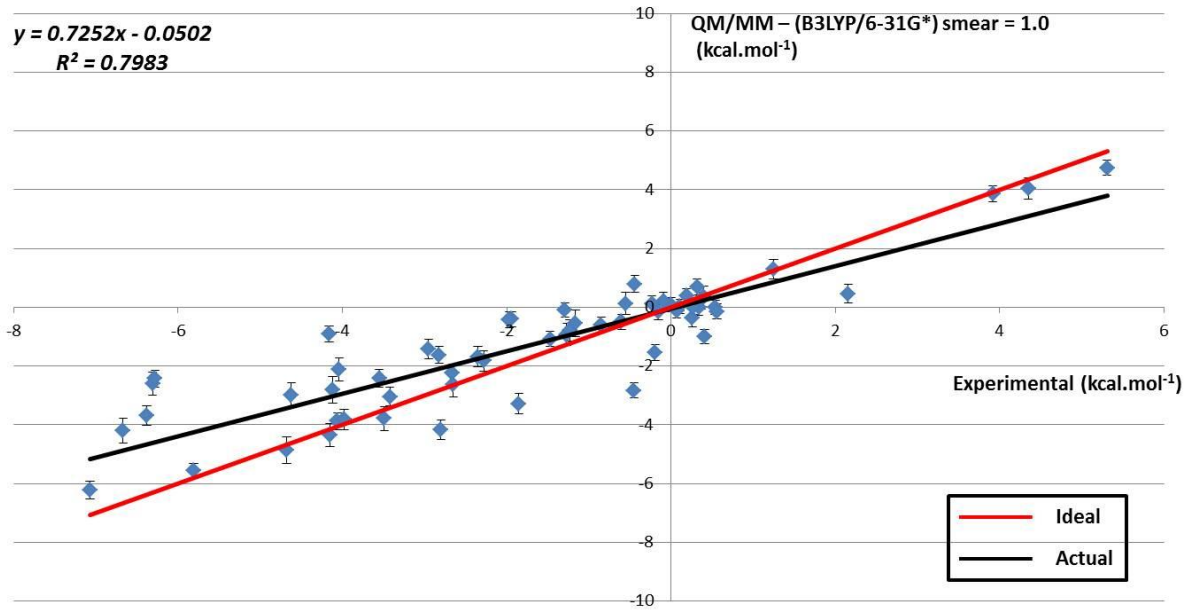


Figure 5.18: MM→QM/MM Gaussian blur (1.0 Å) versus Guthries 's experimental [127] relative hydration free energies. The red line represents ideal correlation (1 to 1) and the black line represents the actual correlation. The error bars shown are calculated from four independent simulations using standard error.

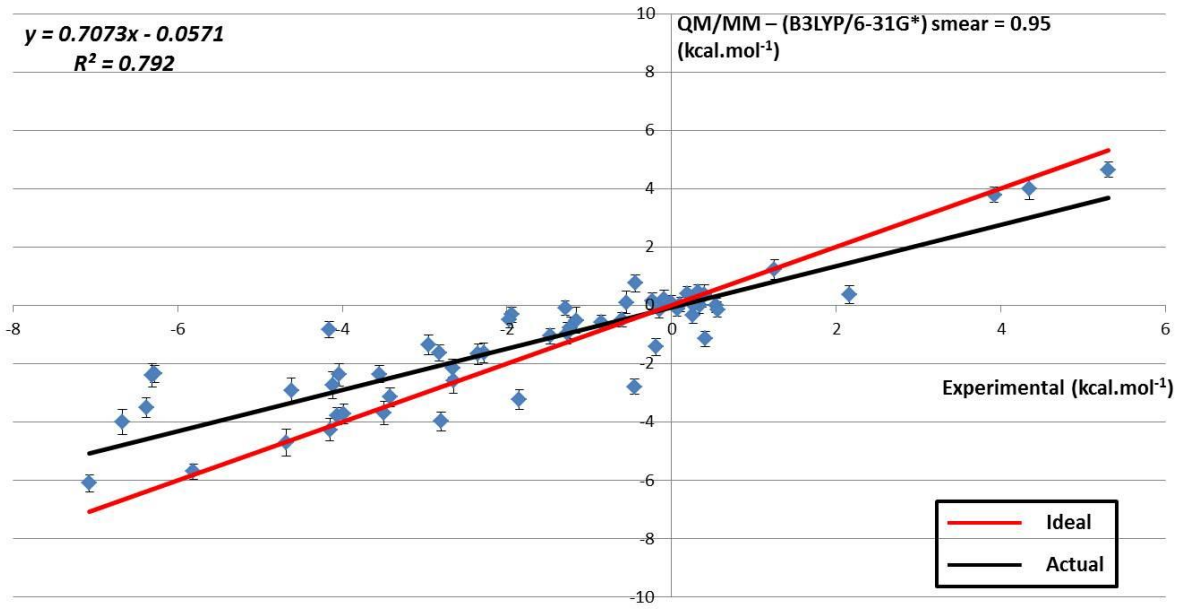


Figure 5.19: MM→QM/MM Gaussian blur (0.95 Å) versus Guthrie's experimental [127] relative hydration free energies. The red line represents ideal correlation (1 to 1) and the black line represents the actual correlation. The error bars shown are calculated from four independent simulations using standard error.

The coefficient of determination for both of the Gaussian blur widths studied is 0.79, which is very similar to the original QM/MM study of 0.86. The large difference is in the slope which shifts from around 1 in the original QM/MM study to 0.7 using the Gaussian blur technique. This suggests that the application of the Gaussian blur does reduce the impact of the point charges, and in fact under-polarises the results for most of this extended dataset. This indicates that this method does not work for all of the perturbations studied. It also highlights that we cannot simply apply a standard Gaussian blur width to a wide range of compounds containing different functionalities, which in turn have different polarisation associated with them. Further investigation is needed to fully understand the impact of Gaussian blur widths upon the calculated QM/MM hydration free energies for compounds containing different functional

groups. It is hoped that future work in this area can provide a useful set of rules for different compound types as currently knowledge *a priori* is needed to use this method and this is not ideal.

5.3 Conclusions

This study focussed on using the QM/MM method presented here to calculate QM/MM corrected hydration free energies for 110 small organic molecules. The results from the MM free energy study showed excellent agreement to experimental and previously published data. The results from the QM/MM corrected hydration free energy study also show very good agreement with experimental data. Yet, several caveats were identified that affect the accuracy of our final QM/MM corrected free energies. In particular outliers all shared several common features; they all involve polar species with the ability to hydrogen bond to the aqueous environment. Further analysis showed that the main driving force behind these results is the embedding scheme utilised within this method. As this method uses snapshots purely from an MM ensemble a close hydrogen bond (between our ligand and water) within the MM ensemble can lead to significant density 'leeching', where the embedded MM point charge pulls density from the QM ligand, in the QM/MM representation. This issue leads to overpolarisation of the QM region.

This study into the QM/MM method's ability to accurately calculate QM/MM corrected relative hydration free energies showed great promise with QM/MM performing as well as MM, which is very surprising as MM forcefields are parameterised for this exact problem; however, identification of structurally similar outliers has exposed the need for a more elegant embedding strategy when placing

the MM system into the QM/MM calculations. A Gaussian blur technique was also identified and tested, which enables the user to blur the embedded MM point charges using a Gaussian distribution. This method shows great promise in being able to negate the over-polarisation effects of the conventional QM/MM implementation, however knowledge is needed a priori as to whether this “smearing” is needed.

6 Calculation of QM/MM Relative Binding Free Energies for 9 CycloOxygenase 2 inhibitors

6.1 Biological Relevance

Prostanoids are a subclass of eicosanoids consisting of the prostaglandins, the thromboxanes, and the prostacyclins. The biosynthesis of prostanoids is induced in different pathological conditions, including inflammatory diseases [139], neurological disorders [140, 141] and cancer [142]. Since the early 1980's two cyclo-oxgenases enzymes, cyclo-oxygenase 1 (COX-1) and Cyclo-Oxygenase 2 (COX-2), have been identified as integral to the production of prostanoids as they are responsible for the production of prostaglandin (H_2), which is the rate-limiting step in the production of prostanoids. Prostanoid biosynthesis is inhibited by non-steroidal anti-inflammatory drugs (NSAIDs) that are widely prescribed as analgesics and anti-inflammatory agents [141]. Their mechanism of action involves inhibition of COX-1 and COX-2 isoenzymes [143]. COX-2, but not COX-1, is characterized by an accessible side pocket that is an extension to the hydrophobic channel [144]. The inhibition of COX-2 is thought to mediate the therapeutic action of NSAIDs, while the inhibition of COX-1 can lead to unwanted side effects, particularly within the gastrointestinal tract [145].

Since the discovery of COX-2 several classes of inhibitors have been designed [141]. These inhibitor sets can generally be classed as first generation COX-2 inhibitors and second generation COX-2 inhibitors. The first class of COX-2 inhibitors designed for treatment of rheumatoid arthritis, osteoarthritis and for pain relief include rofecoxib

and celecoxib which are diaryleterocyclic derivatives containing phenylsulphone and phenylsulphonamide moieties (Figure 6.1).

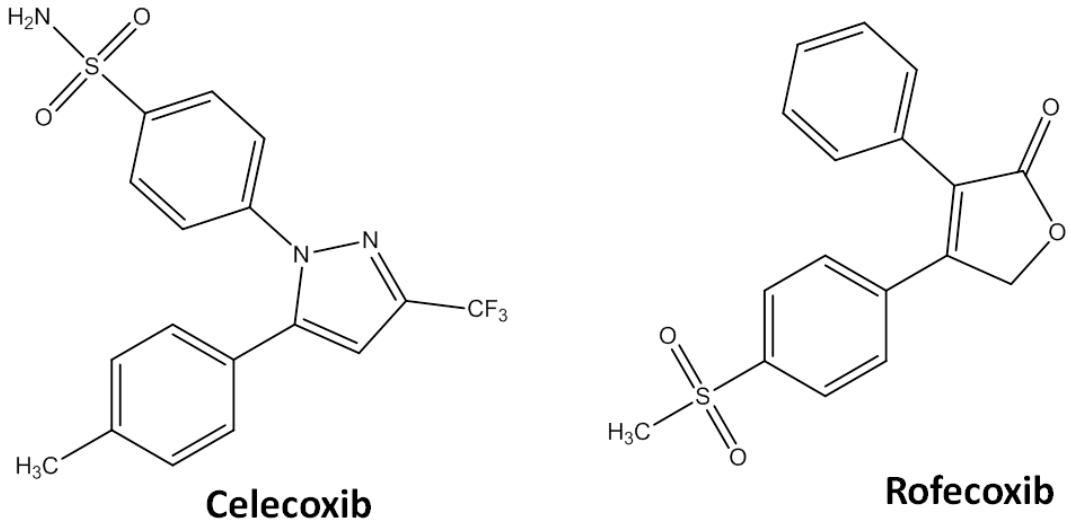


Figure 6.1. Structures of two first class derivative COX-2 inhibitors; Celecoxib and Rofecoxib.

Previous efforts have been made within our group to predict the relative binding free energies for 9 celecoxib based derivatives to the COX-2 enzyme (Figure 6.2) [146]. The results from this study showed a correlation (R^2) of 0.83 between MM(AMBER99/GAFF/AM1-BCC)-RETI calculated binding free energies and experimental data.

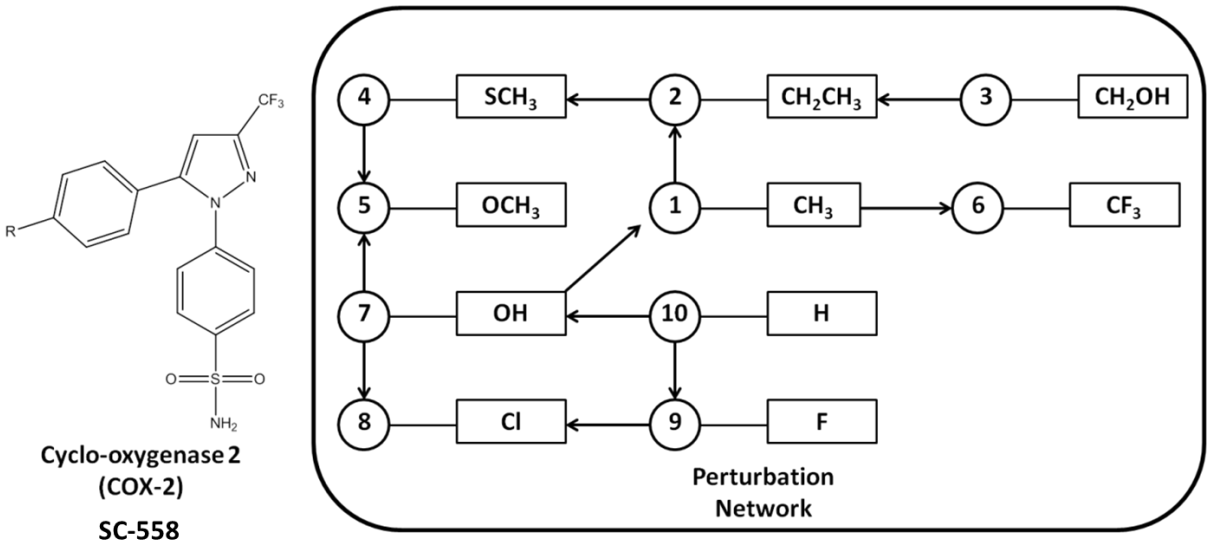


Figure 6.2. Perturbation network for the set of COX-2 ligand perturbations studied here.

This dataset was chosen for this study to understand if our QM/MM method could match or even out-perform classical methods for predicting binding free energies, for which, in this case, MM MM(AMBER99/GAFF/AM1-BCC)-RETI already gives very good correlation.

6.2 System Preparation

Protein – ligand setup

The PDB structure of murine COX-2 (PDB code 1CX2) [147] was selected as a starting point for this study. In this structure the polar hydrogen atoms had already been assigned by the crystallographers. Non-polar hydrogen atoms were added to this structure using the Reduce software package [148]. Previous theoretical studies [146] have shown that the sulphonamide conformation in celecoxib of this COX-2 crystal structure is incorrect. As was done previously, the N—S—C—C torsion around this functional group was rotated to interact favourably with neighbouring residues and a

nearby haem group was removed as it did not have any direct interactions with the binding site. The protein was parameterised using the AMBER99 force field [13], inhibitors were parameterised with the GAFF force field [14] and the partial atomic charges were derived using the AM1-BCC method [124], as implemented in the AMBER 10 suite. To avoid bad steric clashes, the protein-ligand complex (1CX2/ligand 2) was minimised in the SANDER module of AMBER 10 using generalised Born force field to represent the solvent. The backbone of the protein was subsequently fixed for Monte Carlo simulations, which were performed using a modified version of ProtoMS2.2 [123]. To reduce computational cost, only protein residues that contained one heavy atom within 15 Å of any representative ligand atom were retained. The resulting protein scoop contained 155 residues. The ligands were modelled in the binding site based upon the binding mode predicted by the docking program GOLD [116], the binding modes for each ligand were generated by Michel *et al.* [146] in the previous binding free energy for this protein-ligand dataset. Crystallographic waters were retained and the complex was hydrated by a sphere of TIP4P [125] water molecules of 22 Å radius and centered on the geometric centre of each ligand studied. To prevent evaporation, a half-harmonic potential with a 1.5 kcal.Å^{-2} force constant was applied to water molecules whose oxygen atom distance to the ligands centre of geometry was greater than 22 Å. A similar sphere of water was used for the unbound state.

Monte Carlo Simulation Protocol

The bond angles and torsions for the side chains of residues within 10 Å of any ligand heavy atom and all bond angles and torsions of the ligand were sampled during the simulation, with ring structures being the only exception. The bond lengths of the

residues and ligand were constrained. The total charge of the system was brought to zero by neutralising lysine residues 511 and 532 lying in the outer 'frozen' part of the scoop. The neutralised lysines were then re-modelled using the AMBER99 forcefield. A 10 Å residue based cut-off was employed in all simulations.

For simulation in the bound state, solvent moves were attempted with a probability of 71.28%, protein side-chain movements with a probability of 24.08% and solute moves with a probability of 4.64%. In the unbound state, solvent moves were attempted 99.07% of the time. Replica exchange moves were attempted every 200000 moves. The solvent was equilibrated for 20 million moves to remove any bad contacts with the solute. The system was then equilibrated at one state (the end state with the larger solute) for 20 million further moves where solute, protein, and solvent moves were attempted. The resulting configuration was distributed over the 16 values for the coupling parameter λ (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82, 0.88, 0.94, and 1.00) and equilibrated for 10 million moves before collecting statistics for 640 million moves (bound) and 320 million moves (free).

QM Single Point Energy Protocol

Configurations from the endpoint (λ=0 and λ=1) of the classical free energy simulations were selected and used as input for DFT-QM/MM single point energy calculations with Gaussian 09 [126]. One QM single point energy calculation with background charges representing the solvent and protein residues within our cut-off (Gaussian keyword 'CHARGE') were performed every 100000^{th} MM/MC moves, with symmetry operations disabled (Gaussian keyword 'NoSymm'). This gave a total of 6400

QM/MM single points for each solute perturbation in the bound state and 3200 QM/MM single points for each solute perturbation in the free state. Gaussian calculations with embedded background charges allow a polarisation of the QM wave function via the MM charges, however no back polarisation of the MM part via the polarised QM wave function was considered.

The QM energies were computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

As our ligands were flexible we needed to compute the QM vacuum energies for each snapshot used. This was carried out in Gaussian 09, but without the use of the 'CHARGE' and 'NoSymm' keywords, which are only necessary if embedding MM point charges in our calculation.

The QM vacuum energies were again computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

6.3 Results & Discussion

6.3.1 MM – RETI Results

The calculated relative binding free energies of 9 celecoxib derivatives are shown in Figure 6.3. The coefficient of determination (R^2) between predicted and experimental binding free energies [146] is 0.79. The mean unsigned error (MUE) is equal to 0.79 kcal.mol^{-1}, this is within "chemical accuracy". The relevant free energies are summarised in Tables 2.1 and 2.3 of Supporting Information 2.

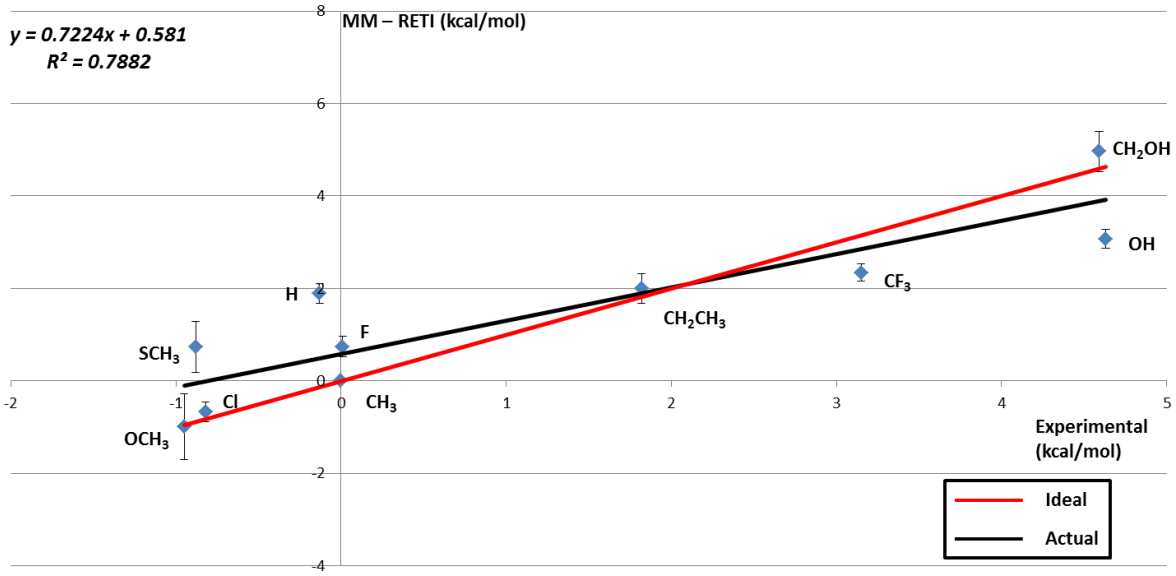


Figure 6.3. MM-RETI results versus experimental data [146]. The red line represents the ideal (1 to 1) correlation and the black line represents the best fit. The error bars shown are calculated from four independent simulations using standard error.

To ensure smooth transition between the two end states (λ=0 and λ=1) the free energy gradients for each perturbation were studied (Figure 6.4).

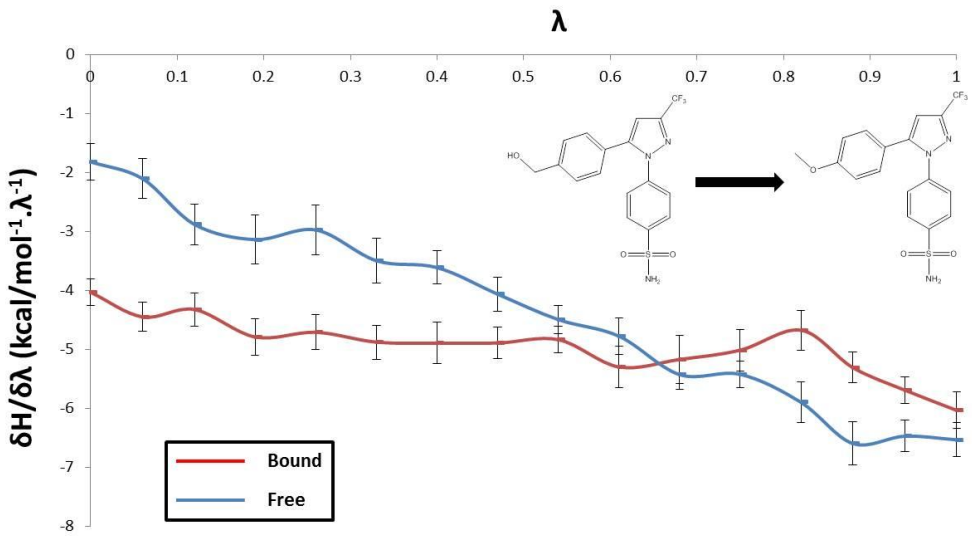


Figure 6.4. Free energy gradients for free (blue line) and bound (red line) for the ligand 7 to ligand 5 perturbation. The error bars shown are calculated from four independent simulations using standard error.

These show that for both the free and bound legs of our free energy simulations the transition across our reaction co-ordinate is very smooth. This indicates that the free energies obtained from our simulations are precise.

To analyse the statistical uncertainty in our free energy simulations we calculated the hysteresis for closing a binding free energy cycle for a set of 4 COX-2 perturbations (7→8, 9→8, 10→9 and 10→7). For the free legs of our simulations (Figure 6.5) the hysteresis was found to be extremely small at just -0.17 (0.46) kcal.mol^{-1} suggesting little statistical uncertainty.

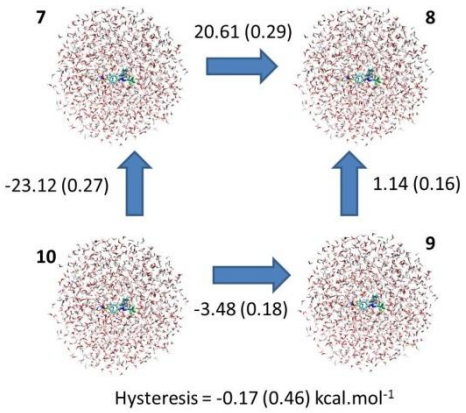


Figure 6.5. MM-RETI free leg hysteresis for four COX-2 perturbations. The standard errors for each step are shown in the brackets.

For the bound legs (Figure 6.6) the hysteresis rises to -0.69 (0.18) kcal.mol^{-1}, which implies that our bound legs contain a higher statistical uncertainty than our free legs, however this is still a very reasonable value.

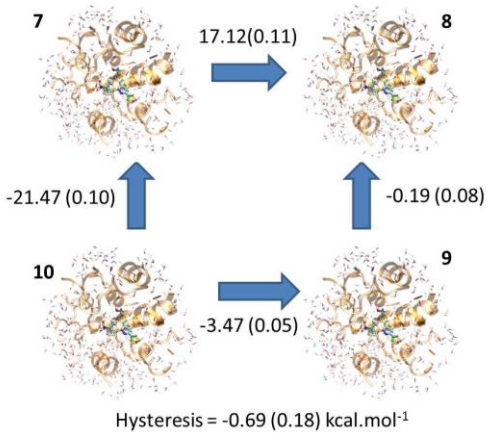


Figure 6.6. MM-RETI bound leg hysteresis for four COX-2 perturbations. The standard errors for each step are shown in the brackets.

The overall hysteresis for closing the cycle for binding free energies is small at just 0.51 (0.49) kcal.mol^{-1}, as illustrated in Figure 6.7. Such hysteresis is an important indicator of the statistical uncertainty in these calculations; however, hysteresis of ≈1kcal.mol^{-1} over four simulations is notably small [11].

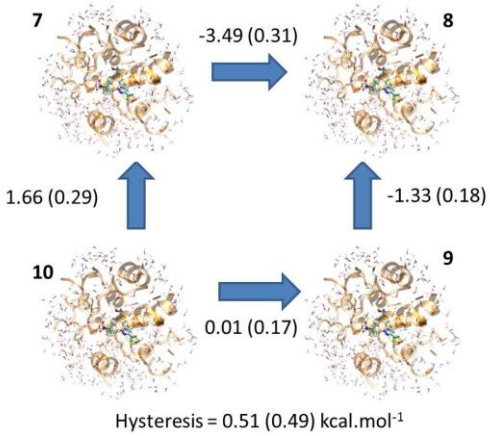


Figure 6.7. MM-RETI total hysteresis for four COX-2 perturbations. The standard errors for each step are shown in the brackets.

To futher quantify our calculated binding free energies, we calculated the Predictive Index (PI) [149], which enables ranking of our inhibtors. For a set of experiemental values, $E(i)$, and corresponding precited scores $P(i)$, the index is as follows:

$$PI = \sum_{j>i}\sum_{i} w_{ij}C_{ij} \Big/ \sum_{j>i}\sum_{i} w_{ij} \tag{6.1}$$

where

$$w_{ij} = |E(j) - E(i)| \tag{6.2}$$

and

$$C_{ij} = \begin{cases} 1\ if\ [E(j) - E(i)]/[P(j) - P(i)] < 0 \\ -1\ if\ [E(j) - E(i)]/[P(j) - P(i)] > 0 \\ 0\ if\ [P(j) - P(i)] = 0 \end{cases} \tag{6.3}$$

$P(i)$ and $E(i)$ refer to the predicted and experimental binding free energy of reference compound 1, whereas $P(j)$ and $E(j)$ are the predicted and experimental binding free energies of the 9 other COX-2 celecoxib derivatives. This index ranges from -1 to +1, depending on how well the predicted binding free energies track the rank order of experimentally obtained binding free energies: +1 arises from perfect prediction, -1 arises from predictions which are always wrong, and 0 arises from predictions which are completely random. This function includes a weighting term that depends upon the difference between experimentally obtained binding free energies. It is also common to use an unweighted predictive measure, known as Kendalls Tau (τI) [150], this index is as follows:

$$\tau I = \sum_{j>i}\sum_{i} C_{ij} \Big/ \sum_{j>i}\sum_{i} w_{ij} \tag{6.4}$$

where C_{ij} is calculated as in equation 6.3 and w_{ij} is calculated as in equation 6.2.

The calculated PI is 0.88, denoting an excellent ability for the MM-RETI free energy simulations to rank the 10 celeoxib inhibitors according to their potency. The τI is 0.56, indicating that without experimental weighting the ranking of our compounds worsens. These results are in excellent agreement with previously published results from Michel *et al.* [146], which is encouraging as both stuides used the same forcefields and charge models. Price and Jorgensen [11] also studied this system and reported results with slightly better agreement with experimenal data, with a MUE of 0.4 kcal.mol^{-1} and a correlation of 0.96 and a PI of 0.96. This improvement may have been cause by their use of a different forcefield (OPLS/AA and CM1A atomic partial charges versus AMBER99/GAFF and AM1-BCC atomic partial charges). Furthermore, within our simulations no water molecules were included within the COX-2 binding site, and depending on the perturbation stuided one or two water molecules were present in the study of Price *et al*. There is no structural evidence to support the presence of water molecules in this buried hydrophobic pocket, and Price *et al*. could not rule out the possibility that these waters were an artefact of the protocol used to generate the watercap for their simulations (see Section 6.3.2 for futher details). Despite these differences the overall predictivity of both methods used is very similar.

In addition to Price *et al.*, a recent study utilising funnnel metadynamics by Limongelli *et al.* [151] suggested that there is a water present in the hydrophobic

pocket of COX-2. Their evidence for the presence of this water was obtained from structural analysis of NSAID's by Selinksky *et al.* [143] where four inhibitors (ibuprofen, methyl flurbiprofen, alclofenac and flurbiprofen) were crystalised with COX-2. This study showed that there could be a water molecule bridging interactions between SER530 and TYR385 in the COX-2 binding site for this set of inhibitors. To investigate the possibility that this water should have been included in our free energy simulations we performed Grand Canonical Monte Carlo (GCMC) on the COX-2 binding site.

6.3.2 Grand Canonical Monte Carlo - Analysis of Waters within COX-2 Binding Site

To simulate a system where the number of particles can vary, it is necessary to utilise the GCMC technique. Orginally formulated by Adams in 1974 [55, 56], the GCMC technique is capable of predicting the location of molecules in both biological and inorganic systems. In contrast to traditional MC and MD simulations, GCMC utilises the μVT ensemble which allows the number of molecules in the system to fluctuate as a function of the applied chemical potential (μ). As such, the methodology is ideally suited to investigating systems where the number of molecules is unknown, such as an apo/pseudo-holo binding site. In a GCMC simulation the moves associated with the canonical ensemble are permitted, alongside three unique moves associated with the μVT ensemble. The first type of move is a particle creation move, whereby the number of molecules in the system increases by one. The second move type involves particle deletion, whereby the number of molecules in the system decreases by one. The final move type is a localised transition move, whereby the inserted molecule(s) are allowed to translate and rotate around the system. The acceptance tests for these moves are shown in equations (6.5, 6.6, 6.7).

$$P_{in} = min\left[1, \frac{\exp(B)}{N+1}\, exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{6.5}$$

$$P_{del} = min\left[1, N exp(-B) exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{6.6}$$

$$P_{dis} = min\left[1, exp\left(\frac{-\Delta E}{k_b T}\right)\right] \tag{6.7}$$

In the above equations, N is the number of particles in the simulation and B is the Adams parameter ($B = \mu'/k_b T + \ln \bar{n}$). $\bar{n}$ is the expected number of particles in the system given the volume of the simulation region and is equal to $\bar{\rho}v$ where $\bar{\rho}$ is the number density of the particle and v the simulation volume [152]. μ' is the excess chemical potential, k_b is the Boltzmann constant and ΔE the change in energy between the new and old states. Historically, B has been used for simualtions instead of μ, for computational simplicity [153]. No explanation has been given for this parameter, although one possible explanation is that it allows the simulation results to be compared to the expected number of molecules in the bulk, $\bar{n}$. Since B and μ_0 differ by a constant, performing a simulation at constant B is equivalent to performing a simulation at a constant chemical potential, μ'.

GCMC- COX-2 simulation protocol

Insertion and deletion attempts were accepted using the Metropolis tests described above. The GCMC simulation was performed on a pseudo-holo structure of COX-2 where ligand 1 was removed from the COX-2 binding site prior to the GCMC simulations.

No formal hardwall region is applied in the GCMC simulations. Although other (bulk) water molecules are prohibited from entering the defined GCMC region, protein atoms are allowed to occupy the same region as the GCMC simulation. As a result a 13.4 x 8.0 x 10.5 $Å^3$ grid was defined around the binding site to obtain suffcient sampling of the binding site region. Each B value was simulated for 40 million MC moves. At the end of each simulation the average population across the entire simlulation was recorded. The decoupling free energy of the water was found using equation 6.8.

$$\Delta G_{dec} = -k_B T \ln\left(\frac{L_{sim}}{L_{ideal}}\right) \tag{6.8}$$

In equation 6.8, L_{sim} is found by intially recording the population at a set B value. This population is converted into a localised concentration by dividing by the simulation volume, and then converting this into a molar concentation using Avogadro's number. L_{ideal}, as shown in equation 6.9, is related to the B value of the simulation and is found using the following.

$$L_{ideal} = 55.56M \text{ x } \exp\left(B - \ln \bar{n}\right) \tag{6.9}$$

In equation 6.9, $\bar{n}$ is the expected number of particles in the system given the volume of the simulation region and is equal to $\bar{p}v$, where $\bar{p}$ is the number density of the particle and v the simulation volume [153].

After calculating ΔG_{dec}, the binding free energy of the waters was found using equation 6.10.

$$\Delta G_{bind} = \Delta G_{dec} + \Delta G_{hyd} \quad (6.10)$$

For each the COX-2 binding site, 9 B values (4, 0, -4, -8, -12, -14, -16, -18, and -20) were simulated to allow for a reliable estimate of the binding free energy. The free energy of hydration, ΔG_{hyd} was taken to be +6.4 kcal.mol^{-1}. [155]

For the GCMC simulations, solvent moves were attempted with a probability of 47.6%, protein side-chain moves with a probability of 6.2%. Insertion and deletions were attempted with an equal probability of 0.8%, with translations and rotations of the GCMC waters attempted with a probability of 44.6%.

GCMC COX-2 Results & Discussion

The results of our GCMC simulations on the pseudo-holo stucture of COX-2 are shown in Table 6.1.

Biasing Potential	Chemical Potential (μ)	Binding Free Energy (kcal.mol^{-1})	No. of Waters
4	0.14	6.53	23.56
0	-2.24	4.16	21.34
-4	-4.63	1.77	18.72
-8	-7.01	-0.61	14.45
-12	-9.40	-2.99	9.21
-14	-10.59	-4.19	3.15
-16	-11.78	-5.38	2.56
-18	-12.97	-6.57	2.17
-20	-14.16	-7.76	0.03

Table 6.1. GCMC results for COX-2 binding site. Results are shown for 9 biasing potentials along with the number of waters present at each bias and the associated chemical potentials and binding free energies of said waters.

These results suggest that as the biasing potential is increased the number of water molecules present in the COX-2 binding site decreases, until at a biasing potential of -20 and a corresponding free energy of -7.76 kcal.mol^{-1} almost no waters occupy the binding site over the course of the simulation.

To understand the implications of these results the binding site and associated GCMC waters were visualised at different biasing potentials. At a bias of -12 roughly 9 GCMC waters are present in the binding site (Figure 6.8).

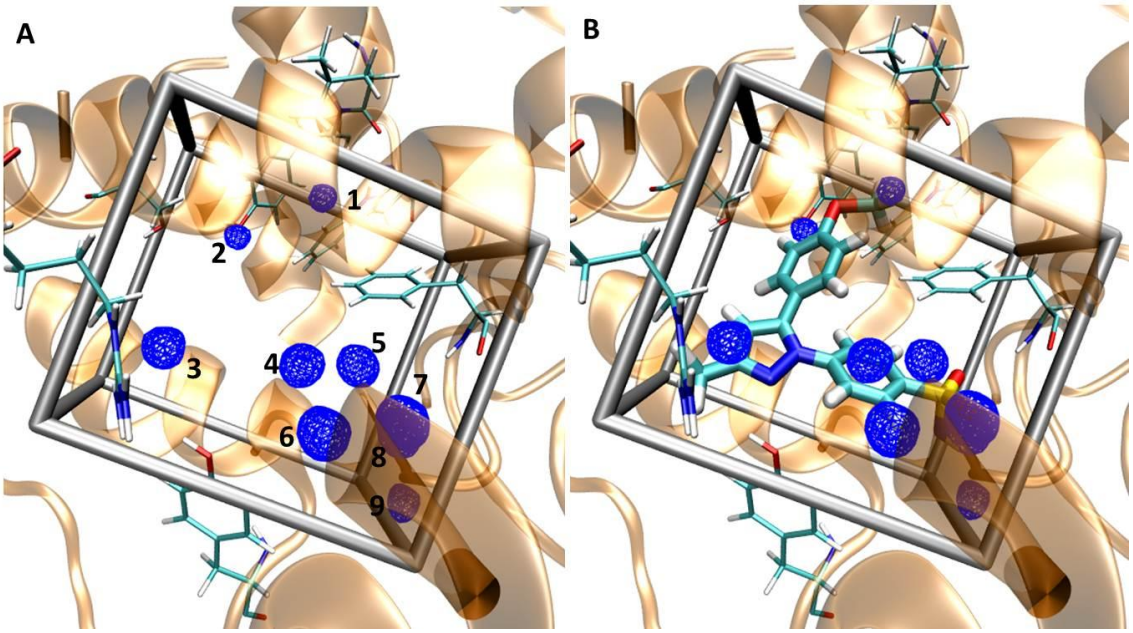


Figure 6.8. Depicts the hydration of the COX-2 binding site at a biasing potential of -12. A shows the 9 waters without any ligand bound, whereas B has ligand 5 superimposed to highlight how the COX-2 inhibtors mirror these water interactions.

The GCMC waters (labelled 1-9) occupy several parts of the binding site. Waters 1 and 2 occupy the buried hydrophobic pocket region which Price and Jorgensen and Limongelli *et al.* suggest could help bridge interactions between the ligand and COX-2 binding site residues. Waters 3-9 occupy the solvent exposed region of the COX-2 binding site. At this biasing potential the binding free energy of these waters is -2.99 kcal.mol^{-1}. With such a weak binding free energy it is unclear whether these waters would be present or displaced upon ligand binding. To further investigate this possibility the COX-2 binding site was visualised at a higher biasing potential of -16 (Figure 6.9), which has a corresponding binding affinity of -5.38 kcal.mol^{-1} for any GCMC waters that are present.

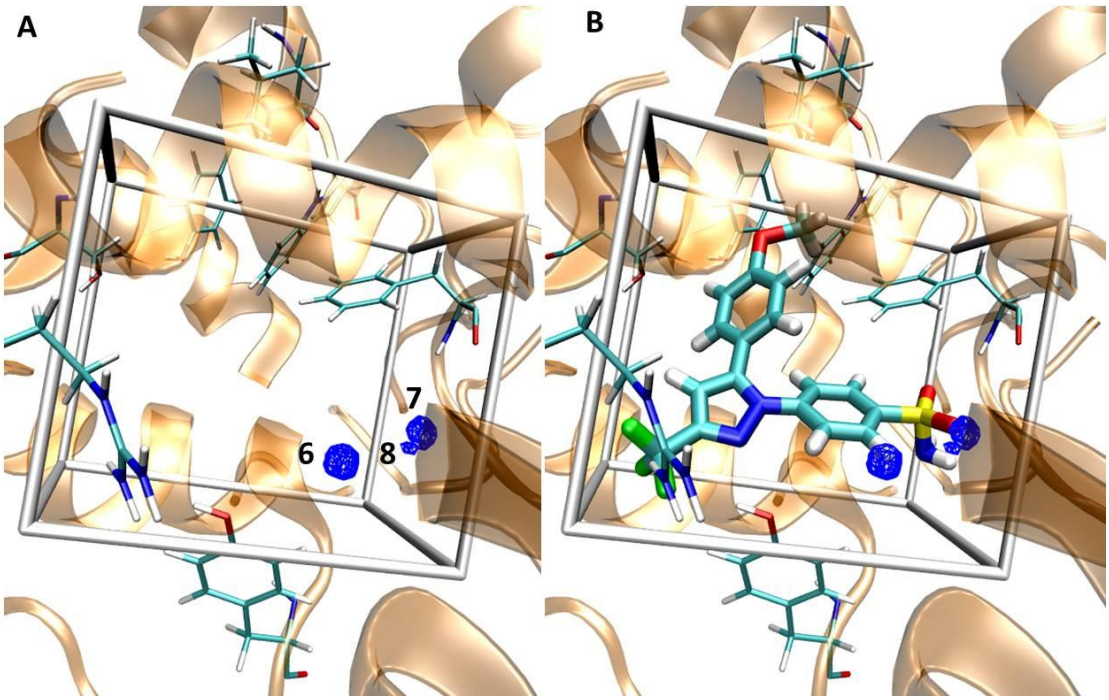


Figure 6.9. Depicts the hydration of the COX-2 binding site at a biasing potential of -20. A shows the 3 waters without any ligand bound, whereas B has ligand 5 superimposed to highlight how the COX-2 inhibtors mirror these water interactions.

It is clear that at this higher bias only three of the previous nine GCMC waters are present. All of these GCMC waters are in the solvent exposed region around, which is occupied by the sulphonamide moeity of the celeoxib–based COX-2 inhibtor. Importantly, no waters are present in the hydrophobic pocket, suggesting that if this region is occupied by water they would be displaced upon ligand binding. This phonomena is common upon ligand binding as it enables a gain in entropic energy as the displaced waters would return to the bulk. The results of this GCMC investigation suggest that there should not be waters present in the hydrophobic pocket of COX-2 and that any which are there would be easily displaced by the binding of celeoxib-based COX-2 inhibtors.

Ideally we would also have performed a GCMC simulation upon a holo COX-2 structure to understand if there is any chance that the COX-2 ligands could stabilse these water molecules. Unfortunately this is outside the scope of this study, but future attempts to perform these simulations would help to shed more light on the hydration patterns within the COX-2 binding site.

6.3.3 MM→QM/MM-FEP Results

The calculated relative MM→QM/MM-FEP binding free energies of 9 celecoxib derivates is shown in Figure 6.10. The coefficeint of determination (R^2) between predicted and experimental binding free energies [146] is 0.49. The MUE is equal to 1.88 kcal.mol^{-1}, which is over 1 kcal.mol^{-1} worse than for the MM-RETI calculated binding free energies. The relevant free energies are summarised in Table 2.2 and 2.3 of Supporting Information 2.

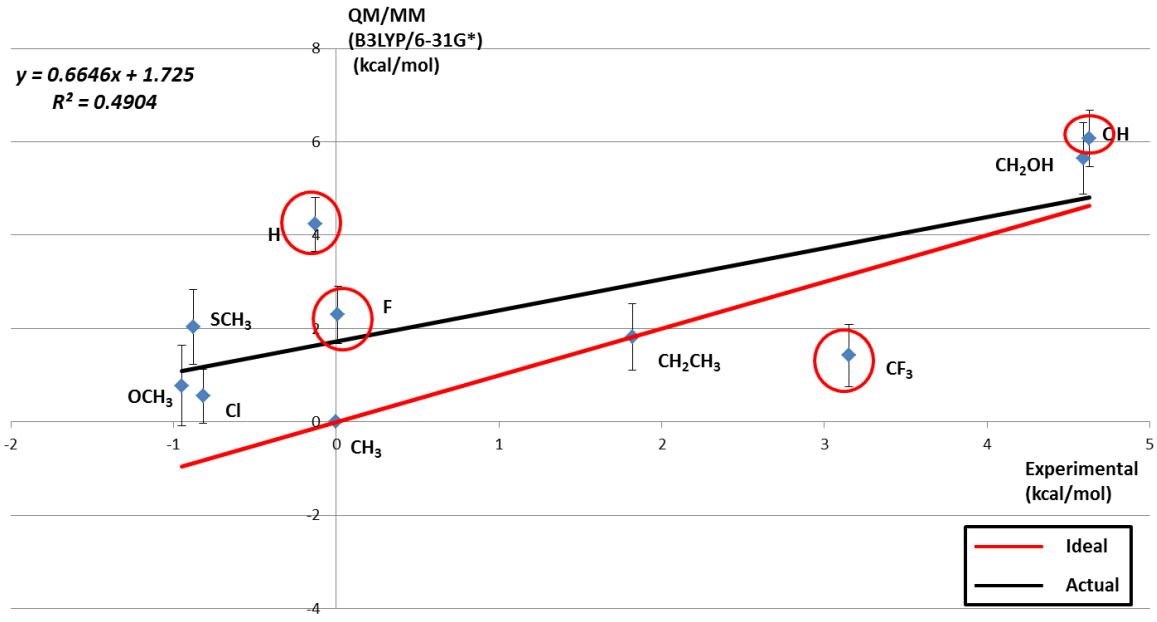


Figure 6.10. MM→QM/MM results versus experimental data [146]. The red line represents the ideal (1 to 1) correlation and the black line represents the actual correlation. The error bars shown were derived from four independent simulations using standard error.

In addition to the loss of accuracy the PI for the MM→QM/MM-FEP binding free energies falls to 0.08, and the τI becomes -0.33. This indicates that when applying our QM/MM corrections to our classically obtained binding free energies our results become random compared to experimental data.

As with our MM-RETI results the hysteresis for closing a binding free energy cycle for a set of 4 COX-2 perturbations (7→8, 9→8, 10→9 and 10→7) was calculated. For the free legs of our MM→QM/MM binding free energy study (Figure 6.11) the hysteresis is 0.77 (1.06) kcal.mol^{-1} this over a 0.5 kcal.mol^{-1} increase from the MM values of -0.17 (0.46). It is also important to note that the error estimate has also increased by 0.6 kcal.mol^{-1}. This indicates that applying our QM/MM corrections leads to higher statistical undertainty and larger errors than compared to standard MM.

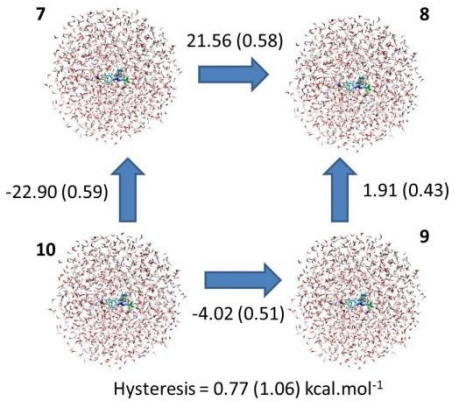


Figure 6.11. MM→QM/MM free leg hysteresis for four COX-2 perturbations. The standard errors for each step are shown in the brackets.

The trend of increased uncertainty continues when considering the bound legs of our MM→QM/MM binding free energy study (Figure 6.12). The hysteresis in the bound legs is -0.94 (0.92) kcal.mol^{-1} which again is an increase from the hysteresis shown in

the MM free energy study of -0.69 (0.18) kcal.mol^{-1}. More noticeable is the increase in the error estimate for the bound leg hysteresis which has increased by 0.74 kcal.mol^{-1}.

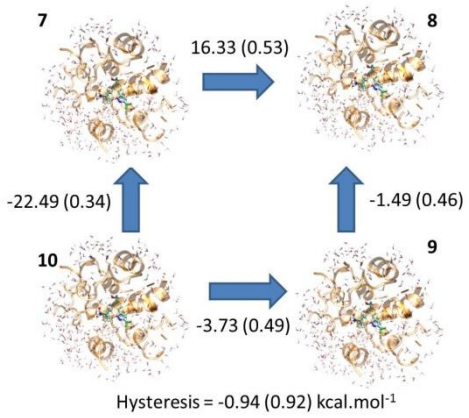


Figure 6.12. MM→QM/MM bound leg hysteresis for four COX-2 perturbations. The standard errors for each step are shown in the brackets.

Combining the free and bound legs leads to the overall hysteresis for this free energy cycle (Figure 6.13). The overall MM→QM/MM hysteresis is -1.71 (1.48) kcal.mol^{-1} which is an increase >1 kcal.mol^{-1} compared to the hysteresis for MM of -0.51 (0.49) kcal.mol^{-1}.

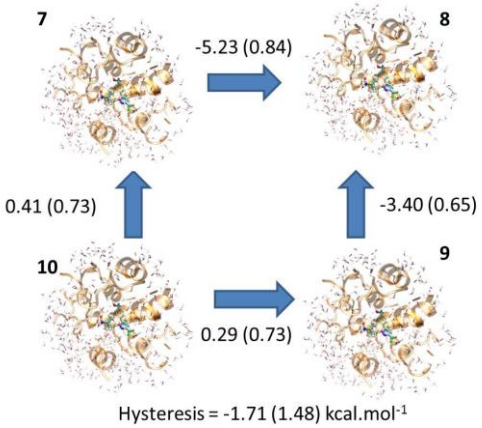


Figure 6.13. MM→QM/MM total hysteresis for four COX-2 perturbations. The standard errors for each step are shown in the brackets.

This large change in hysteresis between MM and MM→QM/MM suggests that we are introducing a large amount of statistical uncertainty, also known as simulation 'noise', through applying our QM/MM corrections.

To understand the changes in accuracy between our MM and MM→QM/MM-FEP binding free energies, in particular those perturbations circled (red circles) in Figure 6.10, we need to analyse the energies produced for each leg of our protein-ligand binding free energy cycle.

For ligand 10 (R=H), ligand 9 (R=F), and ligand 7 (R=OH) the breakdown of MM and QM/MM energies is shown below (Figures 6.14 – 6.16).

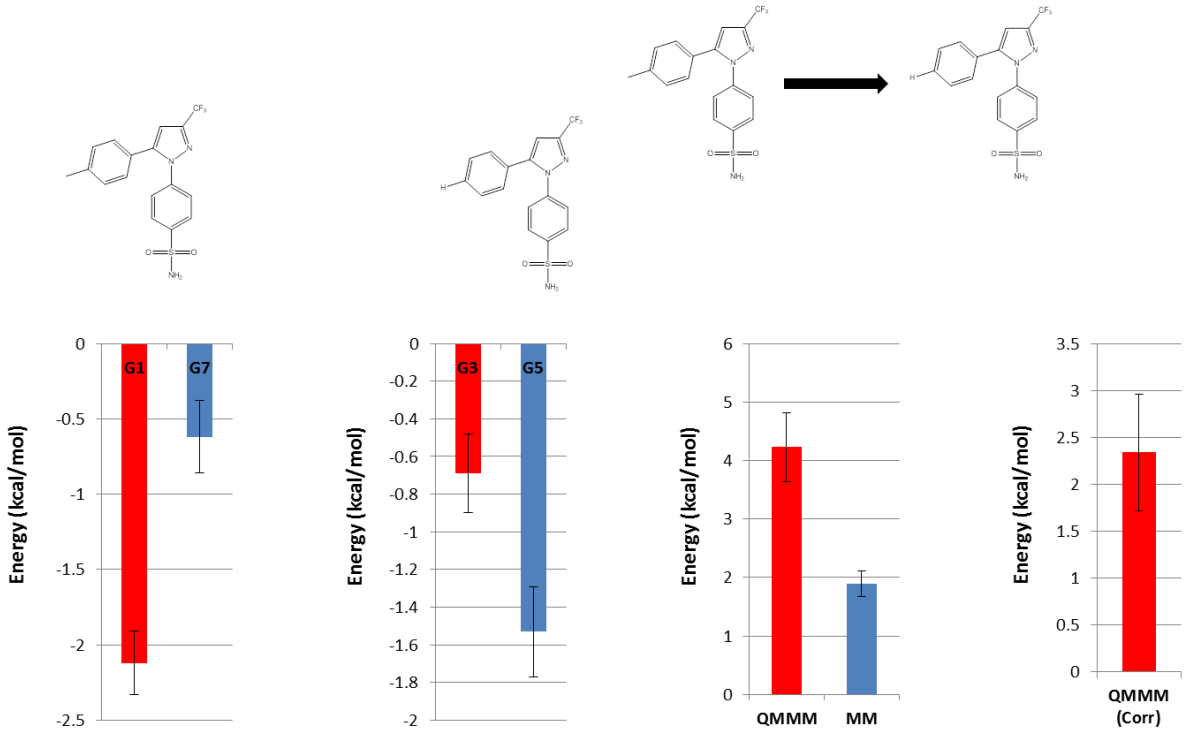


Figure 6.14. MM→QM/MM free energy breakdown for ligand 1 → ligand 10 perturbation. G1 refers to the QM/MM bound leg correction for ligand 1 and G7 refers to the QMM free leg correction for ligand 1. G3 is the QM/MM bound leg correction for ligand 10 and G5 the QM/MM free leg correction for ligand 10. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

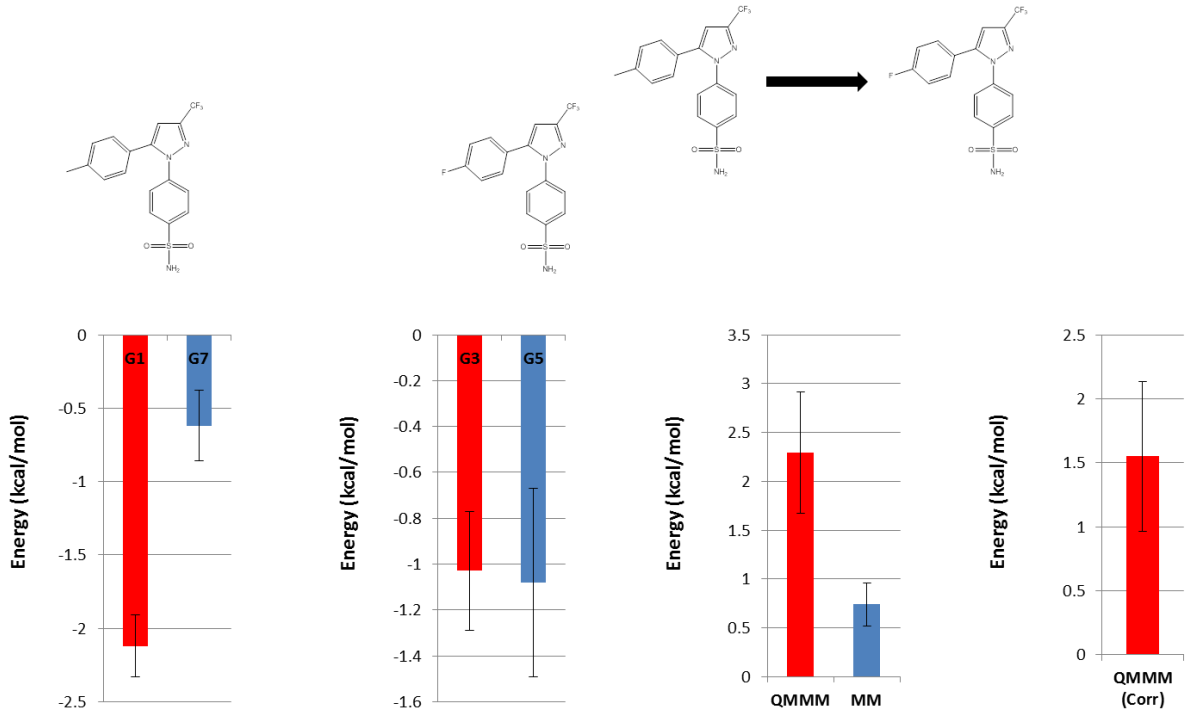


Figure 6.15. MM→QM/MM free energy breakdown for ligand 1 → ligand 9 perturbation. G1 refers to the QM/MM bound leg correction for ligand 1 and G7 refers to the QMM free leg correction for ligand 1. G3 is the QM/MM bound leg correction for ligand 9 and G5 the QM/MM free leg correction for ligand 9. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

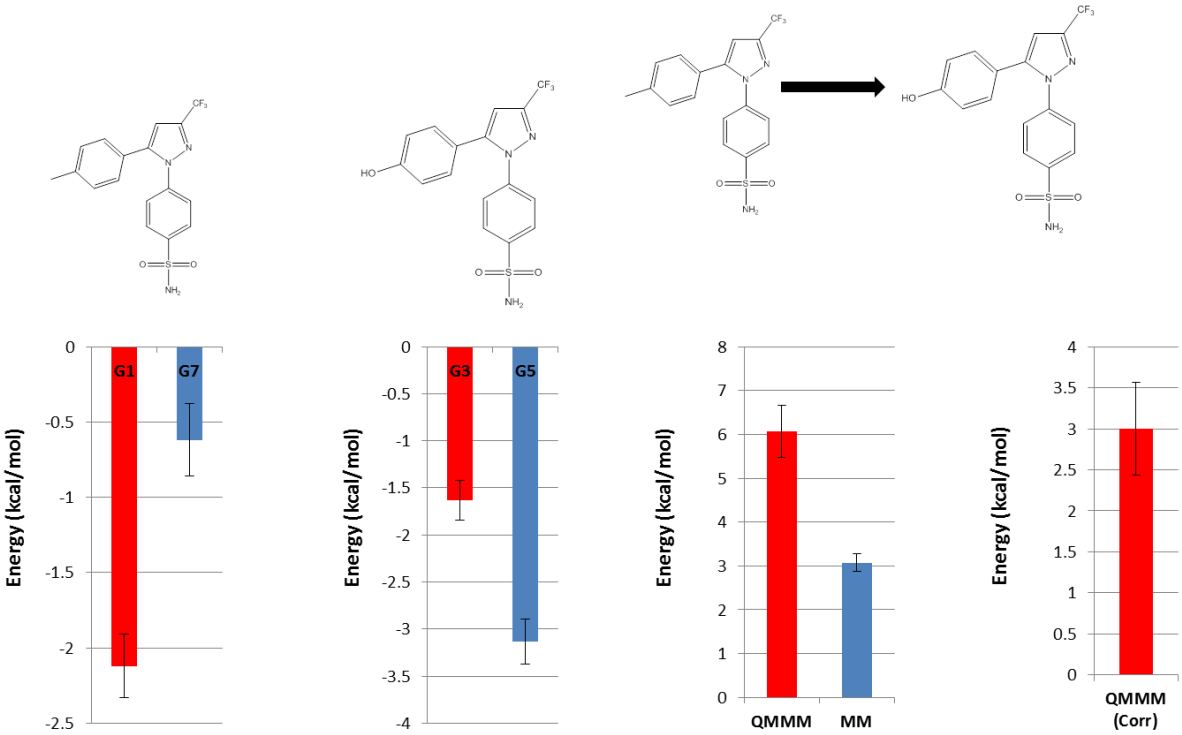


Figure 6.16. MM→QM/MM free energy breakdown for ligand 1 → ligand 7 perturbation. G1 refers to the QM/MM bound leg correction for ligand 1 and G7 refers to the QMM free leg correction for ligand 1. G3 is the QM/MM bound leg correction for ligand 7 and G5 the QM/MM free leg correction for ligand 7. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

Ligands 10, 9, and 7 follow the same trend by which the QM/MM correction free energies show that each of these ligands are less favoured in the COX-2 binding site and more favoured in aqueous solution when compared to the reference ligand (ligand 1). For the bound state QM/MM corrections, this trend may be caused by the size of the pocket into which we are perturbing in the COX-2 binding site. This pocket (Figure 6.17) is a small hydrophobic pocket between TRP384 and TYR385.

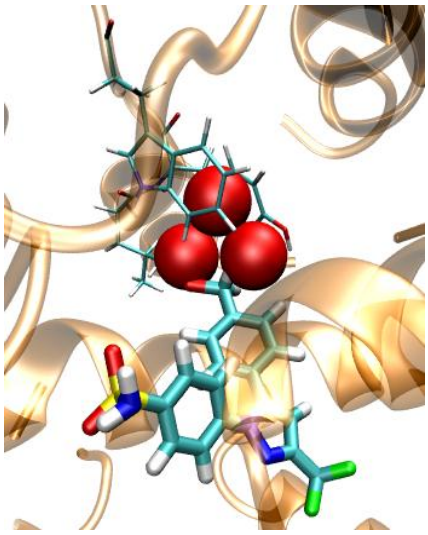


Figure 6.17. The structure of COX-2 (1CX2) and the hydrophobic pocket in which all perturbations are directed. The COX-2 inhibitor (ligand 7) is shown (thick licorice) key binding site residues TRP384 and TYR385 are shown (thin licorice) and the secondary structure. is also shown (orange ribbons) The size of this pocket is shown by three 2 $Å^3$ spheres (red spheres).

Ligand 1 contains a methyl in the R position which is able to fit into this hydrophobic pocket leading to a favourable QM/MM correction equal to -2.23 $kcal.mol^{-1}$. For ligands 10, 9 and 7 the R group is perturbed to a H, F, and OH respectively. All of these changes are to less hydrophobic substituents, which do not 'fit' as well as the original methyl into this pocket. Owing to this the substituted R groups are further from the protein residues, and hence further from the point charges in our QM/MM representation. This could be leading to our QM ligands becoming less polarised by the protein environment, and hence a lower QM/MM correction is obtained for these perturbations. For example, ligand 10 obtains a bound QM/MM correction of -0.69 $kcal.mol^{-1}$ with ligand 9 and 7 obtaining a QM/MM correction of -1.12 $kcal.mol^{-1}$ and -1.61 $kcal.mol^{-1}$ respectively. These changes in QM/MM energy for the bound state

suggests that the size of the perturbed group plays a crucial role in determining the overall MM→QM/MM binding free energy for each ligand.

In the free states ligands 10, 9, and 7 have R groups which can interact more strongly, or simply allow water molecules to be closer to the ligand than the methyl group of ligand 1. As was shown in the hydration free energy study, close contacts between water molecules and polarisable groups in our QM ligand can lead to a greater polarisation effect on our QM ligand, leading to larger QM/MM corrections for these ligand molecules. If we consider the QM/MM correction for the free energy of hydration of our reference ligand -0.77 kcal.mol^{-1} then we can see that ligands 10 and 9 obtain slightly more favourable QM/MM energies of -1.53 kcal.mol^{-1} and -1.35 kcal.mol^{-1} respectively. In the case of ligand 7, this change is much more dramatic with a QM/MM correction of -3.05 kcal.mol^{-1} for the free state. This indicates that polar substituents in position R will obtain a much more favourable QM/MM correction in the free state. This supports previous hydration free energy results, which suggest that polar QM compounds become highly polarised by the MM environment (partial charges), leading to large QM/MM corrections for such ligands.

In contrast, ligand 6 (CF_3) shows more favourable QM/MM corrections for both bound and free states when compared to ligand 1 (Figure 6.18).

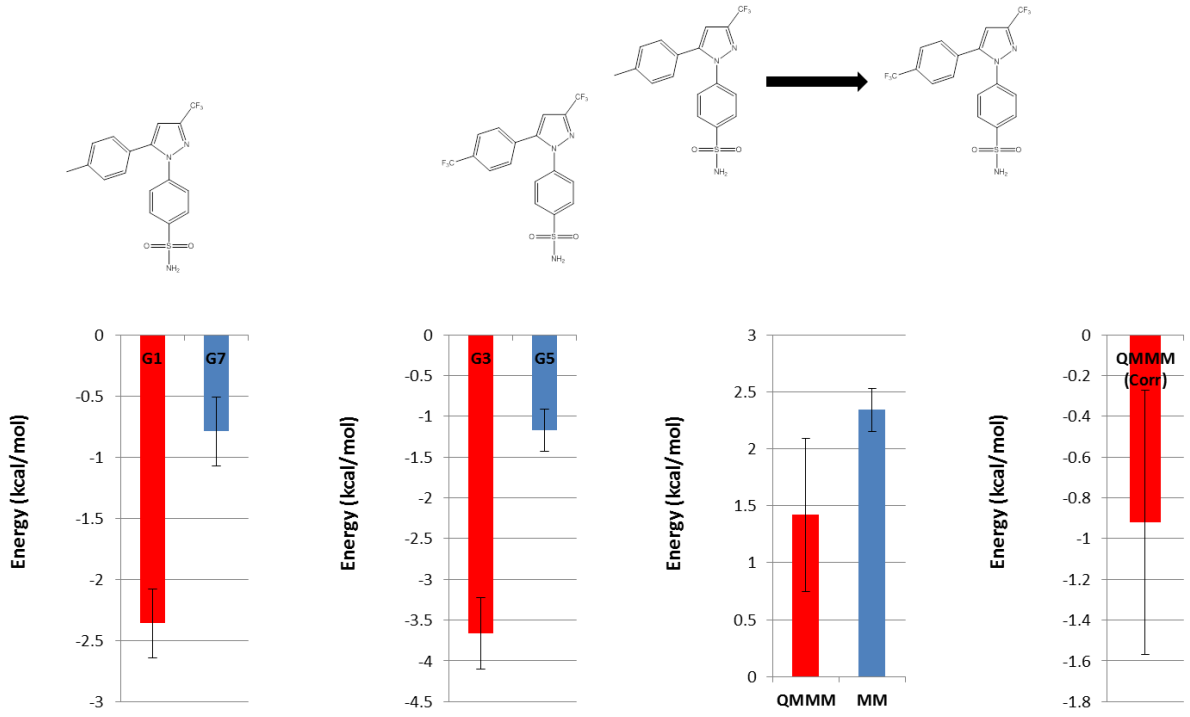


Figure 6.18. MM→QM/MM free energy breakdown for ligand 1 → ligand 6 perturbation. G1 refers to the QM/MM bound leg correction for ligand 1 and G7 refers to the QMM free leg correction for ligand 1. G3 is the QM/MM bound leg correction for ligand 6 and G5 the QM/MM free leg correction for ligand 6. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

The QM/MM correction for the bound state is equal to -3.66 kcal.mol^{-1}, which is the largest QM/MM correction for this dataset, and it is also 1.40 kcal.mol^{-1} more negative than the reference ligand. This could be an artefact of the slightly polar nature of the OH of TYR385 interacting with the polar hydrophobic fluorine of the CF_3 (Figure 6.19).

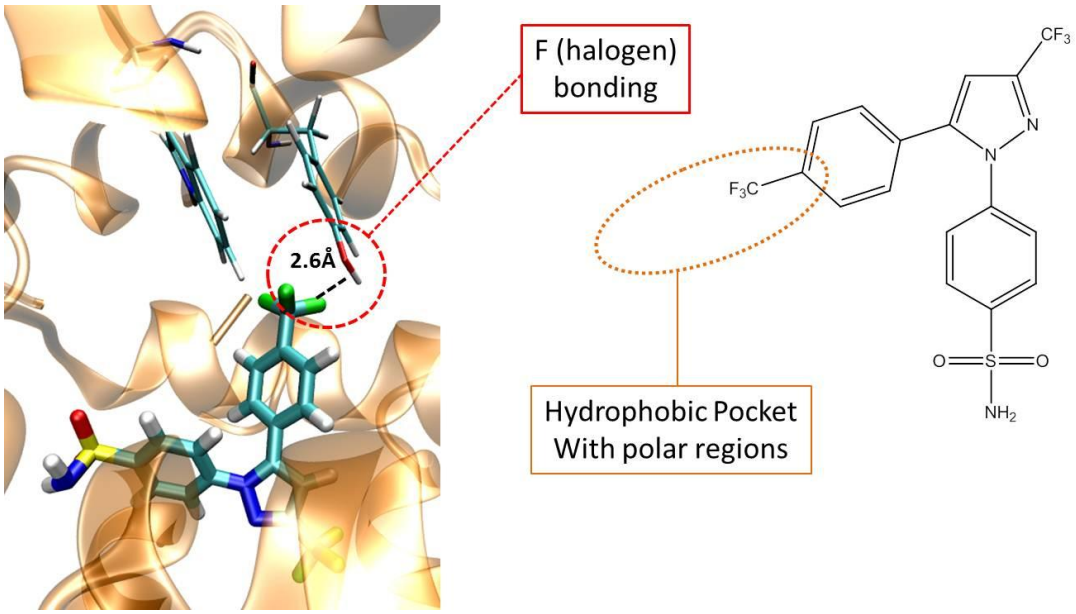


Figure 6.19. Interactions between ligand 6 (thin licorice) and key binding site residues TRP384 and TYR385 (thin licorice) taken from a representative snapshot of this simulation.

This could imply that our QM/MM approach can capture the 'polar-hydrophobicity' of fluorine, although more observations are needed before any such conclusion can be drawn.

In contrast to these poor results, there are some successes in this dataset. For example, ligand 2 ($R=CH_2CH_3$) performs well in both MM and MM→QM/MM free energy studies. Analysis of the energetic components of this perturbation highlights the minor change between MM and QM/MM representations of our system (Figure 6.20).

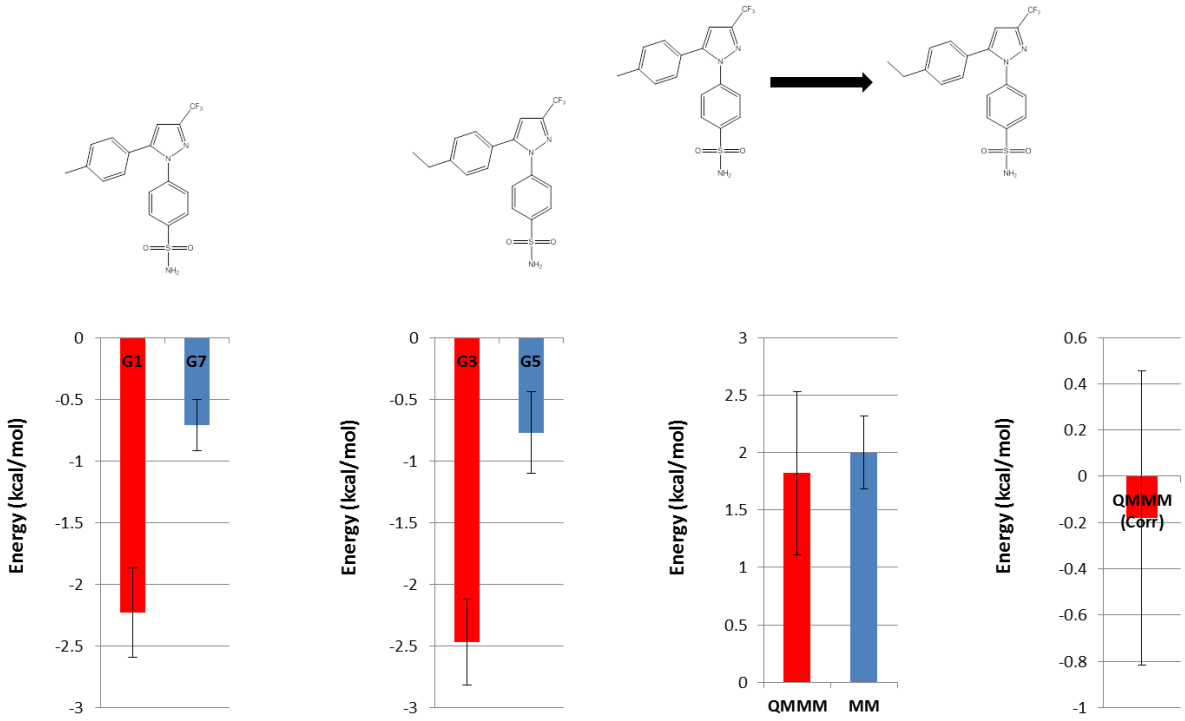


Figure 6.20. MM→QM/MM free energy breakdown for ligand 1 → ligand 2 perturbation. G1 refers to the QM/MM bound leg correction for ligand 1 and G7 refers to the QMM free leg correction for ligand 1. G3 is the QM/MM bound leg correction for ligand 2 and G5 the QM/MM free leg correction for ligand 2. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

It is clear to see that for the bound and free QM/MM corrections for ligand 2, -2.47 kcal.mol^{-1} and -0.77 kcal.mol^{-1}, there is very little change compared to the reference compound, -2.23 kcal.mol^{-1} and -0.71 kcal.mol^{-1}. These small corrections can be attributed to the non-polar nature of the R group as this produces a QM/MM correction which only slightly prefers the bound state and shows very similar values in the free state when compared to the reference compound.

Therefore, we can see from our results that very small non-polar changes to our ligand can lead to good MM→QM/MM corrections. Unfortunately, it appears that any change involving large differences in polarisation appear to perform poorly. This could indicate several problems: first, the results could be suffering due to inaccuracies in the forcefield. Second, the sampling of the system may not be sufficient or the single step QM/MM approach taken here may not be accurate enough to capture the subtleties of each ligand interactions with the COX-2 binding site. Lastly, it may indicate the need to include key binding site residues in the QM/MM representation of the system in order to accurately capture the changes in polarisation and other key interactions of each differing R group with this region.

6.4 Protein-ligand Charge Perturbations

As this QM/MM approach neglects any QM/MM sampling, the pathway-independence of the free energies obtained must be investigated. This is achieved through the use of charge perturbation pathways (see section 4.2). For COX-2, three compounds were selected to perform charge perturbations (Figure 6.21).

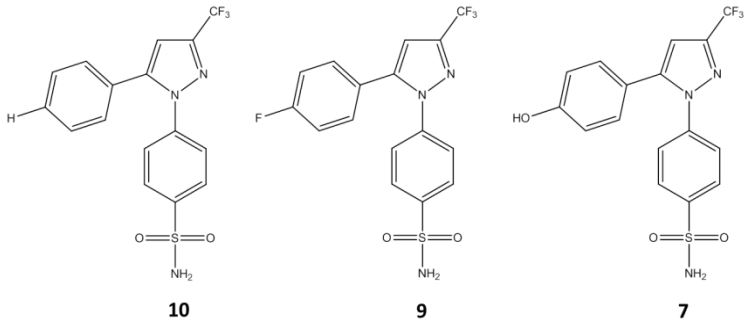


Figure 6.21. Three COX-2 ligands chosen for charge perturbation analysis.

Monte Carlo Simulation Protocol

Generating alternative pathways by scaling solute charges up or down can be used to validate the pathway independence of calculated free energies. Alternative configurations were generated by performing RETI [51, 52] calculations in which the solute charges were scaled up. For protein-ligand systems the charges of the solute-environment interactions were scaled, while the charges used for the solute internal energy computation remained at their un-scaled level (λ=0). 16 λ windows (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82, 0.88, 0.94, and 1.00) were used to ensure smooth transition between the two end states. The following charge scale factors were investigated: 1.01, 1.05, 1.07, 1.10, 1.15, 1.20 (N.B. 1.00 implies a simulation with non-perturbed charges). 10 million equilibration moves were performed before collecting statistics for 320 million moves in the bound legs and 160 million in the free legs, with each free energy simulation repeated 4 times. The RETI values shown are the mean of these four repeats and the standard error for these different runs is also shown.

QM/MM Single Point Energy Protocol

As with the 'normal' perturbations, configurations from the endpoint (λ=0 and λ=1) of the classical free energy simulations were selected and used as input for DFT-QM/MM single point energy calculations with Gaussian 09 [126]. One QM/MM single point energy calculation was performed every 100000th MC move. Therefore we took 3200 QM/MM configurations for the bound legs and 1600 QM/MM configurations for the free legs per repeat. In total we performed 12800 QM/MM single point energy

calculations for our bound legs (at λ=0 and λ=1) and 6400 QM/MM single point energy calculations for the free legs (at λ=0 and λ=1).

The QM energies were computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

Results & Discussion

The free energies obtained from the charge perturbations for COX-2/ligand 10 are shown in Figures 6.22 – 6.23.

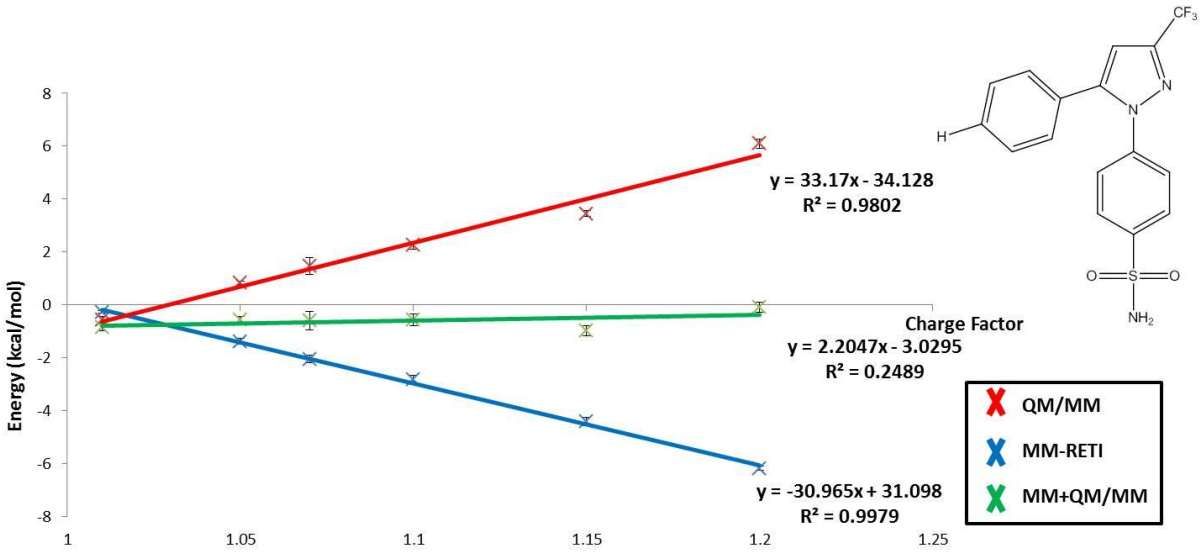


Figure 6.22. Charge perturbation results for ligand 10 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

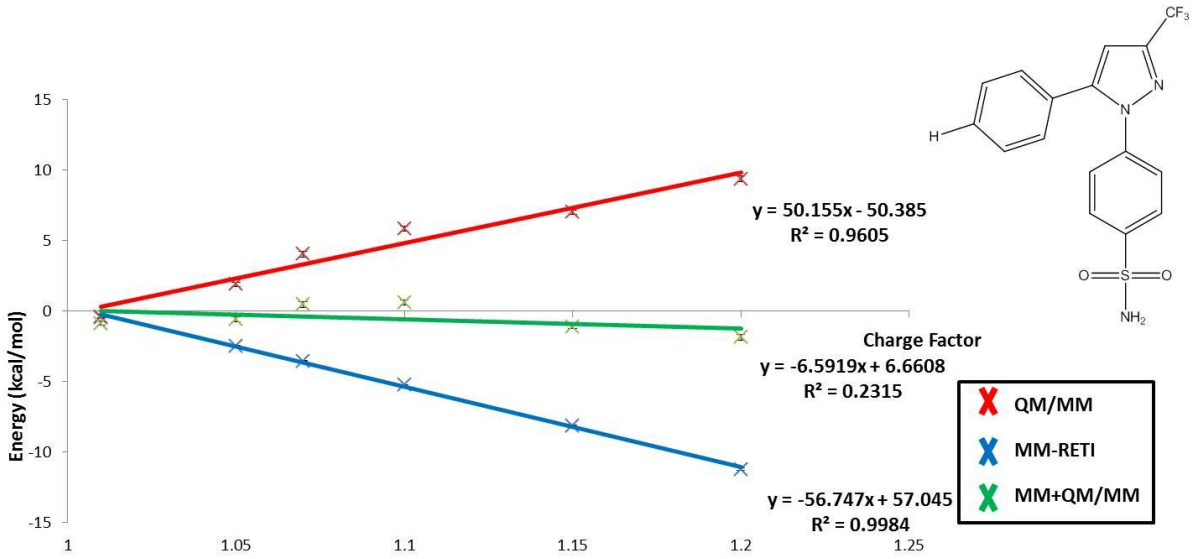


Figure 6.23. Charge perturbation results for ligand 10 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

In Figures 6.22 – 6.23, the sums of the charge perturbed free energy cycles are generally small (small slopes of the green fitted lines). However, if the free energies are pathway independent, the mean values of these sums calculated over all scale factors must be equal to that of the non-charge perturbation MM→QM/MM free energies. The relevant free energies are summarised in Tables 2.4 and 2.5 in Supporting Information 2. The above condition is not fulfilled by ligand 10 in the free state. For ligand 10 in the free legs the average free energy of cycle closure is -0.57 (0.26) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -1.58 (0.29) kcal.mol^{-1} it is clear that our cycle does not obtain the same value. For the bound state of ligand 10 the above condition is fulfilled. Ligand 10 obtains an average free energy cycle closure of -0.61 (0.15), which is very similar to the original

MM→QM/MM free energy of -0.73 (0.31) kcal.mol^{-1}. The poor agreement for ligand 10 in the free state would suggest that additional sampling is needed in order to obtain more precise results for this ligand in the aqueous phase.

The free energies obtained from the charge perturbations for COX-2/ligand 9 are shown in Figures 6.24 – 6.25.

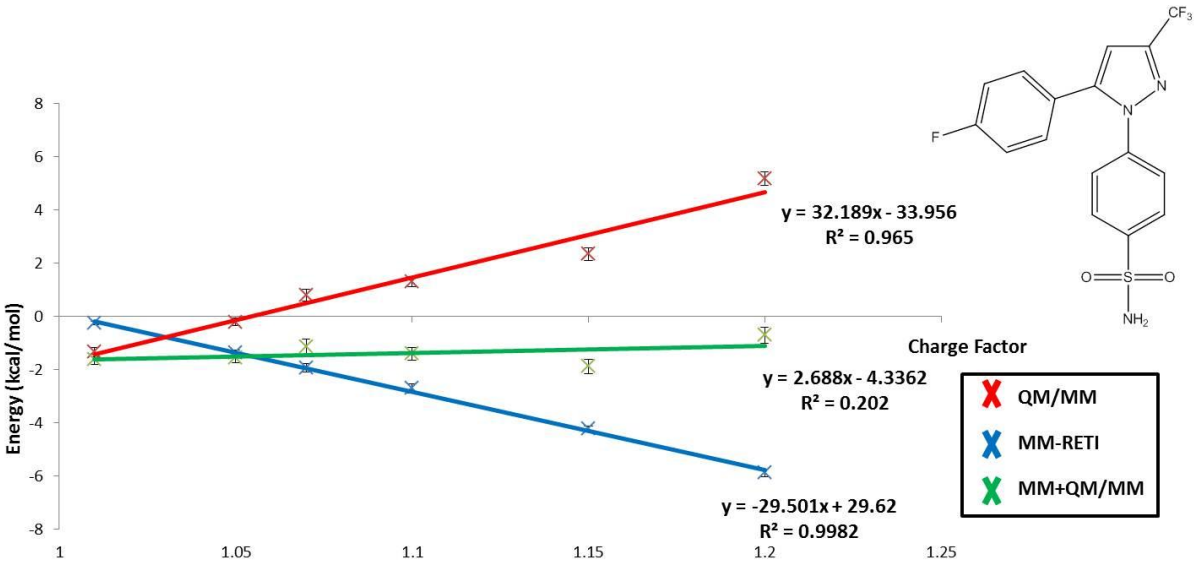


Figure 6.24. Charge perturbation results for ligand 9 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

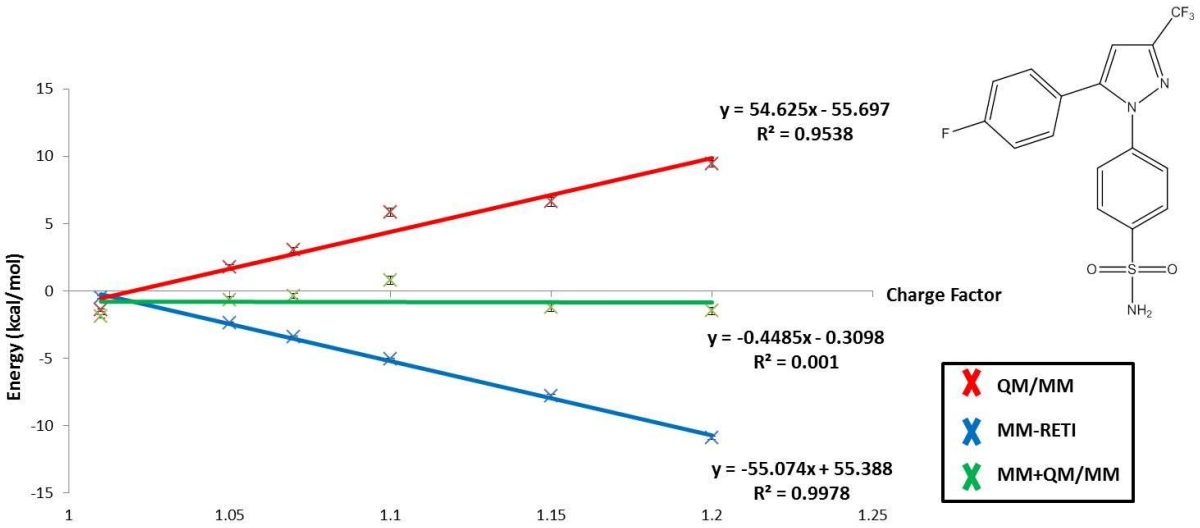


Figure 6.25. Charge perturbation results for ligand 9 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

In Figures 6.24 – 6.25, the sums of the charge perturbed free energy cycles are generally small (small slopes of the green fitted lines). However, if the free energies are pathway independent, the mean values of these sums calculated over all scale factors must be equal to that of the non-charge perturbation MM→QM/MM free energies. The relevant free energies are summarised in Tables 2.6 and 2.7 in Supporting Information 2. The above condition is fulfilled by ligand 9 in the free state. For ligand 9 in the free legs the average free energy of cycle closure is -0.81 (0.26) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -1.38 (0.25) kcal.mol^{-1} it is clear that our cycle does. For the bound state of ligand 9 the above condition is fulfilled. Ligand 9 obtains an average free energy cycle closure of -1.38 (0.25), which is very similar to the original MM→QM/MM free energy of -1.07 (0.36) kcal.mol^{-1}.

The free energies obtained from the charge perturbations for COX-2/ligand 7 are shown in Figures 6.26 – 6.27.

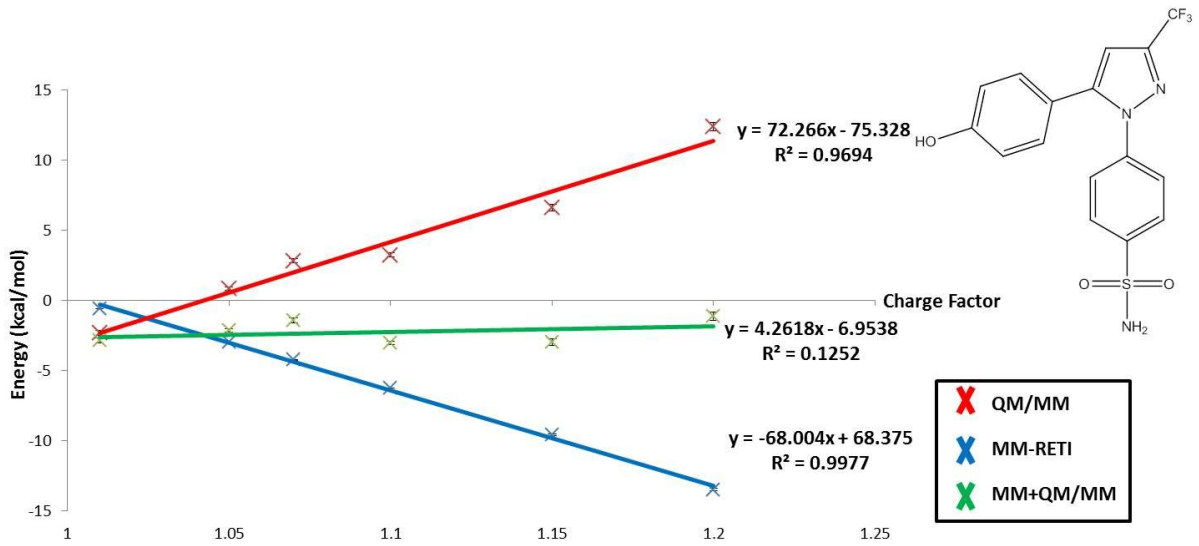


Figure 6.26. Charge perturbation results for ligand 7 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

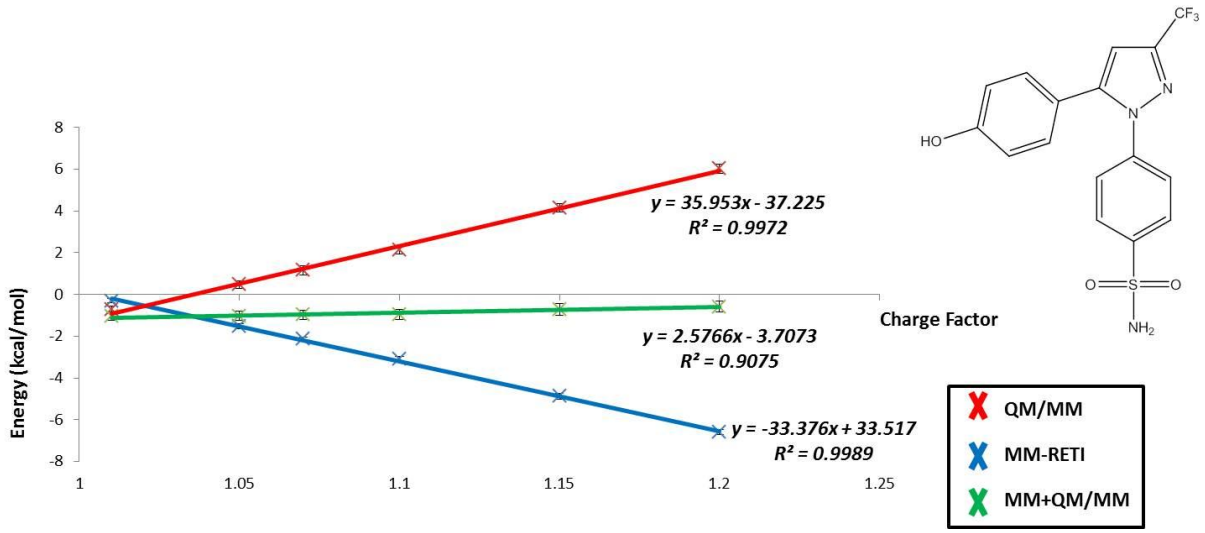


Figure 6.27. Charge perturbation results for ligand 7 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

In Figures 6.26 – 6.27 the sums of the charge perturbed free energy cycles are generally small (small slopes of the green fitted lines). However, if the free energies are pathway independent, the mean values of these sums calculated over all scale factors must be equal to that of the non-charge perturbation MM→QM/MM free energies. The relevant free energies are summarised in Tables 2.8 and 2.9 in Supporting Information 2. The above condition is not fulfilled by ligand 7 in the free state. For ligand 7 in the free legs the average free energy of cycle closure is -2.24 (0.28) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -3.09 (0.28) kcal.mol^{-1} it is clear that our cycle does not obtain the same value. For the bound state of ligand 7 the above condition is again not fulfilled. Ligand 7 obtains an average free energy cycle closure of -0.88 (0.24), which is not similar to the original MM→QM/MM free energy of -1.62 (0.26) kcal.mol^{-1}. The poor agreement for ligand 7

in the free and bound states would suggest that additional sampling is needed in order to obtain more precise results for this ligand in both the aqueous and bound free energy legs.

6.5 Conclusions

The aim of this study was to understand if we could apply a simple and fast QM/MM method to obtain accurate MM→QM/MM binding free energies for a set of COX-2 inhibitors. The results from this study suggest that application of the QM/MM method employed to correct classically obtained MM free energies leads to large inaccuracies between MM→QM/MM calculated free energies and experiment. Unfortunately any good QM/MM corrected free energies appear to be fortuitous, suggesting that either the sampling of our system was not performed for long enough or that our single step QM/MM approach may not be accurate enough to capture the subtleties of the ligand interactions with the COX-2 binding site. This hypothesis is supported via the charge perturbation study, which suggests that the pathway independence of the MM→QM/MM free energies is not maintained for several of the examples studied. Again, this problem could potentially be remedied by extending the sampling of the system; however, due to the already extensive sampling of the system it can be suggested that increasing the sampling may not be the best route to take due to the large timescales involved in enhancing this process. These poor results could also have been caused by inaccuracies in the forcefield used to describe our protein. This could imply that the inclusion of key binding site residues as part of the QM region in the QM/MM representation of the system could be needed in order to more accurately capture the changes in polarisation and other key interactions of each differing R

group with this region. However, this would lead to greater difficulties due to the complex nature of such QM/MM coupling schemes.

7 Calculation of QM/MM Binding Free Energies for 9 Neuraminidase Inhibitors

7.1 Biological Relevance

The impact of the influenza virus is felt every year, with over 20% of the world's population contracting the virus every winter [155]. Recent cases, including both swine and avian flu outbreaks, have heightened the awareness to the life threatening nature of a pandemic flu outbreak. Although vaccination is the primary treatment for influenza, there are a number of likely situations which make vaccination inadequate and effective anti-viral drugs would become essential to minimize the impact of any influenza outbreak. This situation is made more complex due to influenza's ability for antigenic drift, which would lead to a significant loss of potency for any pre-designed vaccines [156]. This makes anti-viral agents an important area of research for a rational approach to treat epidemic influenza and a critical area of planning for any influenza pandemics.

Currently there are four anti-viral drugs on the market for the treatment of influenza; the adamantanes (adamantine and rimantadine) and the newer class of neuraminidase inhibitors zanamivir (Relenza) and osteltamivir (Tamiflu) (Figure 7.1). The adamantanes interfere with the viral uncoating inside the cell [155]. They are only effective against influenza A and are associated with several toxic side effects and with rapid emergence of drug resistant strains [157].

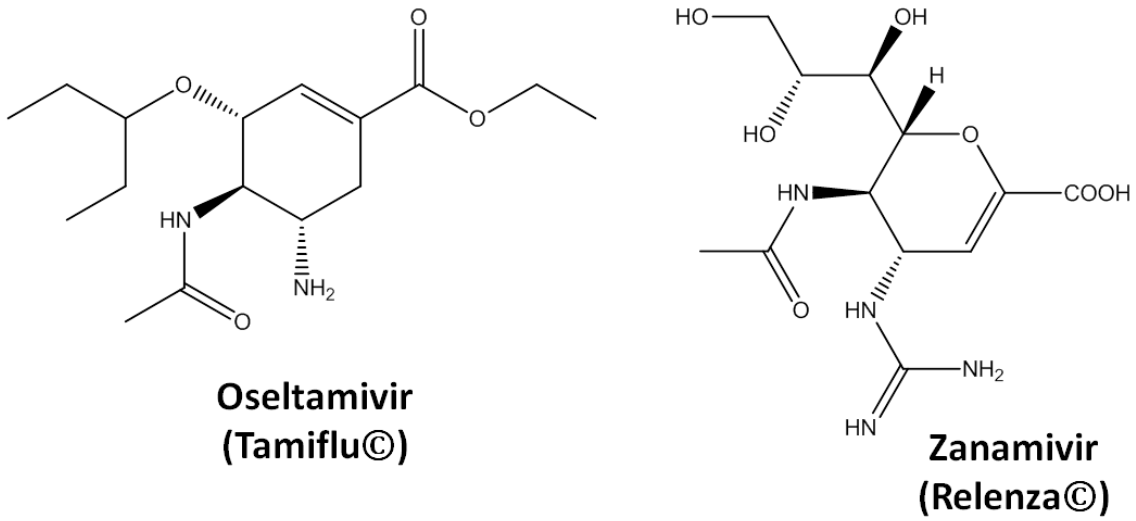


Figure 7.1. Structures for two neuraminidase inhibitors, Oseltamivir (Tamiflu) from Roche and Zanamivir (Relenza) from GlaxoSmithKline.

The newer class of neuraminidase inhibitors interfere with the release of progeny influenza virus from infected host cells. This process prevents infection of new host cells and thereby halts the spread of infection in the respiratory tract. In contrast to the adamantane inhibitors, the neuraminidase inhibitors are associated with very little toxicity and are far less likely to promote drug resistant strains of influenza [158].

Previous efforts have been made within our group to predict the relative binding free energies for 9 neuraminidase inhibitors (Figure 7.2) [146]. The results from this study showed a correlation (R^2) of 0.79 between MM(AMBER99/GAFF/AM1-BCC)-RETI calculated binding free energies and experimental data.

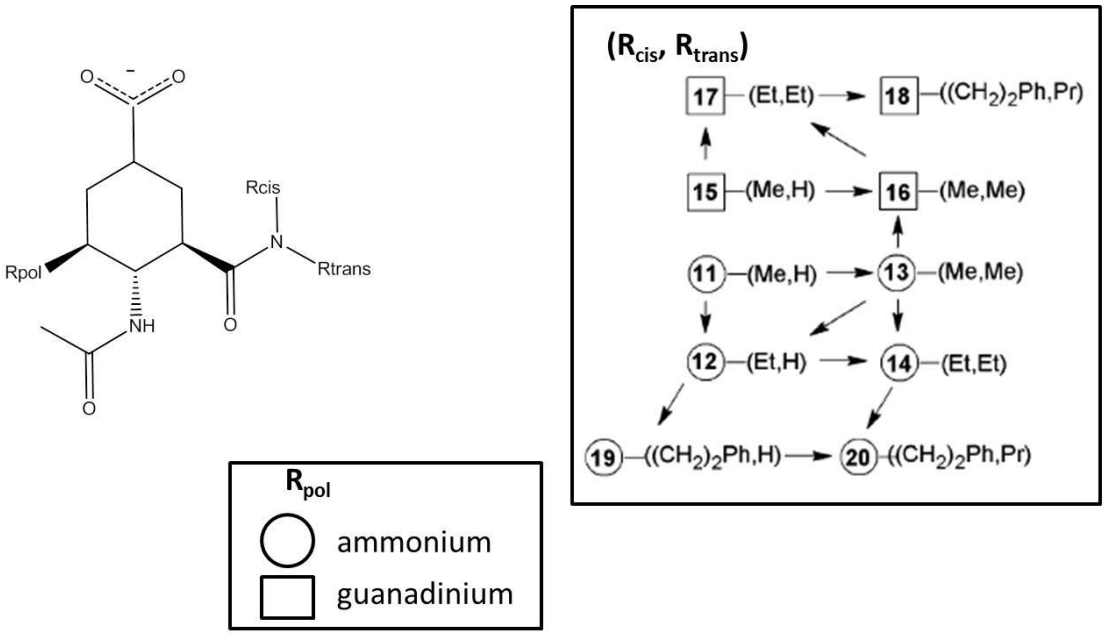


Figure 7.2. Perturbation network for the set of neuraminidase inhibitors studied here. The boxed and circled numbers denote the nature of Rpol. The boxed numbers are guanadinium based inhibitors, whereas the circled numbers are for ammonium based inhibitors. The corresponding R, R, represent the R groups at Rcis and Rtrans.

7.2 System Preparation

Protein – ligand setup

The PDB structure of N9 neuraminidase (PDB code 1BJI) [159] was selected as a starting point for this study. Hydrogen atoms were added to this structure using the Reduce software package [148]. The protonation states of histidines were determined via visual inspection. The protein was parameterised using the AMBER99 force field, [13] inhibitors were parameterised with the GAFF force field [14] and the partial atomic charges were derived using the AM1-BCC method [124], as implemented in the AMBER 10 suite. To avoid bad steric clashes, the protein-ligand complex (1BJI/ligand 20) was minimised in the SANDER module of AMBER 10 with a generalised Born solvent model. The backbone of the protein was subsequently fixed for Monte Carlo

simulations, which were performed using a modified version of ProtoMS2.2 [123]. To reduce computational cost, only protein residues that contained one heavy atom within 15 Å of any representative ligand atom were retained. The resulting protein scoop contained 145 residues. The ligands were modelled in the binding site based upon the binding mode predicted by the docking program GOLD [116], the binding modes for each ligand were generated by Michel *et al.* [146] Crystallographic waters were retained and the complex was hydrated by a sphere of TIP4P [125] water molecules of 22 Å radius and centred on the geometric centre of the ligand. To prevent evaporation, a half-harmonic potential with a 1.5 kcal.Å^{-2} force constant was applied to water molecules whose oxygen atom distance to each ligands centre of geometry was greater than 22 Å. A similar sphere of water was used for the unbound state.

Monte Carlo Simulation Protocol

The bond angles and torsions for the side chains of residues within 10 Å of any ligand heavy atom and all bond angles and torsions of the ligand were sampled during the simulation, with ring structures being the only exception. The bond lengths of the residues and ligand were constrained. The total charge of the system was brought to zero by neutralising lysine residues 273 and 432 lying in the outer 'frozen' part of the scoop. The neutralised lysines were then re-modelled using the AMBER99 forcefield. A 10 Å residue based cut-off was employed in all simulations.

For simulation in the bound state, solvent moves were attempted with a probability of 76.21%, protein side-chain movements with a probability of 21.79% and solute moves with a probability of 6.42%. In the unbound state, solvent moves were attempted

99.06% of the time. Replica exchange moves were attempted every 200000 moves. The solvent was equilibrated for 20 million moves to remove any bad contacts with the solute. The system was then equilibrated at one state (the end state with the larger solute) for 20 million further moves where solute, protein, and solvent moves were attempted. The resulting configuration was distributed over the 16 values for the coupling parameter λ (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82, 0.88, 0.94, and 1.00) and equilibrated for 10 million moves before collecting statistics for 640 million moves (bound) and 320 million moves (free).

QM/MM Single Point Energy Protocol

Configurations from the endpoint (λ=0 and λ=1) of the classical free energy simulations were selected and used as input for DFT-QM/MM single point energy calculations with Gaussian 09 [126]. One QM single point energy calculation with background charges representing the solvent and protein residues within our cut-off (Gaussian keyword 'CHARGE') were performed every 100000^{th} MM MC moves, with symmetry operations disabled (Gaussian keyword 'NoSymm'). This gave a total of 6400 QM/MM single points for each solute perturbation in the bound state and 3200 QM/MM single points for each solute perturbation in the free state. Gaussian calculations with embedded background charges allow a polarisation of the QM wave function via the MM charges, however no back polarisation of the MM part via the polarised QM wave function was considered.

The QM energies were computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

As the ligands were flexible we needed to compute the QM vacuum energies for each snapshot used. This was performed in Gaussian 09, but without the use of the 'CHARGE' and 'NoSymm' keywords, which are only necessary if embedding MM point charges in our calculation.

The QM vacuum energies were again computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

7.3 Results & Discussion

7.3.1 MM – RETI Results

The calculated relative MM-RETI binding free energies for the series of 9 neuraminidase inhibitors are shown in Figure 7.3. The coefficient of determination (R^2) between predicted and experimental binding free energies [146] is 0.79. The relevant free energies are summarised in Tables 3.1 and 3.3 of Supporting Information 3.

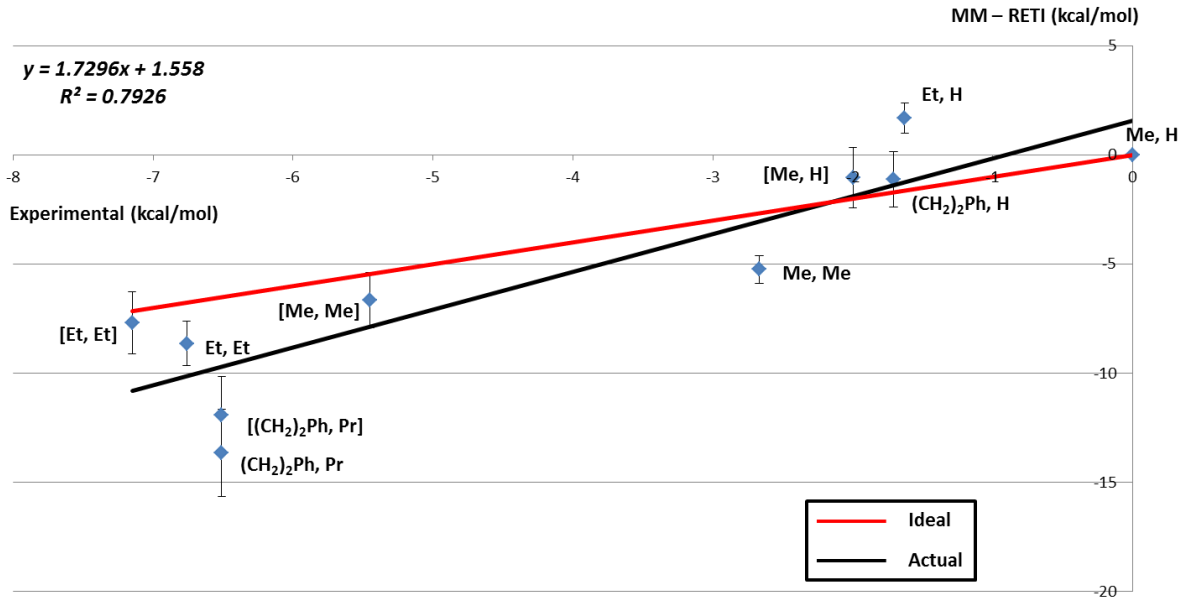


Figure 7.3. MM-RETI results versus experimental data [146]. The red line represents the ideal (1 to 1) correlation and the black line represents the best fit. The error bars shown were computed from four independent simulations suing standard error.

The MUE is equal to 2.63 kcal.mol^{-1}, which is relatively large. This discrepancy is caused by the the binding free energies of two potent inhibitors, 18 and 20, being overestimated. If these compounds are excluded, the MUE drops to 1.58 kcal.mol^{-1}. Despite this, the predictions for the dataset do accurately follow experimentally observed trends with the PI equal to 0.92. The τI is 0.78, again showing that an unweighted ranking of our calculated binding free energies still accurately captures experimentally observed trends.

To ensure smooth transitions between the two end states (λ=0 and λ=1) the free energy gradients for each perturbation were studied (Figure 7.4).

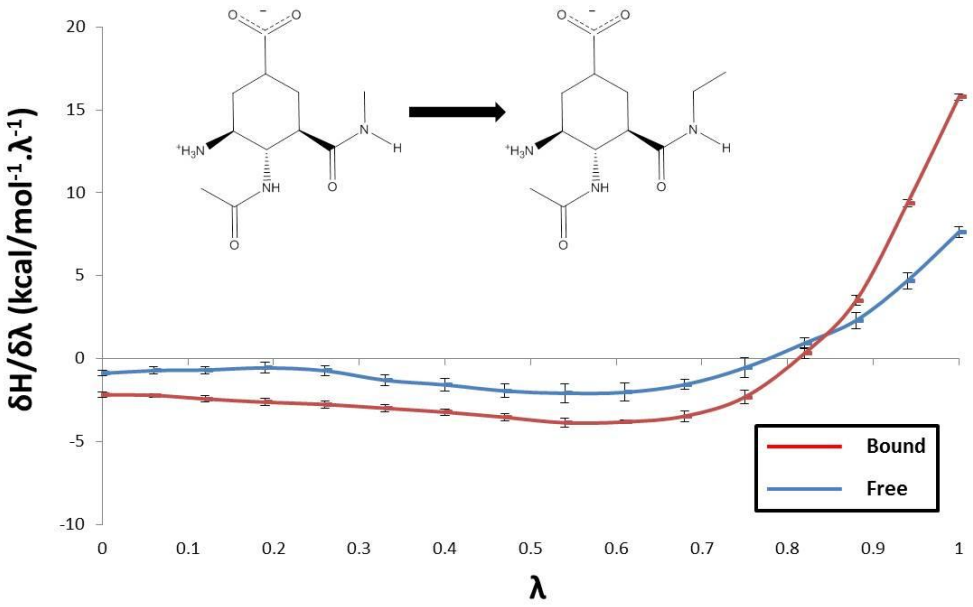


Figure 7.4. Free energy gradients for both free (blue line) and bound (red line) for the ligand 11 to ligand 12 perturbation. The error bars shown were computed from four independent simulations using standard error.

These show that for both the free and bound legs of the free energy simulations the transition across the reaction co-ordinate is very smooth. This indicates that the free energies obtained from the simulations are precise.

To analyse the statistical uncertainty in the free energy simulations the hysteresis for closing a binding free energy cycle for a set of 4 neuraminidase perturbations (12→14, 14→20, 19→20 and 12→19) was calculated. For the free legs (Figure 7.5) the hysteresis is very large at 3.15 (1.93) kcal.mol^{-1}. This shows that there is a high level of statistical uncertainty for the free legs of the MM-RETI free energy cylcles.

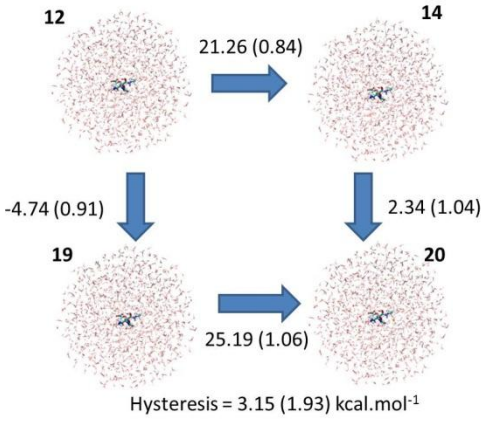


Figure 7.5. MM-RETI free leg hysteresis for four neuraminidase perturbations. The standard errors for each step are shown in the brackets.

The high level of uncertainty shown in the free legs is mirrored in the bound legs (Figure 7.6). The hysteresis is 2.89 (1.59) kcal.mol^{-1} which is very large.

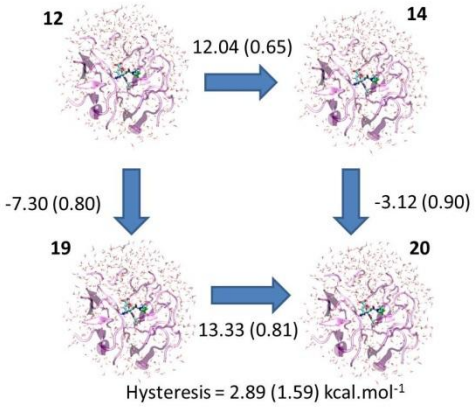


Figure 7.6. MM-RETI bound leg hysteresis for four neuraminidase perturbations. The standard errors for each step are shown in the brackets.

When combining the free and bound legs the overall hysteresis for these four pertubations is obtained (Figure 7.7).

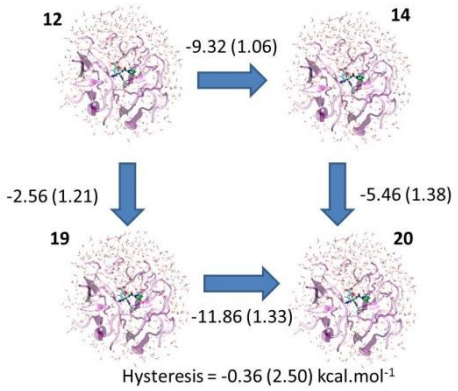


Figure 7.7. MM-RETI total hysteresis for four neuraminidase perturbations. The standard errors for each step are shown in the brackets.

The hysteresis for closing this binding cycle is small at just -0.36 (2.50) kcal.mol^{-1}, as illustrated in Figure 7.7. This is a notably small value, indicating that the statistical uncertainty in our free energy simualtions is low. This is a fortunate result as these errors indicate that this value can vary significantly. However, it is important to note the large hysteresis in the free and bound legs individually, which indicates that there is favourable error cancellations when combining them into a overall free energy cycle.

A difficulty arises in the perturbation of compound 13 into compound 16, where the ammonium is perturbed into a guanadinium. Crystallographic evidence suggests that the bulkier guanadinium group of 16 must expel a cyrstallographic water that is present when 13 is bound to neuraminidase (Figure 7.8) [160].

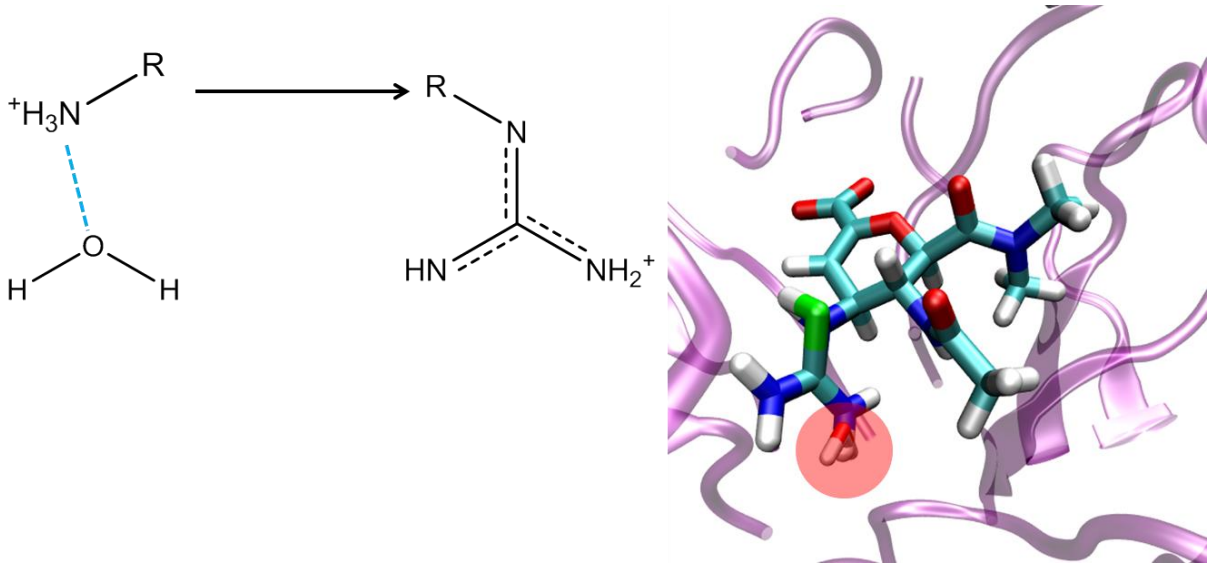


Figure 7.8. The displacement of a water molecule when perturbing Rpol from ammonium to guanadinium (red sphere on right hand side).

This would require the annihilation of the crystallographic water prior to the perturbation of 13 to 16. Such free energy simulations require a more elaborate treatmeant that is outside of the scope of the free energy calculations performed here. However, this system was studied by Michael Bodnarchuk using GCMC to calculate the binding affinities for waters bound to the holo form of Neuraminidase with an identical forcefield and protein motion used in this study [154]. The authors reported a binding affinity of -5.4 (1.1) kcal.mol^{-1} for this water within the neuraminidase binding site. This value was subsequently added to all of the necessary calculated protein-ligand binding free energies.

7.3.2 MM→QM/MM–FEP Results

The calculated relative MM→QM/MM binding free energies for the series of neuraminidase inhibitors is shown in Figure 7.9. The coefficient of determination (R^2) between predicted and experimental binding free energies [146] was 0.79. The

relevant free energies are summarised in Tables 3.2 and 3.3 of Supporting Information 3.

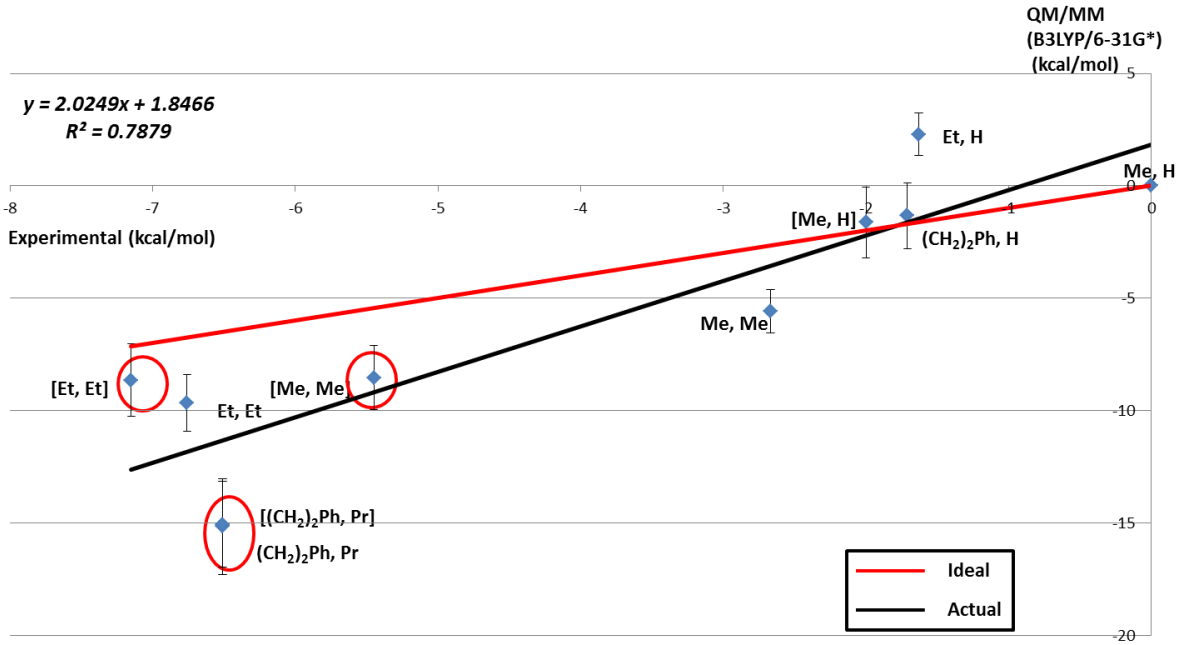


Figure 7.9. MM→QM/MM results versus experimental data [146]. The red line represents the ideal (1 to 1) correlation and the black line represents the actual correlation. The error bars shown were computed from four independent simulations using standard error.

The mean unsigned error (MUE) is equal to 3.59 kcal.mol^{-1}, which is over 1 kcal.mol^{-1} greater than the MUE for the MM-RETI calculated binding free energies. This discrepancy is caused by the MM→QM/MM binding free energies of two potent inhibitors, 18 and 20, which are futher overestimated in QM/MM compared to MM calculated free energies. If these compounds are excluded the MUE drops to 2.15 kcal.mol^{-1}. Despite this, the predicitons for the dataset do accurately follow experimaentally observed trends with the PI equal to 0.92. The τI is 0.78, again showing that an unweighted ranking of our calculated binding free energies still accurately captures experimentally observed trends.

As with our MM-RETI results the hysteresis for closing a binding free energy cycle for a set of 4 neuraminidase perturbations (12→14, 14→20, 19→20 and 12→19) was calculated. For the free legs of the MM→QM/MM binding free energy study (Figure 7.10) the hysteresis is 3.67 (2.23) kcal.mol^{-1} this over a 0.5 kcal.mol^{-1} increase from the MM values of 3.15 (1.93). This indicates that applying the QM/MM corrections leads to higher statistical undertainty and larger errors compared to standard MM.

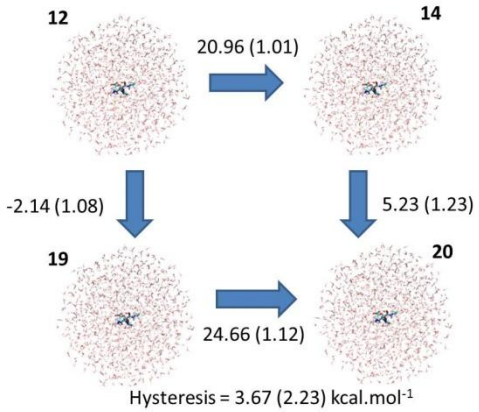


Figure 7.10. MM→QM/MM free leg hysteresis for four neuraminidase perturbations. The standard errors for each step are shown in the brackets.

When considering the bound legs of the MM→QM/MM binding free energy study (Figure 7.11) the hysteresis in the bound legs is 2.50 (3.70) kcal.mol^{-1} which is a slight decrease in the hysteresis shown in the MM free energy study of 2.89 (1.59) kcal.mol^{-1}. More noticeable is the increase in the error estimate for the MM→QM/MM bound leg hysteresis which has increased by 2.11 kcal.mol^{-1}.

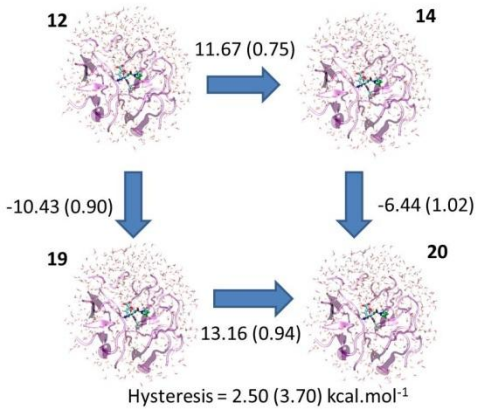


Figure 7.11. MM→QM/MM bound leg hysteresis for four neuraminidase perturbations. The standard errors for each step are shown in the brackets.

Combining the free and bound legs leads to the overall hysteresis for this free energy cycle (Figure 7.12). The overall MM→QM/MM hysteresis is -1.17 (3.82) kcal.mol^{-1} which is an increase 0.81 kcal.mol^{-1} compared to the hysteresis for MM of -0.36 (2.50) kcal.mol^{-1}.

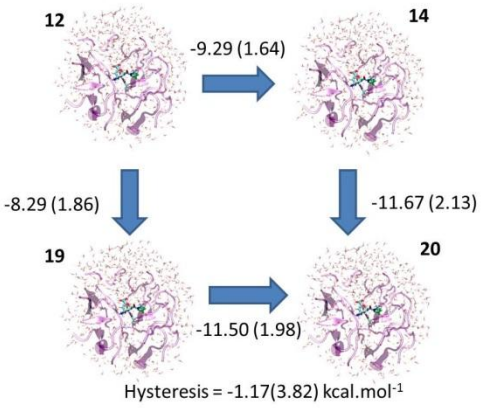


Figure 7.12. MM→QM/MM total hysteresis for four neuraminidase perturbations. The standard errors for each step are shown in the brackets.

This change in hysteresis between MM and MM→QM/MM suggests that the addition of the QM/MM corrections to the classically obtained free energies introduces statistical uncertainty. It also highlights the increased error estimations in the MM→QM/MM free energies, which is again indicative of adding 'noise' to the simulations through the use of QM/MM corrections.

To understand the changes in accuracy between our MM and MM→QM/MM-FEP binding free energies, in particular those perturbations circled (red circles) in Figure 7.9, analysis into the energies produced for each leg of the protein-ligand binding free energy cycle is needed.

The QM/MM corrections calculated for this dataset were very large, and much larger than for other datasets studied. For ligand 20 (Figure 7.13) it is clear that the size of the QM/MM corrections can lead to large changes between MM and MM→QM/MM binding free energies.

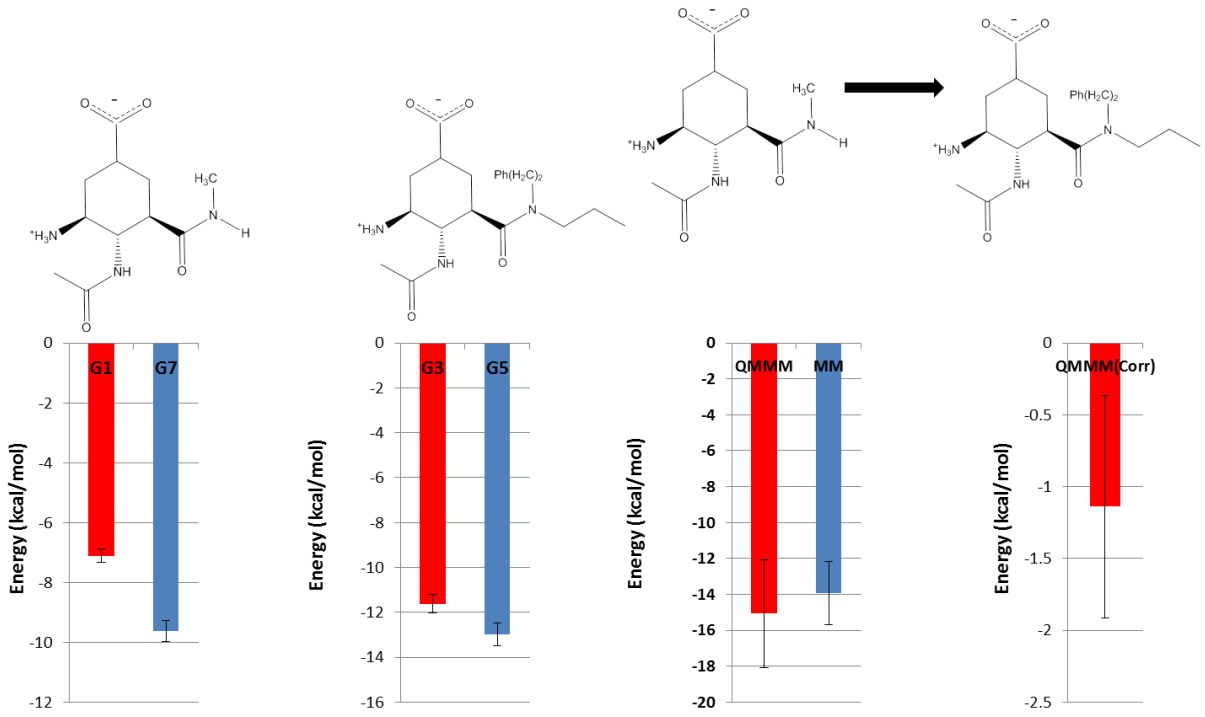


Figure 7.13. MM→QM/MM free energy breakdown for ligand 11 → ligand 20 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 11 and G7 refers to the QMM free leg correction for ligand 11. G3 is the QM/MM bound leg correction for ligand 20 and G5 the QM/MM free leg correction for ligand 20. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

The QM/MM corrections for ligand 20 in the bound and free legs are -11.62 kcal.mol^{-1} and -13.11 kcal.mol^{-1} respectively. These corrections are considerably larger than those for the reference compound (ligand 11), whose QM/MM corrections are -7.24 kcal.mol^{-1} (bound) and -9.77 kcal.mol^{-1} (free). The overall QM/MM correction for this perturbation is -1.15 kcal.mol^{-1}. This makes our MM→QM/MM binding free energy equal -15.07 kcal.mol^{-1} which is further from the experimental value of -6.51 kcal.mol^{-1} than our MM-RETI value of -13.92 kcal.mol^{-1}. This is not a very surprising result as the

MM-RETI simulations displayed over-polarised results for this dataset which lead to inaccurate results and by performing the QM/MM corrections it appears that the QM/MM adds to this over-polarisation effect, causing the MM→QM/MM results to become more inaccurate when compared to experiment. This is particularly prevelant in the free legs of the QM/MM simulations. This trend of large QM/MM corrections for free states is observed across all of the ligands studied here.

In addition to ligand 20, there is also a large negative shift between MM and MM→QM/MM for ligand 18 (Figure 7.14).

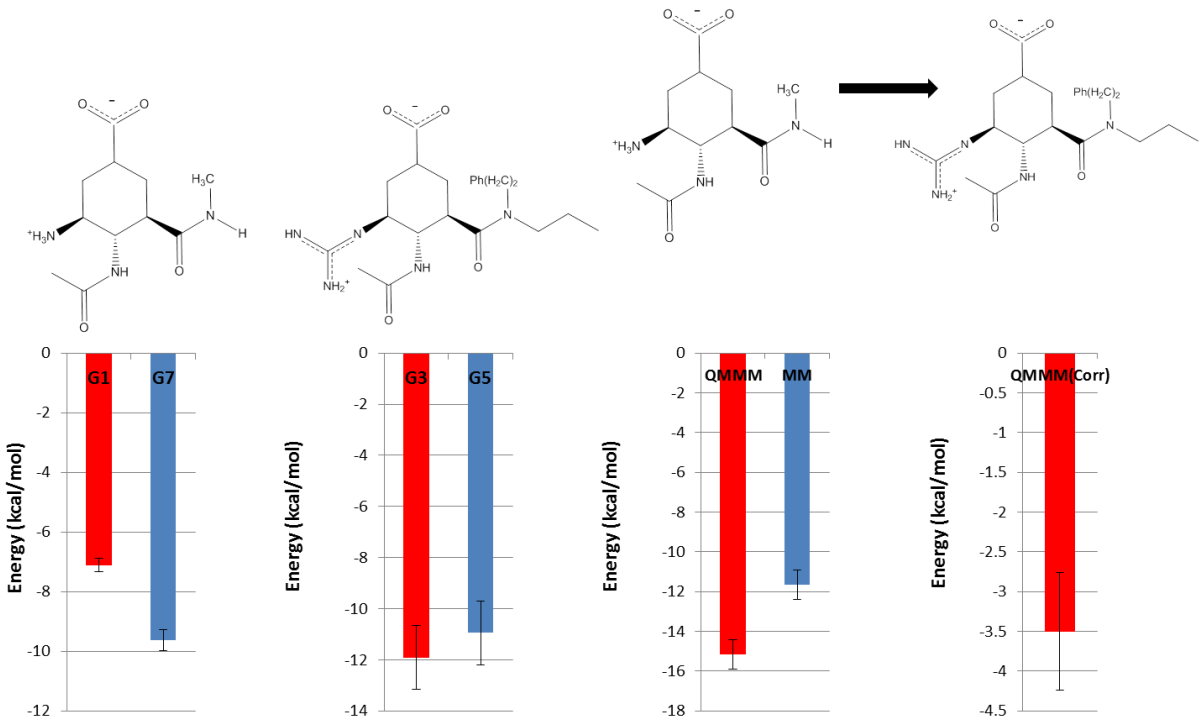


Figure 7.14. MM→QM/MM free energy breakdown for ligand 11 → ligand 18 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 11 and G7 refers to the QMM free leg correction for ligand 11. G3 is the QM/MM bound leg correction for ligand 18 and G5 the QM/MM free leg correction for ligand 18. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

The QM/MM corrections for the bound and free legs of ligand 18 are -11.91 kcal.mol^{-1} and -10.94 kcal.mol^{-1} which again show a large negative increase from the reference compound. The main driving force behind these changes is thought to be the perturbation of the ammonium group to guanadinium as this group forms a salt bridge to an aspartic acid (ASP325) within the neuraminidase binding site. This large shift in binding group is also thought to be the reason for the QM/MM binding free energy correction being more negative than that of the free state. This could also be caused by the extensive interactions between the large phenyl group and the neuraminidase binding site (Figure 7.15).

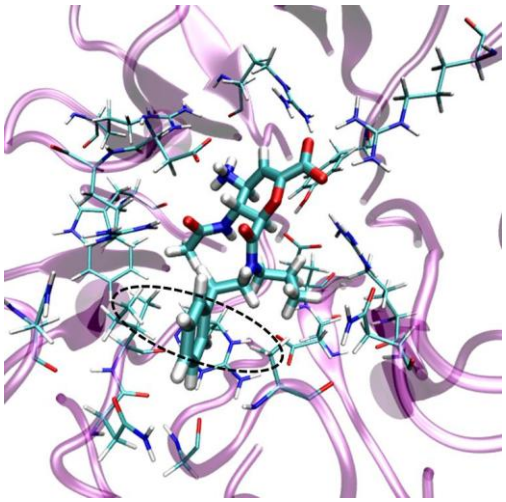


Figure 7.15. Interactions of large phenyl based inhibitor 19 (thick licorice) with neuraminidase (1BJI) binding site. Key binding site residues are shown in licorice and the secondary structure is shown as purple ribbons.

These large shifts in QM/MM corrections give an overall QM/MM correction of -3.5 kcal.mol^{-1} causing the overall MM→QM/MM binding free energy to become -15.06 kcal.mol^{-1} which is far more negative than the original MM-RETI free energy value of -11.66 kcal.mol^{-1}. Much like ligand 20, this large negative shift in binding free energy

leads to our MM→QM/MM becoming less accurate compared to experimental data for ligand 18 (-6.51 kcal.mol^{-1}).

The assumption that the large phenyl substituent can lead to large negative QM/MM corrections, especially for the bound legs, is strengthened when investigating the results of ligand 19 (Figure 7.16).

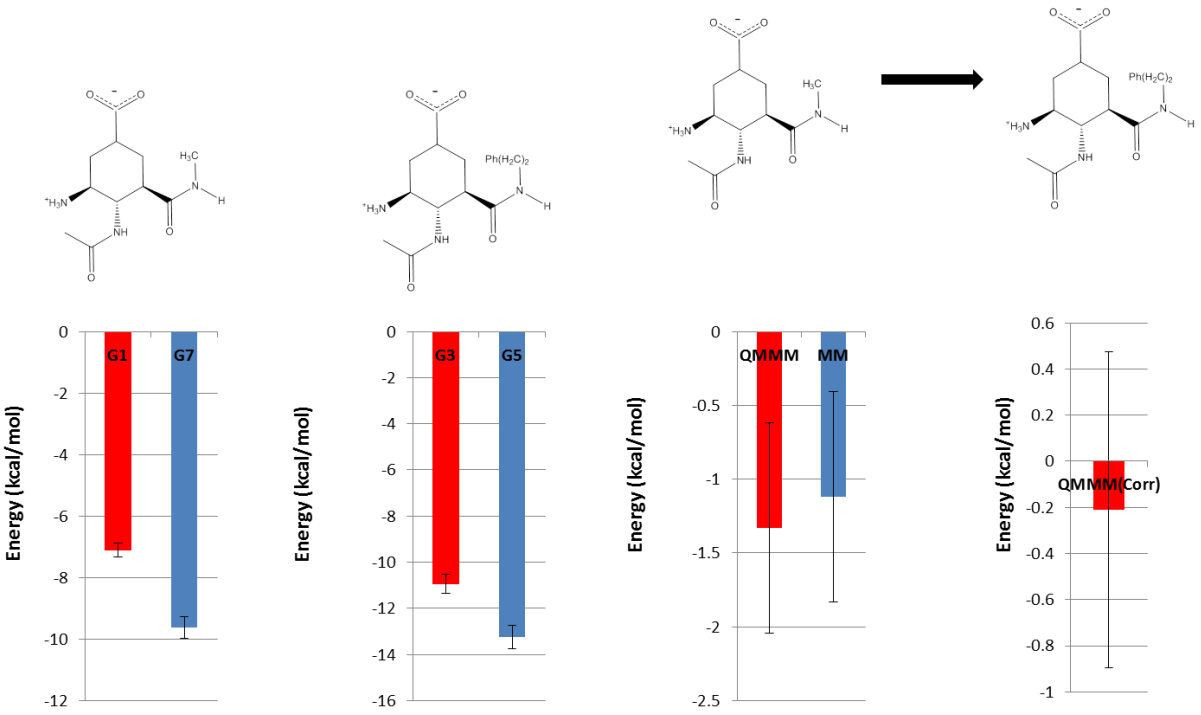


Figure 7.16. MM→QM/MM free energy breakdown for ligand 11 → ligand 19 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 11 and G7 refers to the QMM free leg correction for ligand 11. G3 is the QM/MM bound leg correction for ligand 19 and G5 the QM/MM free leg correction for ligand 19. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

The QM/MM correction for the bound and free states are -11.45 kcal.mol^{-1} and -13.52 kcal.mol^{-1} respectively. This gives an overall QM/MM correction of -0.21 kcal.mol^{-1}.

This small correction value can be attributed to this perturbation consisting of the addition of a non-polar group.

The addition of a guanadinium group to the ligands also shows consistency in giving a large negative shift in free energy between MM and MM→QM/MM. For example, ligand 16 (Figure 7.17) the ammonium group is replaced by a guanadinium, which leads to a large negative shift in free energy between MM and MM→QM/MM.

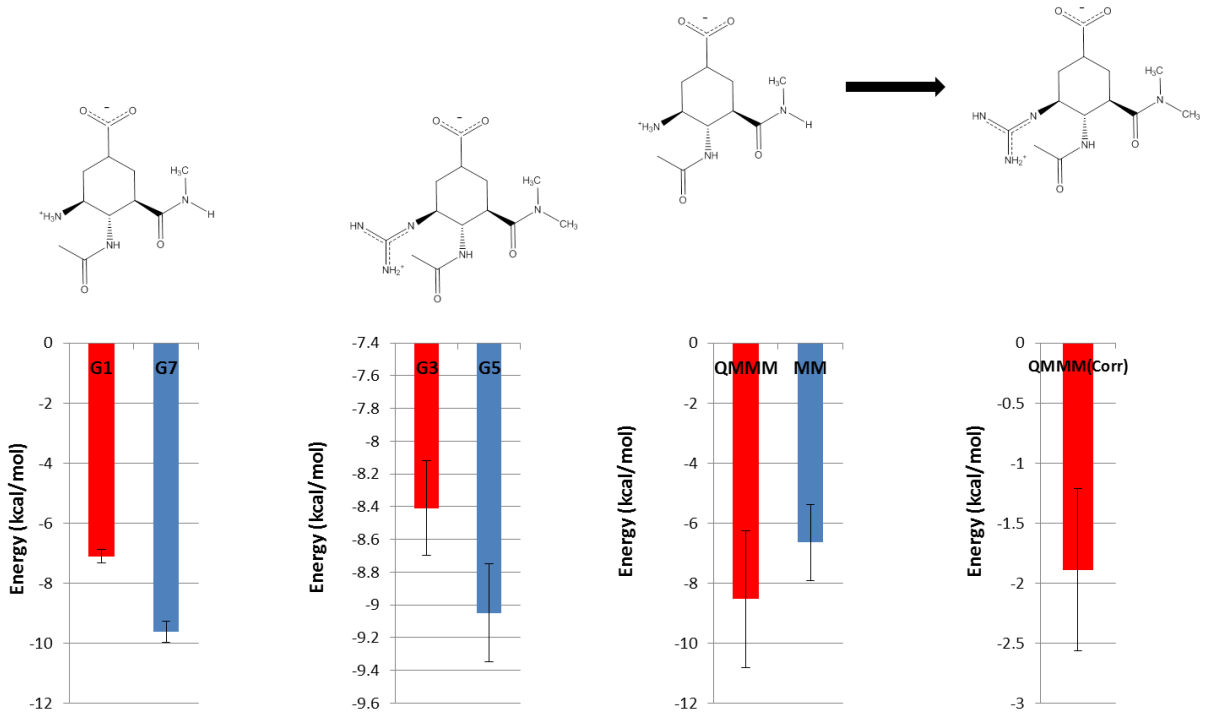


Figure 7.17. MM→QM/MM free energy breakdown for ligand 11 → ligand 16 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 11 and G7 refers to the QMM free leg correction for ligand 11. G3 is the QM/MM bound leg correction for ligand 16 and G5 the QM/MM free leg correction for ligand 16. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four indepnedent simulations using standard error.

The QM/MM corrections for the bound and free states are -8.40 kcal.mol^{-1} and -9.05 kcal.mol^{-1} respectively. This gives and overall QM/MM correction of -1.65 kcal.mol^{-1}. This shifts our MM-RETI free energy of -6.65 kcal.mol^{-1} to -8.30 kcal.mol^{-1} for MM→QM/MM. As is the case for all observations in this dataset the MM→QM/MM shifts the binding free energy further from the experimental value (-5.54 kcal.mol^{-1}) than MM-RETI predicted binding free energies.

Therefore, the QM/MM free energy study into 9 neuraminidase inhibitors has highlighted the difficulty in obtaining accurate free energies for extremely polar proteins with very polar ligands. Despite the dataset maintaining the PI between MM versus experiment and QM/MM versus experiment, there is a large rise in MUE which suggests that application of QM/MM corrections introduces a large amount of error into the predicted binding free energies. These errors arise in both the free and bound legs of our free energy calculations, but are particularly prevalent in the free legs. This is due to the extremely complex polarisation associated with neuraminidase and its inhibitors. The 9 inhibitors studied here are zwitterionic which leads to very large QM/MM corrections being obtained in this study. In fact, very few studies have attempted to describe zwitterions using QM/MM and none have attempted to perform free energy calculations utilising them [161]. Ideally, the Gaussian Blurring technique described in Section 5.2 would be used here to understand if this could potentially minimise the impact of embedding the MM environment in our QM/MM simulations. Unfortunately, this is outside of the scope of this current study, but future work is suggested to follow this route for such polar systems.

7.4 Protein-ligand Charge Perturbations

As this QM/MM approach neglects any sampling of the QM/MM state, the pathway-independence of the free energies obtained is subsequently checked. This is achieved through the use of charge perturbation pathways. For neuraminidase, three compounds were selected to perform charge perturbations (Figure 7.18).

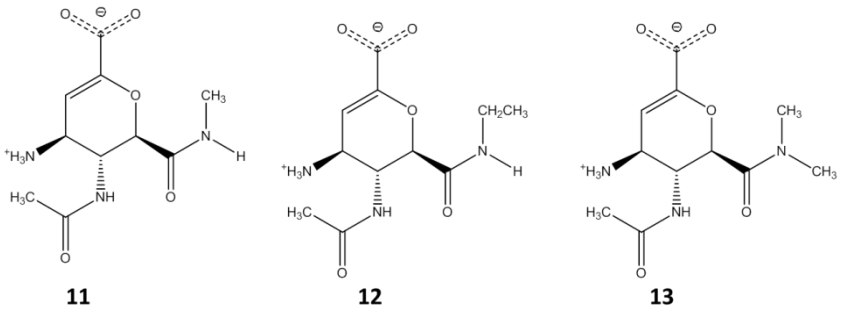


Figure 7.18. Three neuraminidase inhibitors chosen for charge perturbation simulations

Monte Carlo Simulation Protocol

Generating alternative pathways by scaling solute charges up or down can be used to validate the pathway independence of calculated free energies. Alternative configurations were generated by performing RETI calculations in which the solute charges were scaled up. For protein-ligand systems the charges of the solute-environment interactions were scaled, while the charges used for the solute internal energy computation remained at their un-scaled level (λ=0). 16 λ windows (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82, 0.88, 0.94, and 1.00) were used to ensure smooth transition between the two end states. The following charge scale factors were investigated: 1.001, 1.005, 1.007, 1.01, 1.015, 1.02 (N.B. 1.00 implies a simulation with non-perturbed charges). 10 million equilibration moves were performed before collecting statistics for 320 million moves in the bound legs and 160

million in the free legs, with each free energy simulation repeated 4 times. The RETI values shown are the mean of these four repeats and the standard error for these different runs is also shown.

QM/MM Single Point Energy Protocol

As with the 'normal' perturbations, configurations from the endpoint (λ=0 and λ=1) of the classical free energy simulations were selected and used as input for DFT-QM/MM single point energy calculations with Gaussian 09 [126]. One QM/MM single point energy calculation was performed every 100000th MC move. Therefore we took 3200 QM/MM configurations for the bound legs and 1600 QM/MM configurations for the free legs per repeat.

The QM energies were computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

Results & Discussion

The free energies obtained from the charge perturbations for neuraminidase/ligand 11 are shown in Figures 7.19 – 7.20.

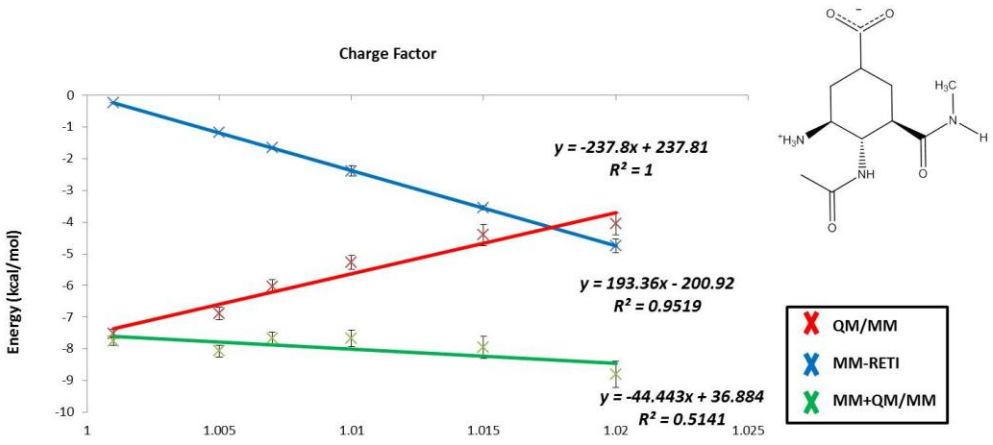


Figure 7.19. Charge perturbation results for ligand 11 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

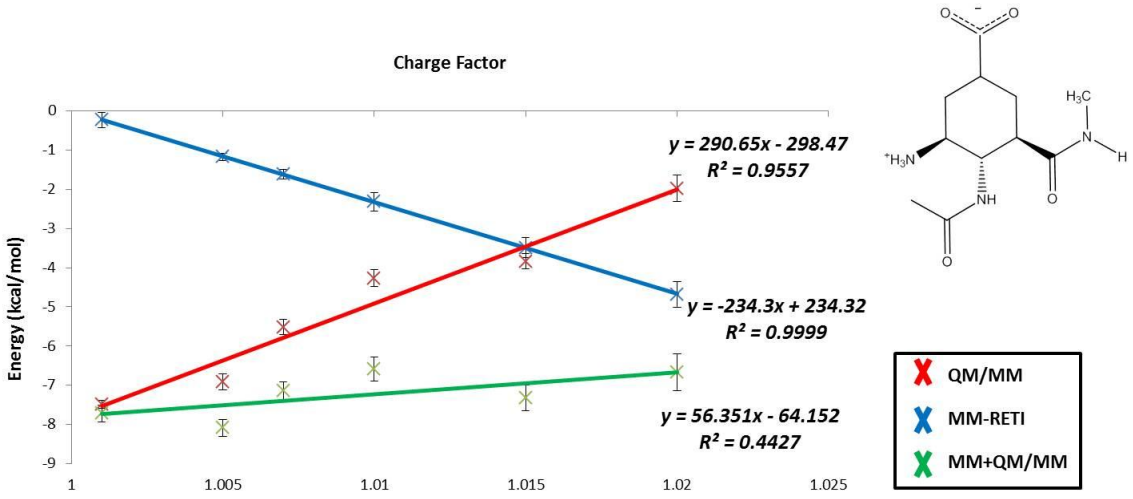


Figure 7.20. Charge perturbation results for ligand 11 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

In Figures 7.19 – 7.20. the sums of the charge perturbed free energy cycles are generally large (small slopes of the green fitted lines). However, if the free energies are

pathway independent, the mean values of these sums calculated over all scale factors must be equal to that of the non-charge perturbation MM→QM/MM free energies. The relevant free energies are summarised in Tables 3.4 and 3.5 in Supporting Information 3. The above condition is not fulfilled by ligand 11 in the free state. For ligand 11 in the free legs the average free energy of cycle closure is -7.99 (0.27) $kcal.mol^{-1}$. Comparing this value to the original MM→QM/MM free energy of -9.68 (0.33) $kcal.mol^{-1}$ it is clear that our cycle does not obtain the same value. For the bound state of ligand 11 the above condition is fulfilled. Ligand 11 obtains an average free energy cycle closure of -7.26 (0.29), which is very similar to the original MM→QM/MM free energy of -7.18 (0.32) $kcal.mol^{-1}$. The poor agreement for ligand 11 in the free state would suggest that additional sampling is needed in order to obtain more precise results for this ligand in the aqueous free energy leg.

The free energies obtained from the charge perturbations for neuraminidase/ligand 12 are shown in Figures 7.21 – 7.22.

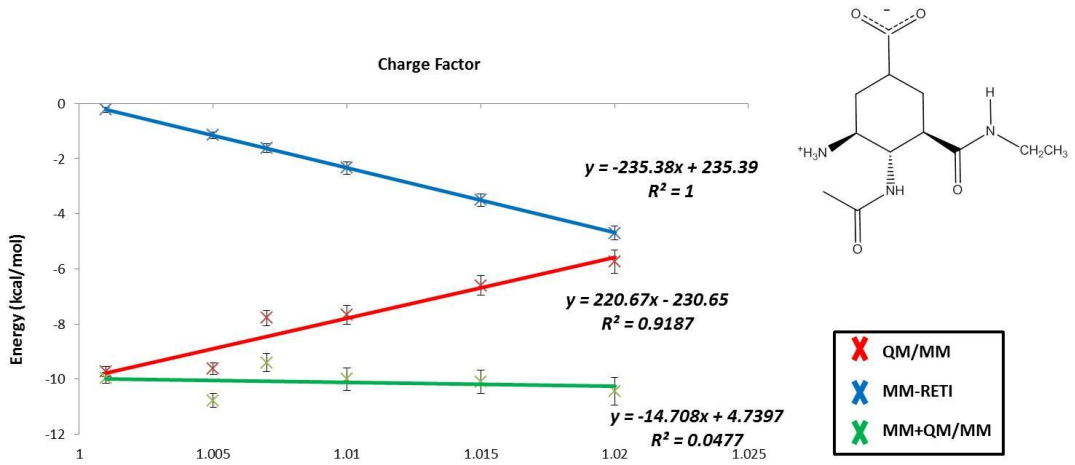


Figure 7.21. Charge perturbation results for ligand 12 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

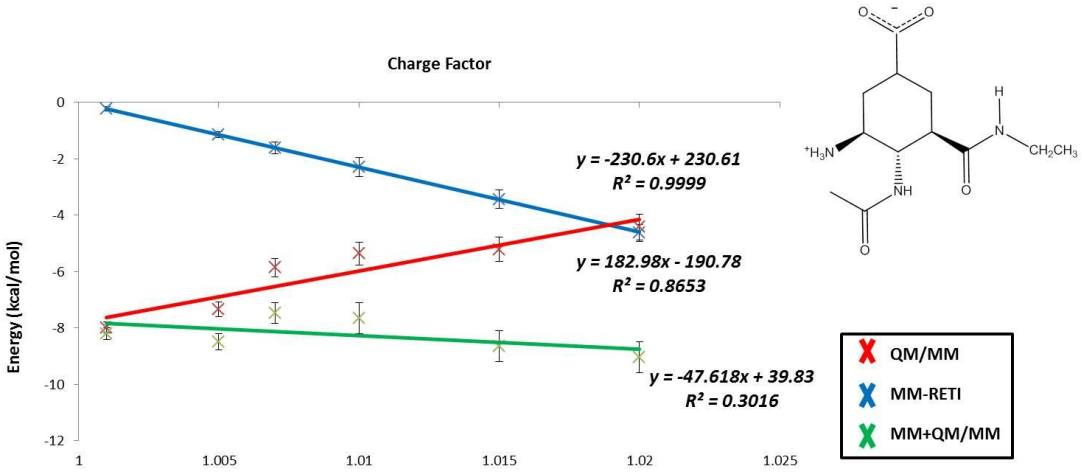


Figure 7.22. Charge perturbation results for ligand 12 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

In Figures 7.21 – 7.22. the sums of our charge perturbed free energy cycles are generally large (small slopes of the green fitted lines). The relevant free energies are

summarised in Tables 3.6 and 3.7 in Supporting Information 3. The above condition is fulfilled by ligand 12 in the free state. For ligand 12 in the free legs the average free energy of cycle closure is -10.11 (0.27) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -9.41 (0.29) kcal.mol^{-1} it is clear that our cycle does obtain the same value. For the bound state of ligand 12 the above condition is fulfilled. Ligand 12 obtains an average free energy cycle closure of -8.25 (0.26), which is very similar to the original MM→QM/MM free energy of -7.82 (0.36) kcal.mol^{-1}.

The free energies obtained from the charge perturbations for neuraminidase/ligand 13 are shown in Figures 7.23 – 7.24.

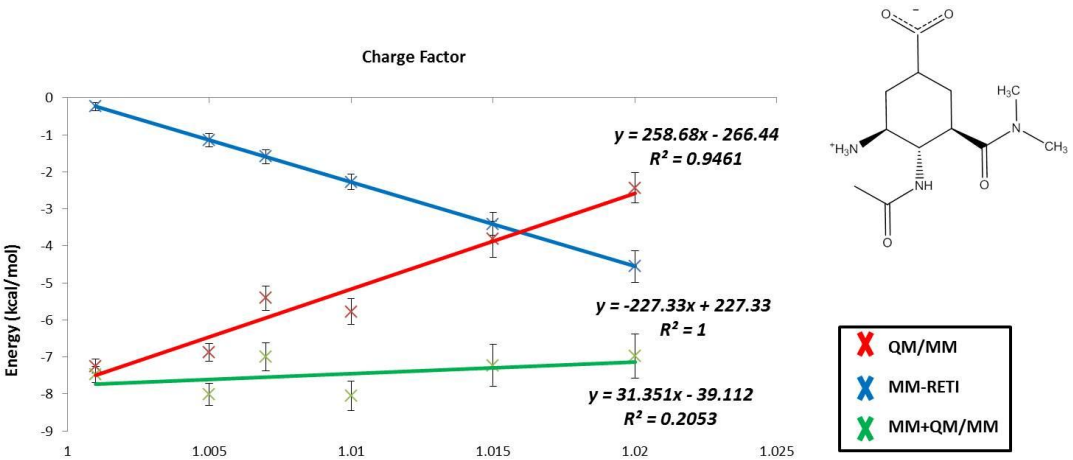


Figure 7.23. Charge perturbation results for ligand 13 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

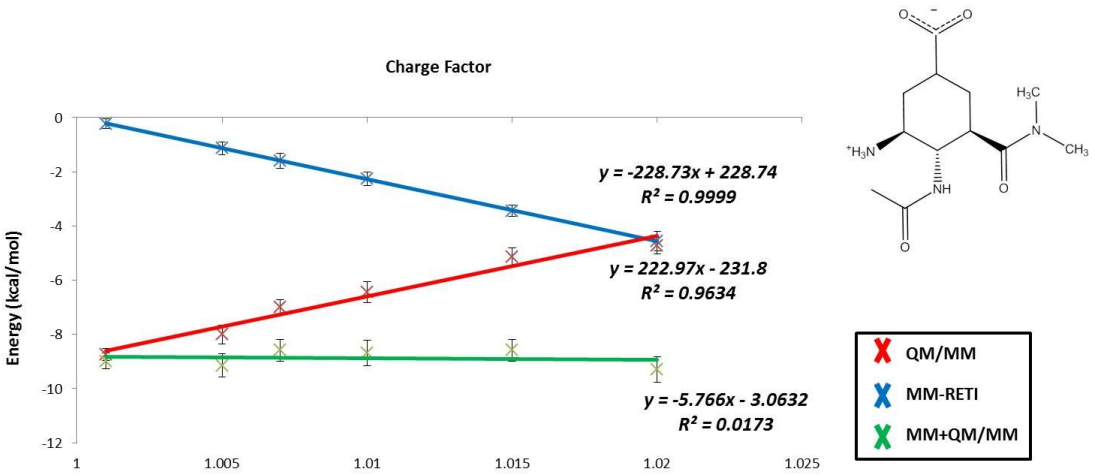


Figure 7.24. Charge perturbation results for ligand 13 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were computed from four independent simulations using standard error.

In Figures 7.23 – 7.24 the sums of our charge perturbed free energy cycles are generally large (small slopes of the green fitted lines). The relevant free energies are summarised in Tables 3.8 and 3.9 in Supporting Information 3. The above condition is fulfilled by ligand 13 in the free state. For ligand 13 in the free legs the average free energy of cycle closure is -9.25 (0.28) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -9.42 (0.31) kcal.mol^{-1} it is clear that our cycle does obtain a similar value. For the bound state of ligand 13 the above condition is not fulfilled. Ligand 13 obtains an average free energy cycle closure of -8.89 (0.25), which shows over 1 kcal.mol^{-1} difference to the original MM→QM/MM free energy of -7.72 (0.26) kcal.mol^{-1}. The poor agreement for ligand 13 in the bound state would suggest that additional sampling is needed in order to obtain more accurate results for this ligand in the bound free energy leg.

7.5 Conclusions

The aim of this study was to assess the applicability of the QM/MM method when applied to a set of neuraminidase inhibitors. The results from this study showed that although the application of the QM/MM method maintained the PI between MM→QM/MM calculated free energies and experiment, in general, the MM→QM/MM calculated free energies showed a higher MUE when compared to the classically obtained MM binding free energies. These errors could be a product of several problems; first, the forcefield used in this study could be describing these systems poorly. Several other neuraminidase studies e.g. by Woods *et al.* [162] and Bonnet and Bryce [163] appear to support this assumption as such polar systems can lead to large errors between predicted and experimental binding free energies. Second, the simulations may not have sampled enough, although due to the already extensive sampling performed here performing more would lead to a greater computational cost, a point which this methodology tries to avoid. Lastly, the single-step QM/MM approach employed may not be sufficient to fully describe the protein-ligand interactions within this system. This is supported by the charge perturbation results which show poor free energy cycle closure for several protein-ligand examples. Ideally, the Gaussian blurring technique describe in section 5.2 would have been applied to these highly polar neuraminidase systems, but unfortunately this is outside the scope of this current study. Future work will be to apply the Gaussian blurring technique to such protein-ligand examples to understand if this can improve the prediction of MM→QM/MM binding free energies for extremely polar protein-ligand complexes. It is also noticable that the MM binding free energies also show large errors due to the highly polarised nature of neuraminidase. This indicates that more advanced

forcefields, i.e. polarisable forcefields which can adjust partial charge parameters for the solute/protein/solvent atoms based upon the interactions they share with each other, may be needed to aid in the description of such polar targets.

8 Calculation of QM/MM Binding Free Energies for 18 Cyclin Dependent Kinase 2 Inhibitors

8.1 Biological Relevance

There are over 500 different kinases within the human body [164], with each kinase responsible for different cellular processes. Cyclin Dependent Kinases (CDKs) such as CDK2 are protein kinases involved in critical cellular processes such as cell proliferation, whose activity relies on associated cyclin subunits. When CDK2 is bound to cyclin E an activated form of the kinase is produced which promotes cell proliferation. Human cancers typically have an over expression of cyclin E, leading to an overabundance of CDK2 and hence promoting tumour development [165]. Therefore, by targeting CDK2 via drug therapy the enzyme can be deactivated, leading to suppression of tumour growth [165].

A noticeable problem with kinase based drug therapy lies in the fact that 60% of kinases share over 85% sequence similar [166]. For example, there are 13 recognised CDKs (CDK1-13) with highly similar structural and sequence properties. As a result careful design of inhibitors is required for each individual target; otherwise undesirable side effects could occur, whereby other structurally related kinases can also be affected by the inhibitor. This problem is ideally suited to computational approaches, since simulating inhibitor effects *in silico* is advantageous over performing expensive *in vivo* experiments. Ligand binding studies have been performed upon CDK2 previously, but the similarity in activities between the inhibitors meant that reliable predictions were not achievable [167]. It has been noted that ligand-kinase complexes typically

contain an unusual interaction; an aromatic CH···→O hydrogen bond. This interaction is thought to be weak, yet it is found in most structures [168].

Understanding such interactions could help to accurately predict the binding free energy of inhibitors to kinases. To examine this problem, QM/MM simulations were performed to predict the binding free energies for a range of CDK2 inhibitors (Figure 8.1)

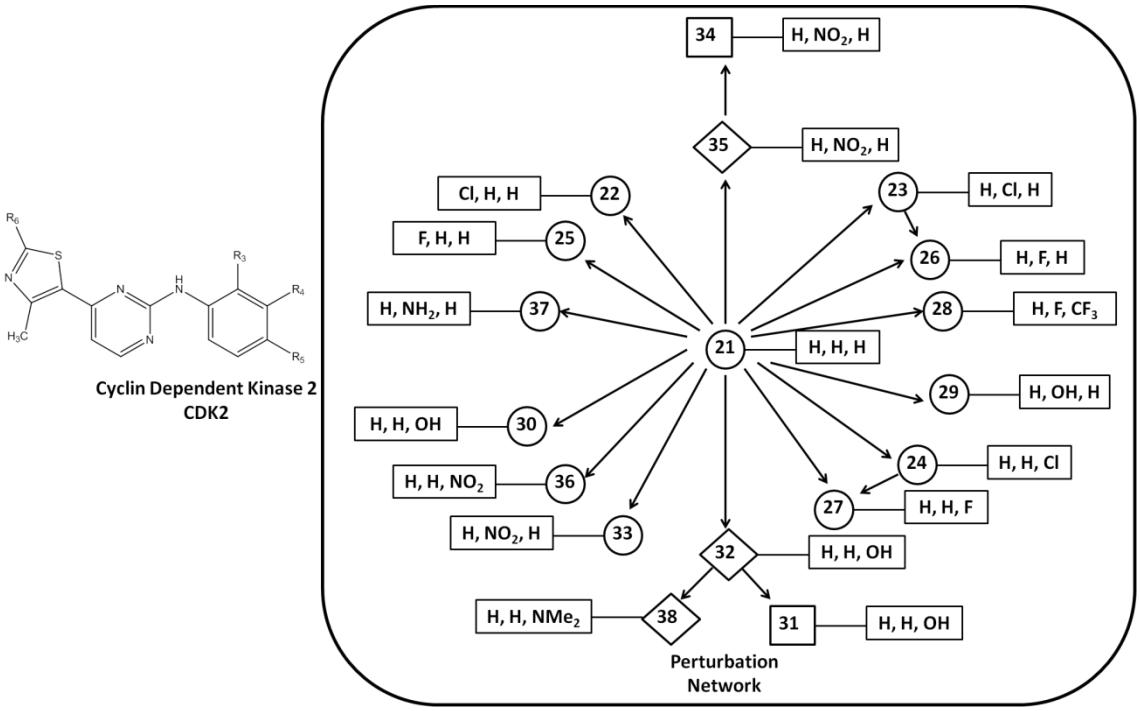


Figure 8.1. Perturbation network for the set of CDK2 perturbations studied here. The circled numbers represent ligands with a methyl at position R_6, the squared numbers represent perturbations with a mono-methylated amino at position R_6 and the diamond numbers represent perturbations with an amino group at position R_6. The R,R,R represent substituents at positions R_3, R_4 and R_5.

The binding free energies for these inhibitors was previously calculated by Michel *et al.* [146], with very poor agreement between for both MM(AMBER99/GAFF/AM1-BCC)-

RETI (GB solvent model) and MM(AMBER99/GAFF/AM1-BCC)-RETI (explicit TIP4P solvent model) calculated binding free energies and experimental results. Therefore this study is aimed towards investigating how our QM/MM method performs on a dataset which MM finds extremely challenging.

8.2 System Preparation

Protein – ligand setup

The PDB structure of human CDK2 extracted from a CDK2/cyclin A complex (PDB code 2C5P) [169] was selected as a starting point for this study. Hydrogen atoms were added to this structure using the Reduce software package [148]. The protonation states of histidines were determined via visual inspection. The protein was parameterised using the AMBER99 force field [13], inhibitors were parameterised with the GAFF force field [14] and the partial atomic charges were derived using the AM1-BCC method [124], as implemented in the AMBER 10 suite. To avoid bad steric clashes, the protein-ligand complex (2C5P/ligand 32) was minimised in the SANDER module of AMBER 10 with a generalised Born solvent model. The backbone of the protein was subsequently fixed for Monte Carlo simulations, which were performed using a modified version of ProtoMS2.2 [123]. To reduce computational cost, only protein residues that contained one heavy atom within 15 Å of any representative ligand atom were retained. The resulting protein scoop contained 115 residues. The ligands were modelled in the binding site based upon the binding mode predicted by the docking program GOLD [116], the binding modes for each ligand were generated by Michel *et al.* [146] Crystallographic waters were retained and the complex was hydrated by a sphere of TIP4P [125] water molecules of 22 Å radius and centred on the geometric

centre of the ligand. To prevent evaporation, a half-harmonic potential with a 1.5 kcal.Å^{-2} force constant was applied to water molecules whose oxygen atom distance to each ligands centre of geometry was greater than 22 Å. A similar sphere of water was used for the unbound state.

Monte Carlo Simulation Protocol

The bond angles and torsions for the side chains of residues within 10 Å of any ligand heavy atom and all bond angles and torsions of the ligand were sampled during the simulation, with ring structures being the only exception. The bond lengths of the residues and ligand were constrained. The total charge of the system was brought to zero by neutralising lysine residues 6, 34, and 56 lying in the outer 'frozen' part of the scoop. The neutralised lysines were then re-modelled using the AMBER99 forcefield. A 10 Å residue based cut-off was employed in all simulations.

For simulations in the bound state, solvent moves were attempted with a probability of 73.4%, protein side-chain movements with a probability of 21.11% and solute moves with a probability of 5.49%. In the unbound state, solvent moves were attempted 99.06% of the time. Replica exchange moves were attempted every 200000 moves. The solvent was equilibrated for 20 million moves to remove any bad contacts with the solute. The system was then equilibrated at one state (the end state with the larger solute) for 20 million further moves where solute, protein, and solvent moves were attempted. The resulting configuration was distributed over the 16 values for the coupling parameter λ (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68,

0.75, 0.82, 0.88, 0.94, and 1.00) and equilibrated for a further 10 million moves before collecting statistics for 640 million moves (bound) and 320 million moves (free).

QM Single Point Energy Protocol

Configurations from the endpoint (λ=0 and λ=1) of the classical free energy simulations were selected and used as input for DFT-QM/MM single point energy calculations with Gaussian 09 [126]. One QM single point energy calculation with background charges representing the solvent and protein residues within our cut-off (Gaussian keyword 'CHARGE') were performed every 100000th MM/MC moves, with symmetry operations disabled (Gaussian keyword 'NoSymm'). This gave a total of 6400 QM/MM single points for each solute perturbation in the bound state and 3200 QM/MM single points for each solute perturbation in the free state. Gaussian calculations with embedded background charges allow a polarisation of the QM wave function via the MM charges, however no back polarisation of the MM part via the polarised QM wave function was considered.

The QM energies were computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

As the ligands were flexible we needed to compute the QM vacuum energies for each snapshot used. This was performed in Gaussian 09, but without the use of the 'CHARGE' and 'NoSymm' keywords, which are only necessary if embedding MM point charges in our calculation.

The QM vacuum energies were again computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

8.3 Results & Discussion

8.3.1 MM - RETI Results

The calculated relative binding free energies of 17 CDK2 inhibitors are shown in Figure 8.2. The coefficient of determination (R^2) between predicted and experimental binding free energies [146] was very poor at just 0.07. The MUE is equal to 4.55 kcal.mol^{-1}, which is extremely disappointing. The PI also showed a lack of correlation to experimentally observed trends with a value of -0.35, and a τI value of -0.05, indicating that our results are at worse than random, and in most cases wrong. The relevant free energies are summarised in Tables 4.1 and 4.3 of Supporting Information 4.

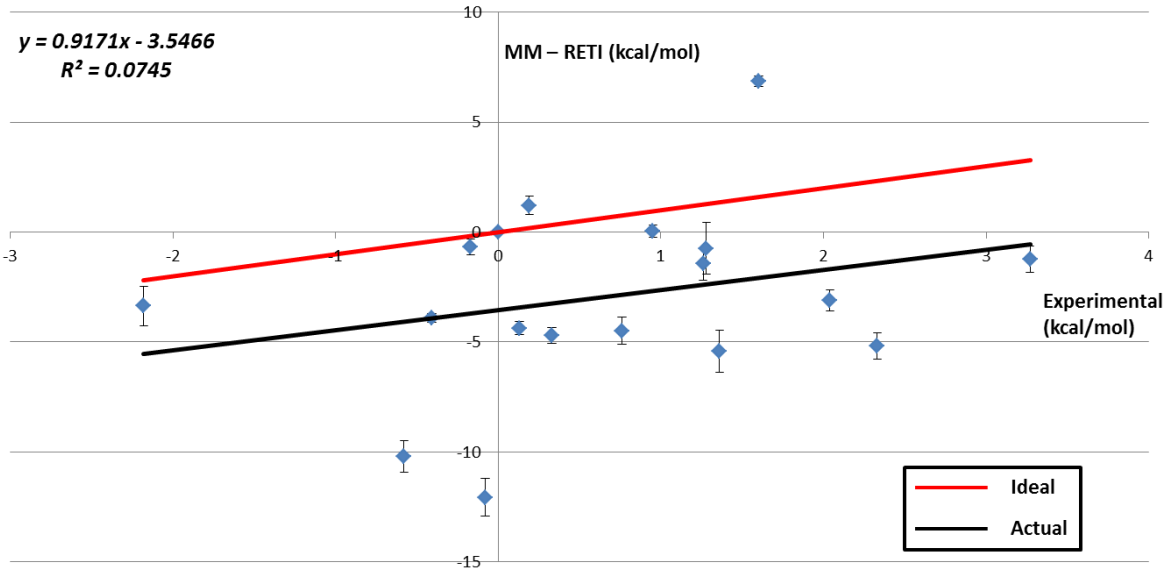


Figure 8.2. MM-RETI results versus experimental data [146]. The red line represents the ideal (1 to 1) correlation and the black line represents the actual correlation. The error bars shown were computed from four independent simulations using standard error.

Several factors make this series of CDK2 inhibitors difficult to predict. First, there are twice as many compounds studied as in previous protein-ligand studies, making this study more challenging due to the additional complexity of the data. Second, the span of experimental binding free energies is smaller (≈5 kcal.mol^{-1}) than for COX-2 (≈6 kcal.mol^{-1}) and neuraminidase (≈7 kcal.mol^{-1}). Lastly, half of the compounds within this dataset have a binding free energy within 1 kcal.mol^{-1} of our reference molecule 21.

To ensure smooth transition between the two end states (λ=0 and λ=1) the free energy gradients for each perturbation were studied (Figure 8.3).

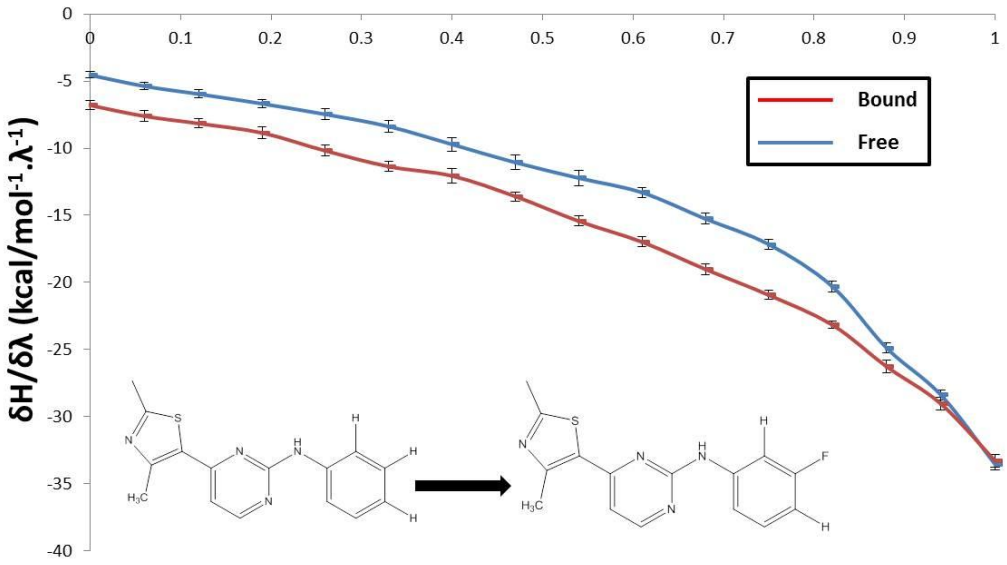


Figure 8.3. Free energy gradients for both free (blue line) and bound (red line) for the ligand 21 to ligand 26 perturbation. The error bars shown were computed from four independent simulations using standard error.

These show that for both the free and bound legs of the free energy simulations the transition across the reaction co-ordinate is very smooth. This indicates that the free energies obtained from these simulations are precise.

To analyse the statistical uncertainty in the free energy simulations the hysteresis for closing two binding free energy cycles for a set of 3 CDK2 perturbations (21→23, 23→26 and 21→26) was calculated. For the free legs of the simulations (Figure 8.4) the hysteresis was found to be extremely small at just -0.38 (0.51) kcal.mol^{-1} suggesting little statistical uncertainty.

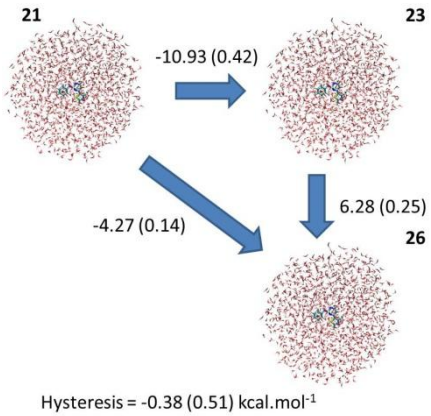


Figure 8.4. MM-RETI free leg hysteresis for three CDK2 perturbations. The standard errors for each step are shown in the brackets.

For the bound legs (Figure 8.5) the hysteresis is 0.04 (0.48), which implies that the bound legs exhibit little statistical uncertainty, much like our free legs.

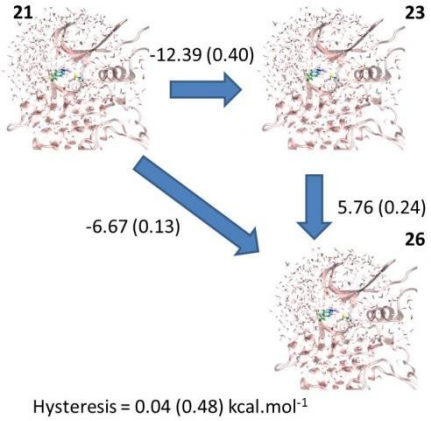


Figure 8.5. MM-RETI bound leg hysteresis for three CDK2 perturbations. The standard errors for each step are shown in the brackets.

The hysteresis for closing both binding cycles is small at just 0.43 (0.81) kcal.mol^{-1} as illustrated in Figure 8.6.

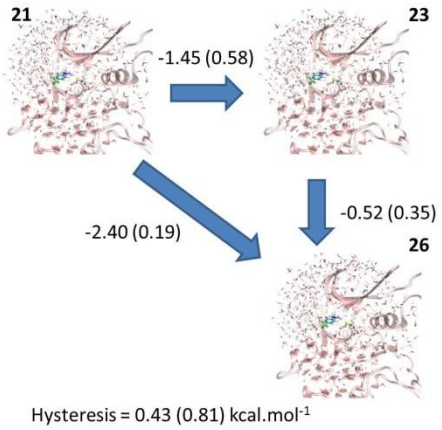


Figure 8.6. MM-RETI total hysteresis for three CDK2 perturbations. The standard errors for each step are shown in the brackets.

This is a notably small value, indicating that the statistical uncertainty in the free energy simualtions is low.

8.3.2 MM→QM/MM–FEP Results

The calculated relative MM→QM/MM-FEP binding free energies of 17 CDK2 inhibitors is shown in Figure 8.7. The coefficient of determination (R^2) between predicted and experimental binding free energies [146] is 0.06. The MUE is equal to 5.38 kcal.mol^{-1}, which is just under 1 kcal.mol^{-1} worse than for the MM-RETI calculated binding free energies. The relevant free energies are summarised in Table 4.2 and 4.3 of Supporting Information 4.

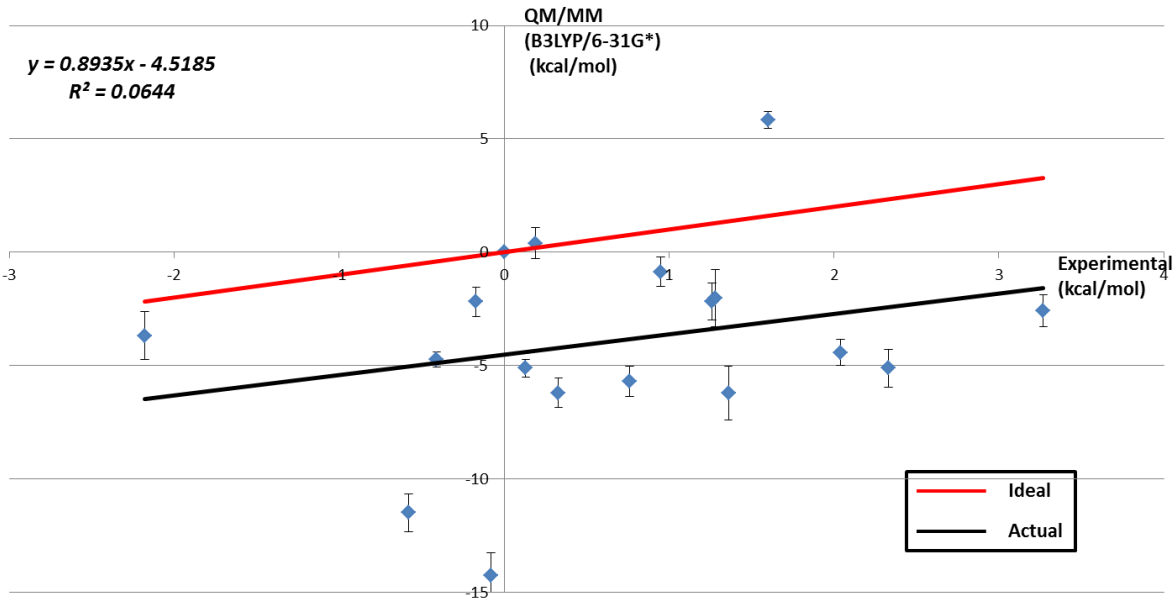


Figure 8.7. MM→QM/MM results versus experimental data [146]. The red line represents the ideal (1 to 1) correlation and the black line represents the actual correlation. The error bars shown were computed from four independent simulations using standard error.

In addition to the poor correlation and large MUE the PI for the MM→QM/MM binding free energies is -0.45, and the τI becomes -0.18. This indicates that by applying the QM/MM corrections to the classically obtained MM binding free energies causes the predictive nature to remain poor, much like the already poor MM-RETI free energies.

As with the MM-RETI results the hysteresis for closing a binding free energy cycle for a set of 3 CDK2 perturbations (21→23, 23→26 and 21→26) was calculated. For the free legs of the MM→QM/MM binding free energy study (Figure 8.8) the hysteresis is 0.65 (0.61) kcal.mol^{-1} this is a small increase from the MM values of -0.38 (0.51). This indicates that applying the QM/MM corrections a slightly higher statistical undertainty and larger errors are produced compared to standard MM.

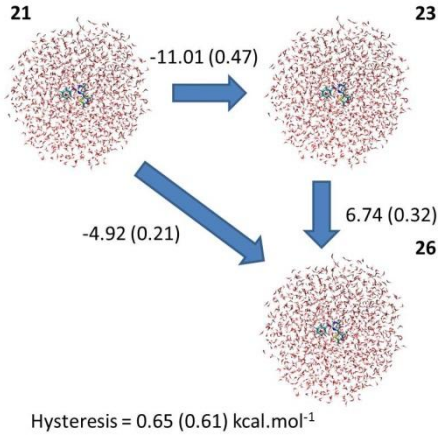


Figure 8.8. MM→QM/MM free leg hysteresis for three CDK2 perturbations. The standard errors for each step are shown in the brackets.

The trend of a slightly increased uncertainty continues when considering the bound legs of our MM→QM/MM binding free energy study (Figure 8.9). The hysteresis in the bound legs is 0.06 (0.64) kcal.mol^{-1} which is a small increase in the hysteresis shown in the MM free energy study of 0.04 (0.48) kcal.mol^{-1}.

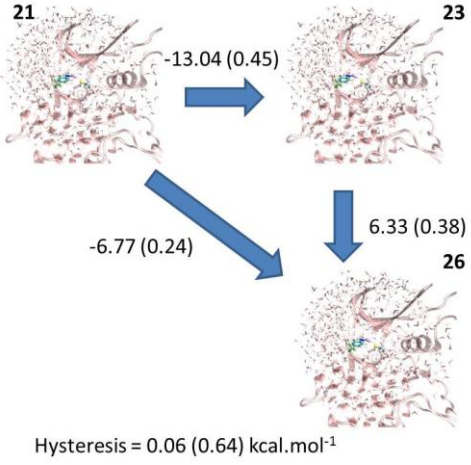


Figure 8.9. MM→QM/MM bound leg hysteresis for three CDK2 perturbations. The standard errors for each step are shown in the brackets.

Combining the free and bound legs leads to the overall hysteresis for this free energy cycle (Figure 8.10). The overall MM→QM/MM hysteresis is -0.59 (1.13) kcal.mol^{-1} which is an insignificant increase compared to the hysteresis for MM of 0.43 (0.81) kcal.mol^{-1} as this is an absolute value so the change of sign is insignificant.

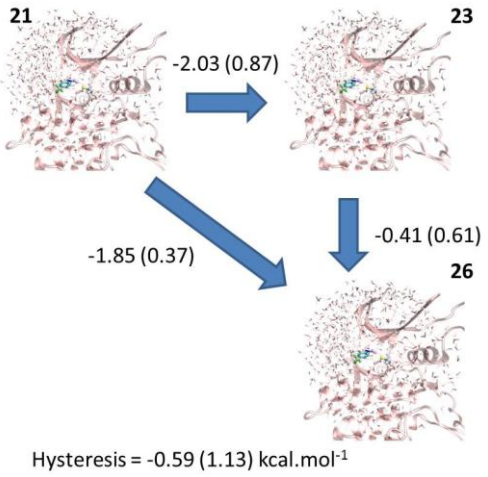


Figure 8.10. MM→QM/MM total leg hysteresis for three CDK2 perturbations. The standard errors for each step are shown in the brackets.

This small change in hysteresis between MM and MM→QM/MM suggests that by performing QM/MM corrections very little statistical uncertainty is produced when compared to a purely MM approach.

To understand the changes in accuracy between our MM and MM→QM/MM-FEP binding free energies we need to analyse the energies produced for each leg of the protein-ligand binding free energy cycle. The perturbations are directed into three areas (R3, R4, and R5).

The perturbations at position R3 are all of the type H→Cl/F, all of these perturbations show a similar trend of obtaining more favourable QM/MM bound state

corrections and less favourable QM/MM free state corrections. For example, ligand 25 (Figure 8.11) the bound QM/MM correction is -0.62 kcal.mol^{-1} which is slightly more favourable than the reference compounds -0.35 kcal.mol^{-1}. Whereas, the free QM/MM correction is less favourable at 0.32 kcal.mol^{-1} compared to the reference compounds -0.44 kcal.mol^{-1}.

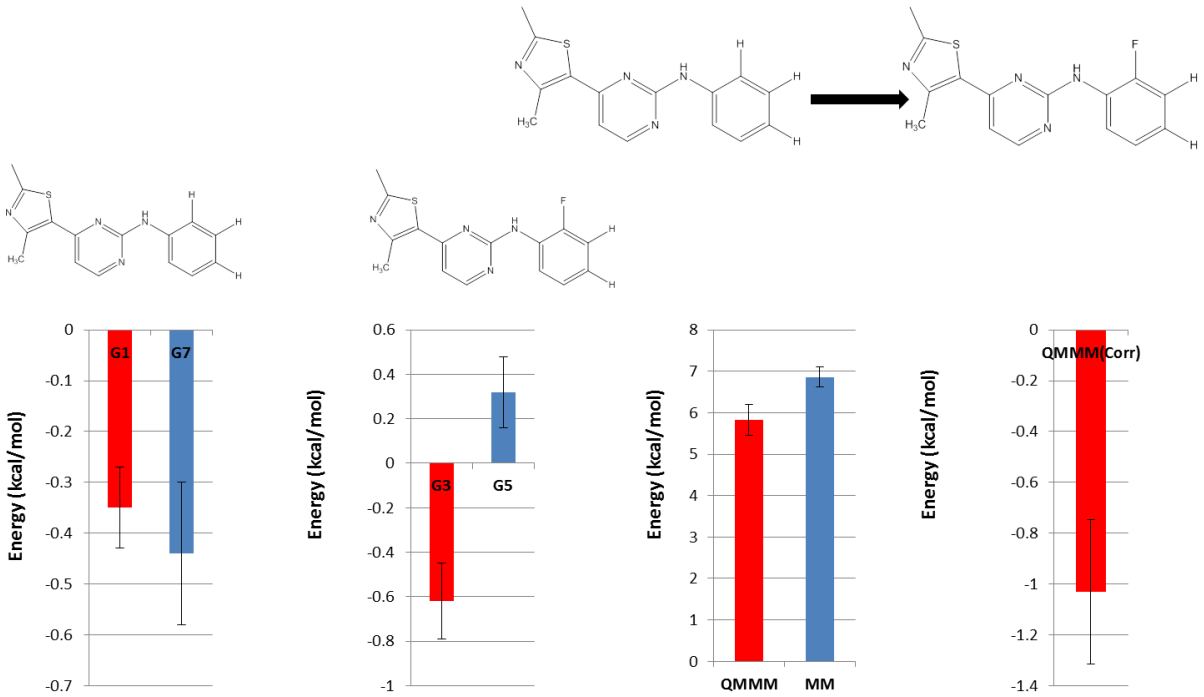


Figure 8.11. MM→QM/MM free energy breakdown for ligand 21 → ligand 25 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 21 and G7 refers to the QMM free leg correction for ligand 21. G3 is the QM/MM bound leg correction for ligand 25 and G5 the QM/MM free leg correction for ligand 25. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

This combination of QM/MM corrections leads to a largely negative overall MM→QM/MM binding free energy for ligand 25 compared to MM. This does not

agree with experimental data (0.24 kcal.mol^{-1}), and it appears that the QM/MM correction obtained is not sufficient to correct for the extremely poor MM-RETI predicted binding free energy, a trend which is seen throughout this CDK2 dataset for perturbations in the R3 position.

In the R4 position there is a greater variance in perturbations attempted as also the perturbation of the methyl group into amino or mono-methylated amino substituents is performed. These additional changes lead to quite substantial differences in QM/MM corrections obtained. For example, ligand 29 (Figure 8.12) shows highly favourable QM/MM correction for both the bound and free states compared to the reference compound.

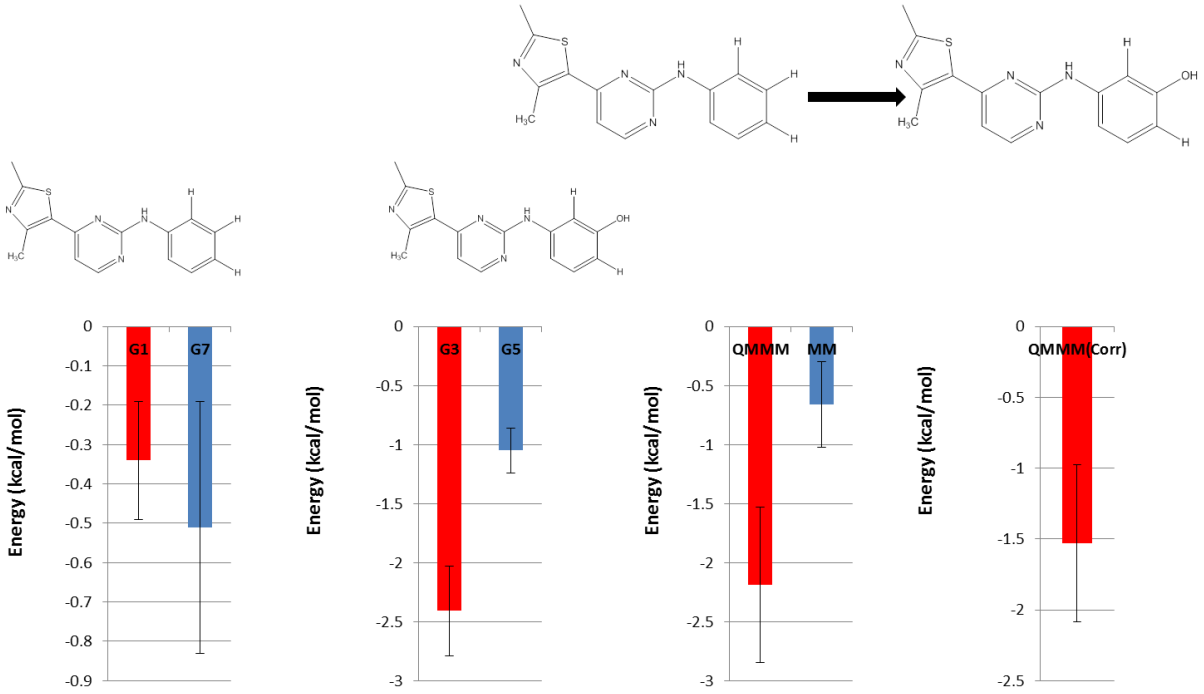


Figure 8.12. MM→QM/MM free energy breakdown for ligand 21 → ligand 29 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 21 and G7 refers to the QMM free leg correction for ligand 21. G3 is the QM/MM bound leg correction for ligand 29 and G5 the QM/MM free leg correction for ligand 29. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

The QM/MM corrections for ligand 29 are -2.41 kcal.mol^{-1} (bound) and -1.05 kcal.mol^{-1} (free). These differences are most likely caused by the polar nature of the OH, leading to a greater interaction between the ligand and polar residues on the edge of the CDK2 binding site, and also water molecules in the free state. This leads to an MM→QM/MM binding free energy which has a large negative shift compared to MM.

Comparing the results of ligand 29 to ligand 31 (Figure 8.13) the results of performing the same perturbation (H → OH) with the addition of perturbing the

methyl group to a mono-methylated amino gives a more favourable QM/MM in both free and bound legs when compared to the reference compound.

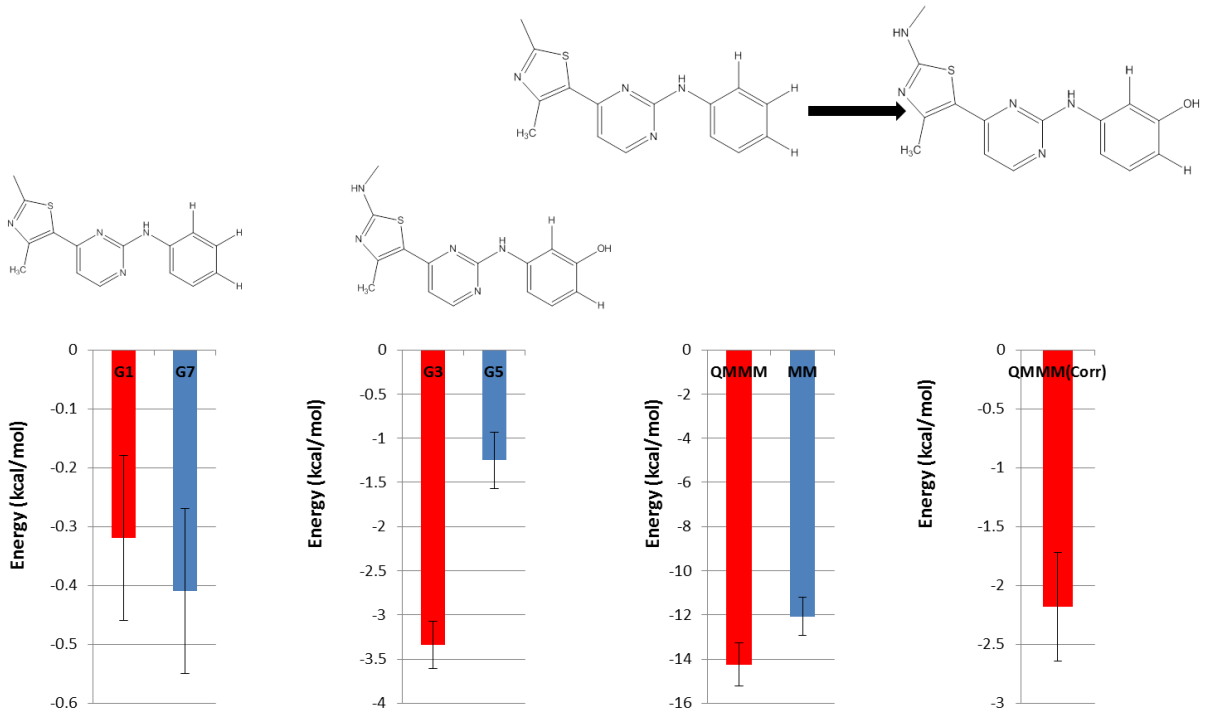


Figure 8.13. MM→QM/MM free energy breakdown for ligand 21 → ligand 31 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 21 and G7 refers to the QMM free leg correction for ligand 21. G3 is the QM/MM bound leg correction for ligand 31 and G5 the QM/MM free leg correction for ligand 31. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

The additional perturbation in ligand 31 leads to bound and free QM/MM corrections of -3.34 kcal.mol^{-1} and -1.25 kcal.mol^{-1} respectively. Again, this leads to large negative shift in the overall MM→QM/MM binding free energy. As with ligand 29 the MM prediction is much more negative than the experimental value. Hence, the MM→QM/MM binding free energy is shifted further from the experimental result.

This suggests that for perturbations in position R4 that the MM free energies are highly inaccurate and any QM/MM corrected free energies appear to make the free energies obtained worse.

As with perturbations in the R4 position, perturbations in the R5 position also have a greater possibility for change. For example, ligand 30 (Figure 8.14) we calculate the QM/MM corrections for the bound and free states are more favourable than those of the reference compound.

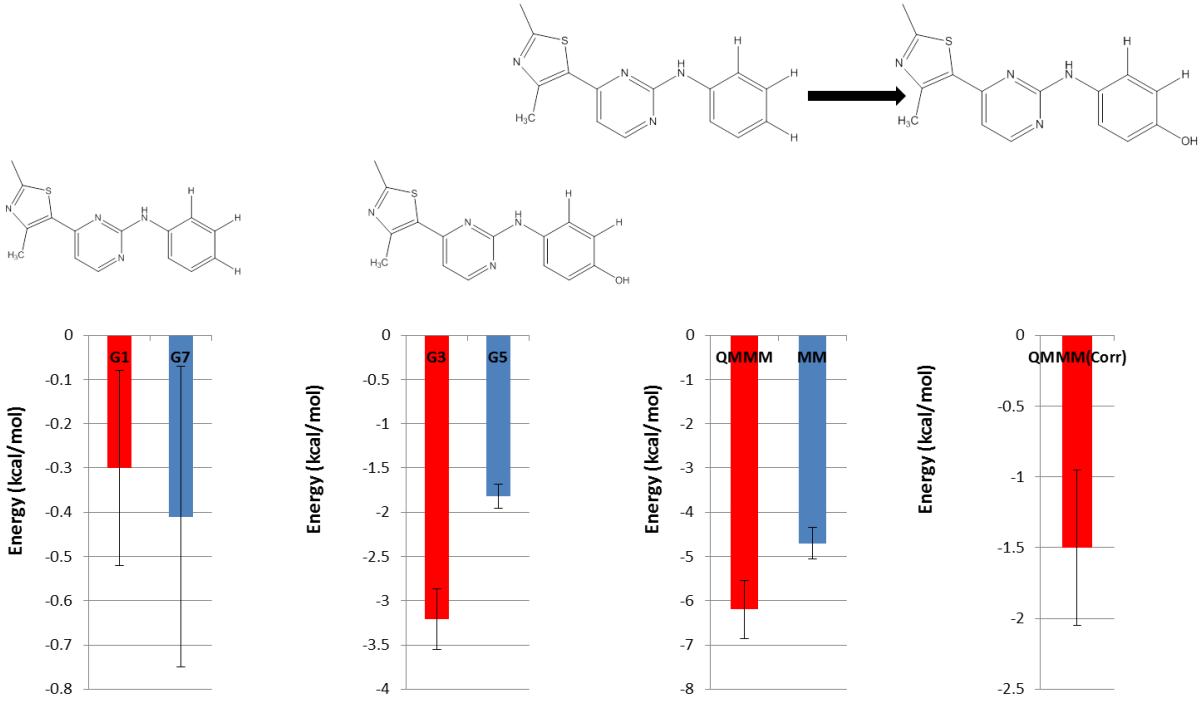


Figure 8.14. MM→QM/MM free energy breakdown for ligand 21 → ligand 30 perturbation. The error bars shown were computed from four independent simulations. G1 refers to the QM/MM bound leg correction for ligand 21 and G7 refers to the QMM free leg correction for ligand 21. G3 is the QM/MM bound leg correction for ligand 30 and G5 the QM/MM free leg correction for ligand 30. The QM/MM and MM free energies are also shown, along with the overall QM/MM correction for this perturbation. The error bars shown are computed from four independent simulations using standard error.

The QM/MM corrections for bound and free states are -3.21 kcal.mol^{-1} and -1.82 kcal.mol^{-1} respectively. These values are very similar to the same perturbation in the R4 position (ligand 29). Similar to ligand 29, our overall QM/MM correction is negative, but as our MM-RETI value (-4.7 kcal.mol^{-1}) was already very inaccurate and too negative compared to experimental data (0.35) kcal.mol^{-1} this leads to an MM→QM/MM binding free energy which is even less accurate than MM.

For the CDK2 dataset it is possible for the QM/MM method used to show some improvement on the original MM-RETI calculated binding free energies, particularly for the perturbations in the R3 position involving H→Cl/F. However, owing to the large inaccuracies between the calculated MM-RETI and experimental data, these corrections are too small to have any statistical significance when comparing the MM→QM/MM results to experimental data. For more polar perturbations, particularly in the R4 and R5 positions, it is clear that the QM/MM corrections are leading to less accurate results when compared to our already very inaccurate MM-RETI calculated values. From investigation of literature sources we identified a study from Heady *et al.* who had identified a potentially key water molecule within the CDK2 binding site [170]. The identification of this water could have large implications on this study, as within this work no waters were placed in this region of the binding site. Therefore, to understand if this water could be bound to the CDK2 inhibitor GCMC simulations were performed on the CDK2 binding site.

8.4 Grand Canonical Monte Carlo – CDK2 binding site

The GCMC simulation was performed on a holo structure of CDK2 where each ligand was bound to the protein during the GCMC simulation. Insertion and deletion attempts were accepted using the Metropolius tests described in section 6.3.2.

GCMC Simualtion protocol

No formal hardwall region is applied in the GCMC simulations. Although other (bulk) water molecules are prohibited from entering the defined GCMC region, protein atoms are allowed to occupy the same region as the GCMC simulation. As a result a 14 x 10 x 5 $Å^3$ grid was defined around the binding site to obtain suffcient sampling of the binding site region. Each B value (see section 6.3.2 for definition of Adams parameter B) was simulated for 40 million MC moves. At the end of each simulation the average population across the entire simlulation was recorded.

For each the CDK2 protein-ligand system, 7 B values (4, 0, -4, -8, -12, -16, and -20) were simulated to allow for a reliable estimate of the binding free energy. The free energy of hydration of water, ΔG_{hyd} was taken to be +6.4 kcal.mol^{-1} [154].

Results & Discussion

The results from the GCMC study of CDK2 are shown below in Figure 8.16 – 8.18, with titration curves and images of hydration patterns at biasing potentials of 4 and -20 displayed.

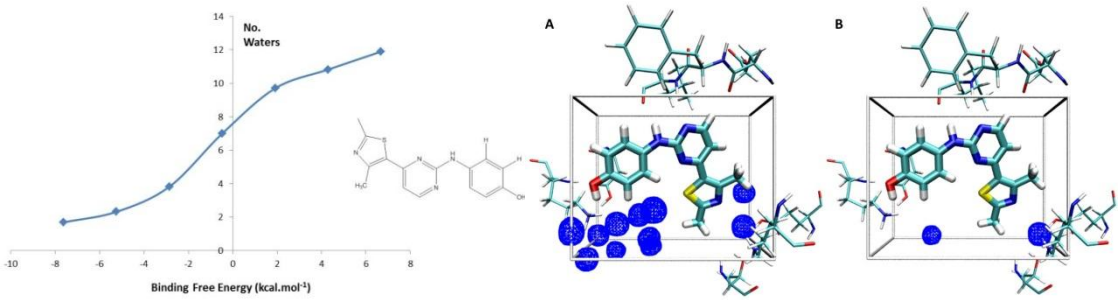


Figure 8.16. Titration curve and images of hydration patterns at biasing potentials of 4 (A) and -20 (B) for ligand 30 in complex with CDK2.

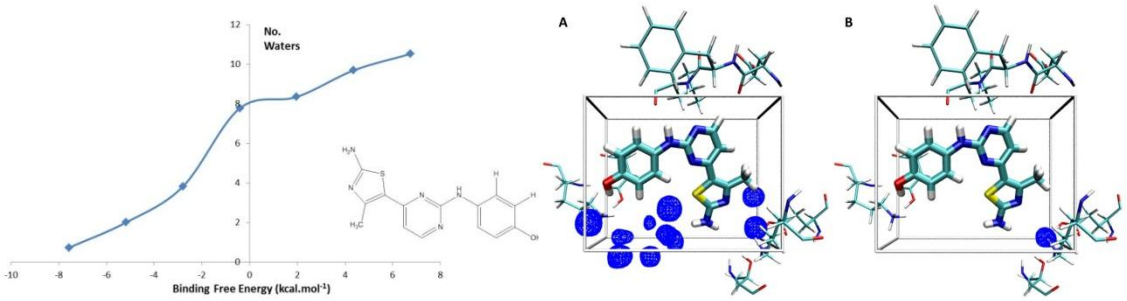


Figure 8.17. Titration curve and images of hydration patterns at biasing potentials of 4 (A) and -20 (B) for ligand 32 in complex with CDK2.

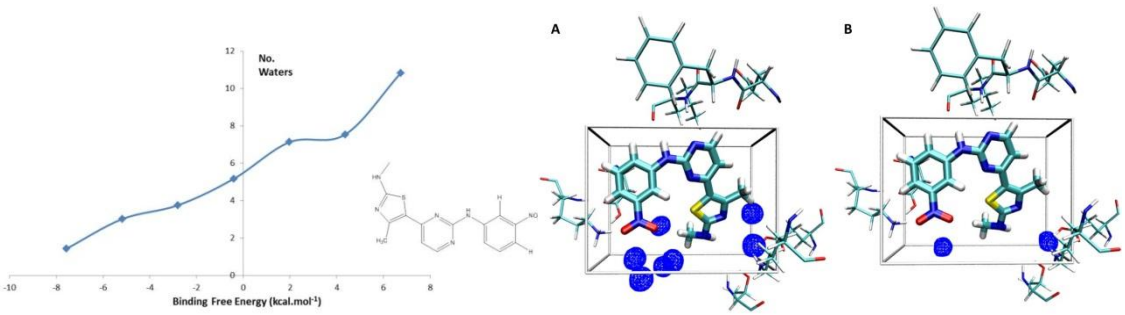


Figure 8.18. Titration curve and images of hydration patterns at biasing potentials of 4 (A) and -20 (B) for ligand 34 in complex with CDK2.

In Figures 8.16 – 8.18 it is clear that at a high B-value (4) there are 10-12 water molecules present within the CDK2 binding site. Most of these appear in the solvent exposed region (left hand side of A), which is a region that is well sampled in the previous protein-ligand binding free energy simulations. In each of the examples there

are two waters in the region that Heady *et al.* suggest that water should be located to bridge interactions between the CDK2 ligand and binding site residues (right hand side of A). At a lower B-value (-20), there are just one/two water molecules present in the CDK2 binding site. One of these one is in the solvent exposed region (left hand side of B), which again is a region we sample well in the previous CDK2 protein-ligand binding free energy simulations. Of greater interest is the water molecule in the region of space identified by Heady *et al.* as the region where water can bridge interactions between the CDK2 ligand and protein (right hand side of B). At this biasing potential the binding free energy of this water is ≈ -8 kcal.mol^{-1} indicating that this water is strongly bound.

The results from our GCMC simulations corroborate the study of Heady *et al.* [170] that there is a key water molecule bridging interactions between the CDK2 ligand and key binding site residues. This indicates that the system setup of our CDK2 study may well be incorrect, and hence this could be a reason for the poor results shown in the MM-RETI binding free energy study. In turn, any impact from this water molecule in MM will have a significant impact on the QM/MM corrections obtained for the bound state of perturbations within this dataset. Unfortunately further investigation into this is outside the scope of this study, however it is hoped that future work will reveal the true significance of water molecules within the CDK2 binding site on calculated binding free energies.

8.5 Protein – Ligand Charge Perturbations

As this QM/MM approach neglects any sampling of the QM/MM ensemble, the pathway-independence of the free energies obtained is validated. This is achieved

through the use of charge perturbation pathways. For CDK2, three compounds were selected to perform charge perturbations (Figure 8.19).

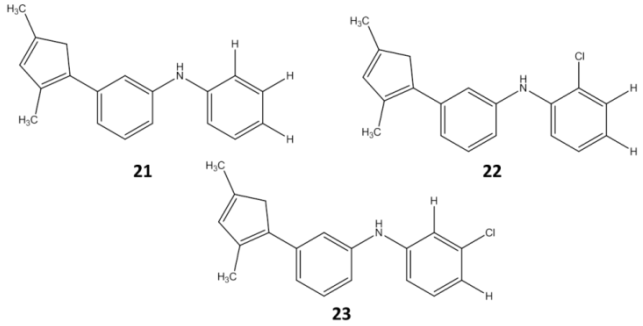


Figure 8.19. Three CDK2 ligands selected for charge perturbation simulations.

Monte Carlo Simulation Protocol

Generating alternative pathways by scaling solute charges up or down can be used to validate the pathway independence of calculated free energies. Alternative configurations were generated by performing RETI calculations in which the solute charges were scaled up. For protein-ligand systems the charges of the solute-environment interactions were scaled, while the charges used for the solute internal energy computation remained at their un-scaled level (λ=0). 16 λ windows (0.00, 0.06, 0.12, 0.19, 0.26, 0.33, 0.40, 0.47, 0.54, 0.61, 0.68, 0.75, 0.82, 0.88, 0.94, and 1.00) were used to ensure smooth transition between the two end states. The following charge scale factors were investigated: 1.01, 1.05, 1.07, 1.10, 1.15, 1.20 (N.B. 1.00 implies a simulation with non-perturbed charges). 10 million equilibration moves were performed before collecting statistics for 320 million moves in the bound legs and 160 million in the free legs, with each free energy simulation repeated 4 times. The RETI values shown are the mean of these four repeats and the standard error for these different runs is also shown.

QM/MM Single Point Energy Protocol

As with the 'normal' perturbations, configurations from the endpoint (λ=0 and λ=1) of the classical free energy simulations were selected and used as input for DFT-QM/MM single point energy calculations with Gaussian 09. One QM/MM single point energy calculation was performed every 100000th MC move. Therefore 3200 QM/MM configurations for the bound legs and 1600 QM/MM configurations for the free legs were used per repeat.

The QM energies were computed using the B3LYP hybrid density functional calculations with the 6-31G* basis set, as implemented in Gaussian 09.

Results & Discussion

The free energies obtained from the charge perturbations for CDK2/ligand 21 are shown in Figures 8.20 – 8.21.

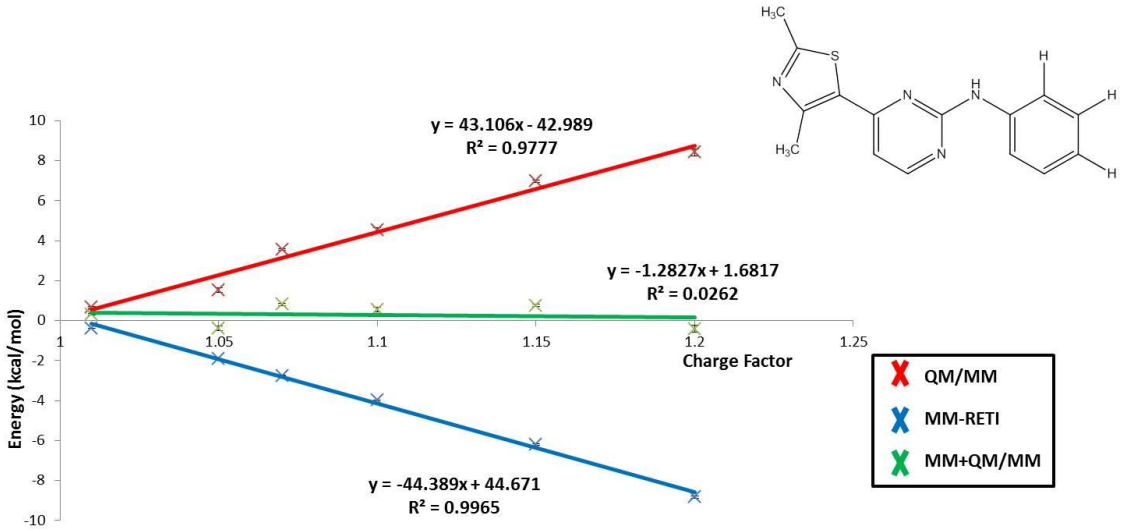


Figure 8.20. Charge perturbation results for ligand 21 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were generated from four independent simulations using standard error.

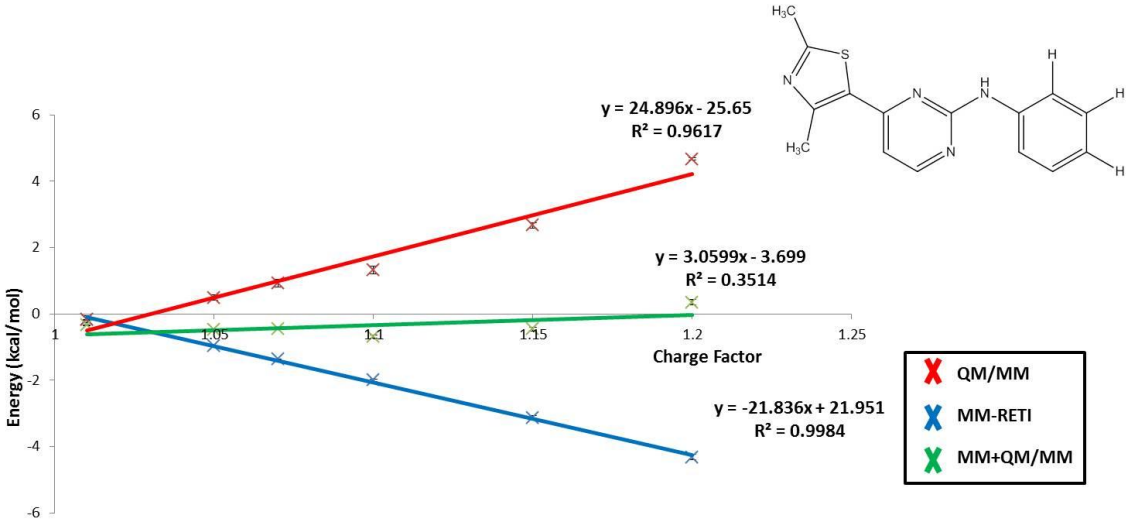


Figure 8.21. Charge perturbation results for ligand 21 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were generated from four independent simulations using standard error.

In Figures 8.20 – 8.21. the sums of our charge perturbed free energy cycles are generally small (small slopes of the green fitted lines). However, if the free energies are pathway independent, the mean values of these sums calculated over all scale factors must be equal to that of the non-charge perturbation MM→QM/MM free energies. The relevant free energies are summarised in Tables 4.4 and 4.5 in Supporting Information 4. The above condition is just fulfilled by ligand 21 in the free state. For ligand 21 in the free legs the average free energy of cycle closure is 0.28 (0.29) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -0.35 (0.31) kcal.mol^{-1} it is clear that our cycle just fails to agree. For the bound state of ligand 21 the above condition is fulfilled. Ligand 21 obtains an average free energy cycle closure of -0.34 (0.26), which is very similar to the original MM→QM/MM free energy of -0.24 (0.29) kcal.mol^{-1}. The less satisfactory agreement for ligand 21 in the free state would suggest that additional sampling is needed in order to obtain more accurate results for this ligand in the aqueous free energy leg.

The free energies obtained from the charge perturbations for CDK2/ligand 22 are shown in Figures 8.22 – 8.23.

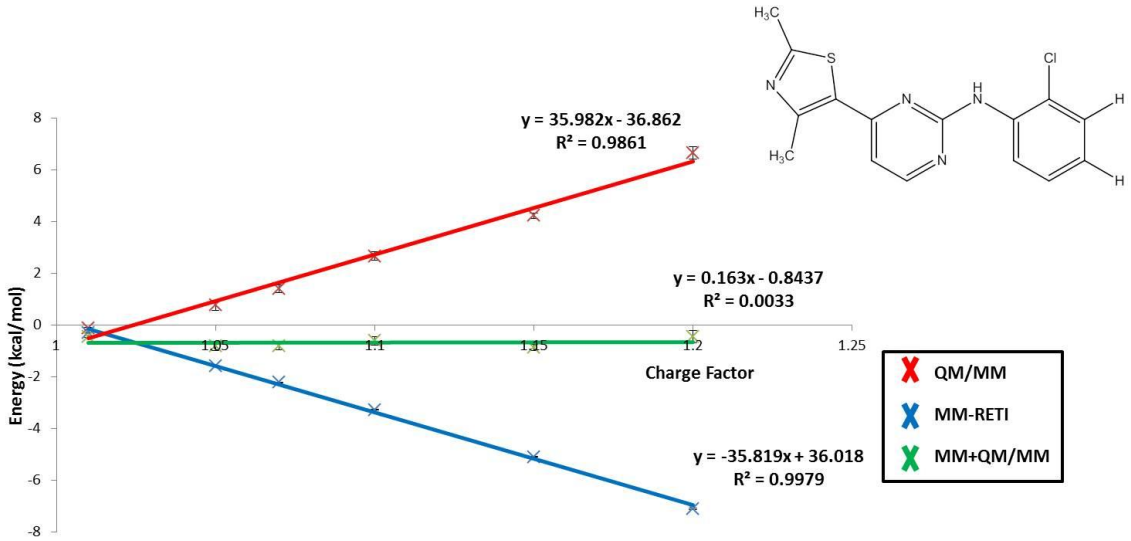


Figure 8.22. Charge perturbation results for ligand 22 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were generated from four independent simulations using standard error.

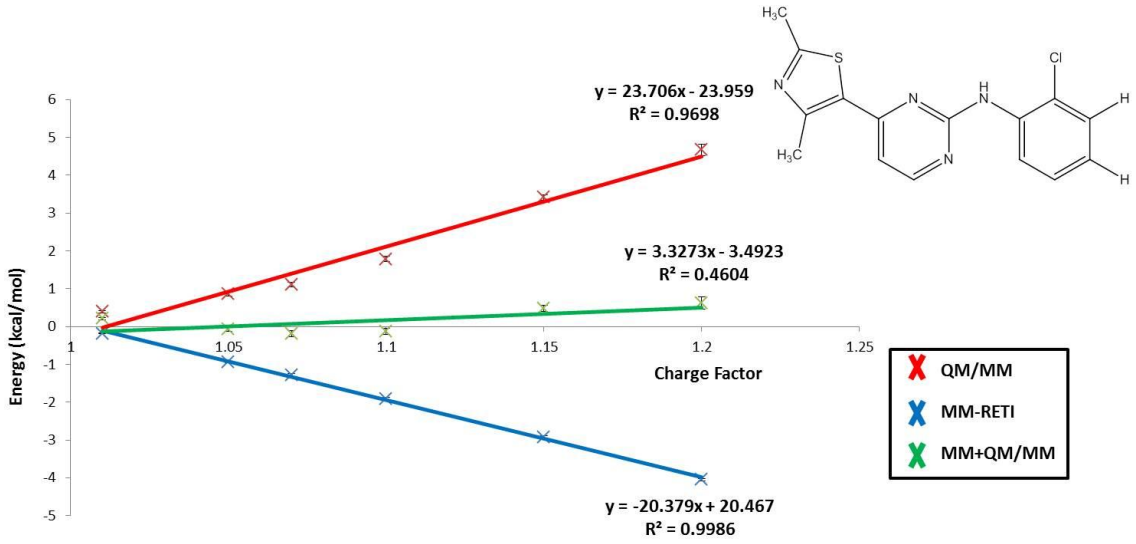


Figure 8.23. Charge perturbation results for ligand 22 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were generated from four independent simulations using standard error.

In Figures 8.22 – 8.23 the sums of our charge perturbed free energy cycles are generally small (small slopes of the green fitted lines). However, if the free energies are pathway independent, the mean values of these sums calculated over all scale factors must be equal to that of the non-charge perturbation MM→QM/MM free energies. The relevant free energies are summarised in Tables 4.6 and 4.7 in Supporting Information 4. The above condition is fulfilled by ligand 22 in the free state. For ligand 22 in the free legs the average free energy of cycle closure is -0.27 (0.23) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -0.72 (0.32) kcal.mol^{-1} it is clear that our cycle obtains a similar value. For the bound state of ligand 22 the above condition is fulfilled. Ligand 22 obtains an average free energy cycle closure of 0.16 (0.25), which is very similar to the original MM→QM/MM free energy of 0.22 (0.22) kcal.mol^{-1}.

The free energies obtained from the charge perturbations for CDK2/ligand 23 are shown in Figures 8.24 – 8.25.

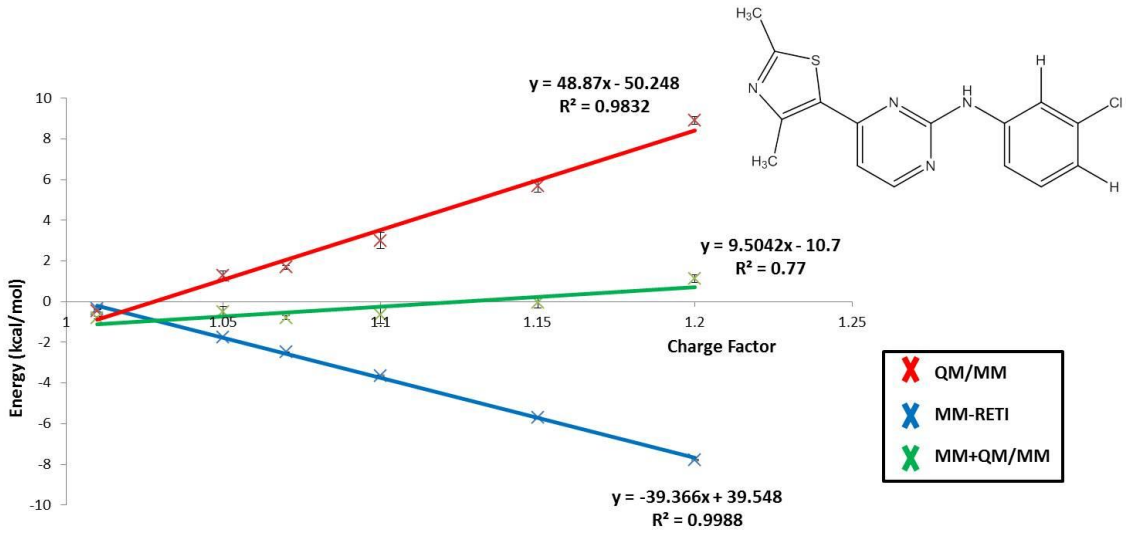


Figure 8.24. Charge perturbation results for ligand 23 in the free legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were generated from four independent simulations using standard error.

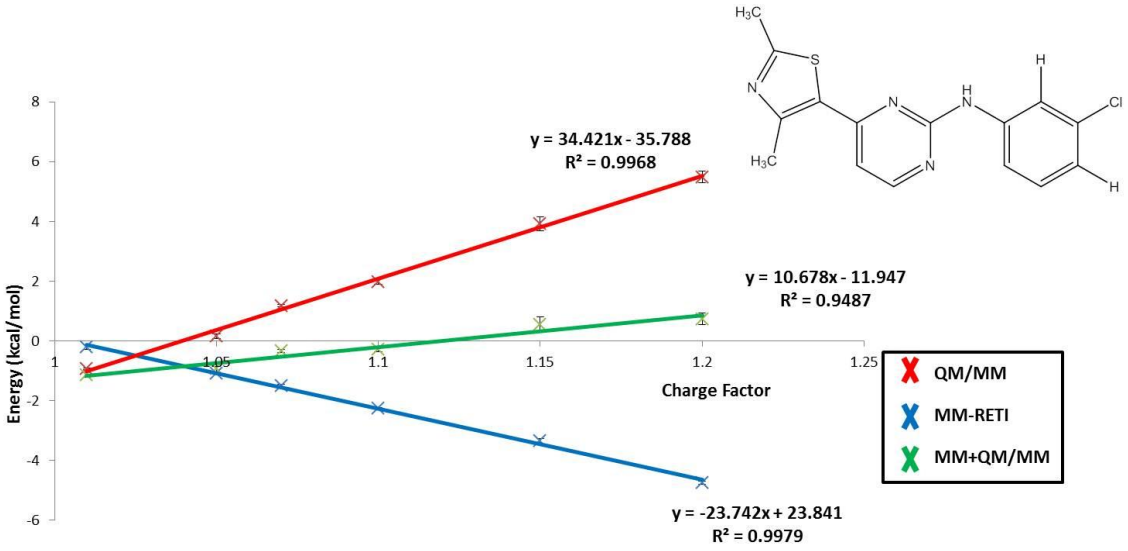


Figure 8.25. Charge perturbation results for ligand 23 in the bound legs. The red line represents the QM/MM free energies, the blue line shows the MM-RETI results and the green line shows the combined MM→QM/MM free energies. The error bars shown were generated from four independent simulations using standard error.

In Figures 8.24 – 8.25 the sums of our charge perturbed free energy cycles are generally small (small slopes of the green fitted lines). The relevant free energies are summarised in Tables 4.8 and 4.9 in Supporting Information 4. The above condition is fulfilled by ligand 23 in the free state. For ligand 23 in the free legs the average free energy of cycle closure is -0.38 (0.32) kcal.mol^{-1}. Comparing this value to the original MM→QM/MM free energy of -0.33 (0.27) kcal.mol^{-1} it is clear that our cycle does obtain a similar value. For the bound state of ligand 23 the above condition is not fulfilled. Ligand 23 obtains an average free energy cycle closure of -0.24 (0.32), which is not similar to the original MM→QM/MM free energy of -1.05 (0.21) kcal.mol^{-1}. The poor agreement for ligand 23 in the bound state would suggest that additional sampling is needed in order to obtain more precise results for this ligand in the bound free energy leg.

8.6 Conclusions

The aim of this study was to understand if applying the simplistic QM/MM method could produce more accurate MM→QM/MM binding free energies for a set of CDK2 inhibitors. The results from the MM binding free energy study were extremely poor and any attempt to perform QM/MM corrections to these classically obtained free energies led to no improvement. Further analysis into poor performance of the free energy studies identified that the system setup was incorrect and a key water molecule that bridges interaction between the inhibitors and CDK2 was not present in this free energy study. This water was found to bind extremely favourably in GCMC calculations on the CDK2 binding site. Therefore, current work aims to perform free

energy simulations with this water present to ascertain if the addition of this can lead to an increase in accuracy between predicted and experimental binding affinities.

9 Conclusions & Future Perspectives

9.1 Conclusions

Here is presented the development and application of a simplified QM/MM method for the calculation of hydration free energies and protein-ligand free energies.

In the hydration free energy study (Section 5) it was seen that this QM/MM method can perform as well as conventional MM. We also identified that the QM/MM method causes large over-polarisation for several compounds, most noticeably those with hydrogen bonding moieties. This is caused by close contacts between the embedded MM point charges and the QM ligand leading to charge exchange between our QM ligand and MM point charges, and hence the QM ligand becoming over-polarised. The impact of this can be minimised with the adaption of this method to include a Gaussian Blurring technique (section 5.2). However, knowledge of the extent of over-polarisation is needed *a priori* in order for the application of this technique to show improvement on the conventional QM/MM approach. It is believed that further investigation into this and other similar methods could yield a more powerful and diverse QM/MM method, which would be able to deal with more polar systems.

For the protein-ligand binding free energy studies many different conclusions can be drawn depending upon the target protein system. For example with COX-2 (section 6), it was found that the QM/MM corrected free energies showed a loss of accuracy and precision compared to conventional MM. This is believed to be caused by the small size of the COX-2 pocket into which all of the perturbations are directed causing close contacts to form between the larger QM ligands and our embedded protein point charges; these close contacts have a similar effect to close contacts found in

solution. Therefore, it would be advisable for such protein-ligand complexes to include these key binding site residues as part of the QM region in order to increase the accuracy in the description of the QM/MM system.

In contrast, for neuraminidase (section 7) the extremely polar nature of the ligands and binding site lead to very large negative QM/MM corrections in both the bound and free legs of the free energy simulations. This leads to a large negative shift in most of the free energies calculated; despite this the correlation is preserved between MM and QM/MM calculated binding free energies and experiment. The MUE increase shows the large amount of over-polarisation experienced in our QM/MM simulations. It is believed that performing a Gaussian Blurring approach to Neuraminidase protein-ligand complexes would be the best route to minimising the impact of these large QM/MM corrections.

CDK2 (section 8) fails to show any correlation between either MM or QM/MM calculated binding free energies and experiment. GCMC analysis (section 8.1.3) into this has shown that key binding site waters were not present in the free energy study presented here. It is hoped that future work including this key water will give improved results for both MM and QM/MM, yet it does show that if MM completely fails to predict binding free energies than our QM/MM method will not be able to improve on this.

Therefore, after applying our QM/MM method to several different case studies it is possible to conclude that for hydration free energies of small organic molecules this method performs well. A Gaussian Blurring approach was also have also implemented, which enables us to adjust the embedding strategy for polar ligands. It has also been shown that the application of the QM/MM method to calculate protein-ligand free

energies generally leads to a loss of accuracy and precision, when compared to MM. This could be indicative that the single-step QM/MM approach taken here is too simplistic for such systems. This theory is supported by the charge perturbations which do show convergence errors in the calculated free energies for protein-ligand systems. However, it is believed that this could be mitigated by the inclusion of key binding site residues as part of the QM region, and also by the use of more sophisticated embedding techniques, such as the Gaussian Blurring technique which was applied to the hydration free energy study. It is also thought that including a way to sample the QM/MM ensemble and so avoid doing the MM→QM/MM perturbation in one step. This could be done using the approach of Woods *et al.* [110] where they simulate using MM and accept the configurations using a Metropolis-Hastings acceptance test to create a more expensive QM/MM ensemble.

9.2 Future Perspectives

The future work in this area has several possibilities which need to be explored in order to obtain a QM/MM method capable of producing more accurate free energies, particularly for protein-ligand systems.

Firstly, the embedding of the MM environment into QM/MM must be investigated further. Studies within this thesis (sections 6, 7 and 8) have shown the need for a more elegant embedding strategy that standard EE when using this method. This could be adapted using the Gaussian blurring technique (section 5.2) as was done for the hydration free energy study within this work. Alternatively, it could be of promise to invest in polarisable MM potentials [17] and hence give rise to a practical PE strategy. This could also be achieved through the use of well parameterized

reference potentials, i.e. the approached taken in EVB theory [104] or MMBIF [106], which could be iteratively updated, such as in the work of Thompson and Schenter [72], and Bakowies and Thiel [73].

Secondly, within this work no coupling of the Lennard-Jones between MM and QM regions of the systems studied was considered. This means that it is possible to explore the possibility of including dispersion within the QM/MM simulations to ascertain if this can benefit this QM/MM method. Studies from Mulholland *et al.* have shown that activation barriers can be shifted by as much a 5 – 10 kcal.mol^{-1} when including dispersion effects when calculating activation barriers for cytochrome P450 reaction mechanisms [91, 92, 93].

Lastly, it appears that for protein-ligand systems that important binding site residues should be included as part of the QM region within QM/MM simulations. This would involve the use of LA's [77, 78, 79, 80] or GHO's [81, 82, 83, 84], which would make the coupling of the QM and MM subsystems more complex. However, it is believed that the use of a more complex coupling scheme combined with dispersion and more elegant embedding strategies is needed to produce free energies with a greater degree of accuracy.

Overall there are several challenges that remain to produce a QM/MM method that can be as versatile and generally applicable as conventional MM forcefields. This thesis has shown that QM/MM can produce results that are very comparable to MM for hydration free energies. It has also been highlighted that for more complex systems, such as protein-ligand systems, QM/MM struggles to achieve the same accuracy as standard MM forcefields. Many suggestions have been included in this

work as to how QM/MM methods can be improved to enable them to achieve and hopefully exceed the accuracy of MM forcefields in the future.

References

1. PDB website:

http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100

(**visited on 30/9/2012**)

2. Song, C.M., Lim, S.J., and Tong, J.C. *Brief Bioinform.* 10(5), 579-591 (**2009**)

3. Rao, S.N., Singh, U.C., Bash, P.A., and Kollman, P.A. *Nature* 328, 551-554 (**1987**)

4. Mizushimag, N., Spellmayer, D., Hironoll, S., Pearlman, D., and Kollman, P.A. *J. Bio. Chem.* 226(13), 11801-11991 (**1995**)

5. Wright, L.B., Walsh, T.R., *J. Phys. Chem. C*. 116 (4), 2933–2945 (**2012**)

6. Sousa, S.F., Fernandes P.A., and Ramos, M.J. *Proteins* 65(1), 15-26 (**2006**)

7. Pospisil, P., Kuoni, T., Scapozza, L., and Folkers, G. *J. Recpet. Signal Transduct. Res.* 22, 141-154 (**2002**)

8. Wang, C.X., Shi, Y.Y., Zhou, F., and Wang, L., *Proteins* 15(1), 5-9 (**1993**)

9. Jiao, D., Zhang, J., Duke, R.E., Li, G., Schnieders, M.J., and Ren, P. *J. Comput. Chem.* 30(11), 1701-1711 (**2009**)

10. Jorgensen, W.L., Price, M.L.P., Price, D.J., Rizzo, R.C., Wang, D., Pierce, A.C., and Tiraso-Rives, J. *Free Energy Calculations in Rational Drug Design* Reddy, M.R., and Elion, M.D. (Eds.), Kluwer, 299-316 (**2001**)

11. Wang, D.P., Rizzo, R.C., Tirado-Rives, J., and Joregensen, W.L. *Bioord. Med. Chem. Lett.* 11, 2799-2802 (**2001**)

12. Wesolowski, S., and Jorgensen W.L. *Bioorg. Med. Chem. Lett.* 12, 267-270 (**2002**)

13. Wang, J., Cieplak, P., and Kollman, P. A. *J. Comp. Chem.* 21(12), 1049–1074 (**2000**).

14. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. *J. Comp. Chem.* 25(9), 1157–1174 (**2004**).

15. Jorgensen, W.L., and Tirado-Rives, J. *J. Am. Chem. Soc.* 110, 1657-1666 (**1988**)

16. van Gunsteren, W.F., and Berendsen, H.J.C., Groningen Molecular Simulation (GROMOS) Library Manual, BIOMOS b.v., Groningen (**1987**)

17.Warshel, A., Kato, M., and Pisliakov, A.V. *J. Chem. Theory Comput.* 3(6), 2034-2045 (**2007**)

18. Duan, Y., Wu, C., Chowdhury, S., Lee., M.C., Xiong, G., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J., and Kollman, P.A. *J. Comput. Chem.* 24, 1999-2012 (**2003**)

19.Hou, Q., Du, L., Gao, J., Liu, Y., and Liu, C. *J. Phys. Chem. B.* 114(46), 15296-15300 (**2010**)

20. Wong, K.Y., and Gao, J., *Biochemistry* 46, 13352-13369 (**2007**)

21. Dubey, K.D., and Ojha, R.P. *J. Bio. Phys.* 37(1), 69-78 (**2011**)

22. Varnai, C., Bernstein, N., Mones, L., and Csanyi G. *J. Phys. Chem. B.* online edition (**2013**)

23. Heimdal, J., and Ryde, U. *Phys. Chem. Chem. Phys.* 14 12592-12604 (**2012**)

24. Voelz, V., Bowman, G. R., Beauchamp, K., and Pande, V. S. *J. Am. Chem. Soc.* 132(5), 1526–1528 (**2010**)

25. Orsi, M. and Essex, J. W. *Soft Matter* 6(16), 3797–3808 (**2010**)

26. Bemporad, D., Luttmann, C., and Essex, J.W. *Biophys. J.* 87(1), 1-13 (**2004**)

27. Leach, A. Molecular Modelling : Principles and Applications. 2nd edition, (**2001**).

28. Allen, M. P. and Tildesley, D. J. Computer Simulation of Liquids. (**1989**).

29. Atkins, P. W. Physical Chemistry. 3rd edition, (**1988**).

30. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. *J. Comp. Chem.* 25(9), 1157–1174 (**2004**).

31. Oostenbrink, C., Villa, A., Mark, A. E., and Van Grunsteren,W. F. *J. Comp. Chem.* 25(13), 1656–1676 (**2004**).

32. Lee, F.S., Chu, Z.T., and Warshel, A. *J. Comp. Chem.* 14, 161 (**1993**)

33. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. *J. Comp. Chem.* 4(2), 187-217 (**1983**)

34. Ponder, J.W., and Chuanjie, W. *J. Phys. Chem. B.* 114(8), 2549-2564 (**2010**)

35. Piquemanl, J-P., Williams-Hubbard, B., Fey, N., Deeth, R.J., Gresh, N., and Giessner-Prette, C., *J. Comput. Chem.* 24, 1963 (**2003**)

36. Le Sueur, C.R., and Stone, A.J. *Molec. Phys.* 78, 1267-1291 (**1993**)

37. Szabo, A. and Ostlund, N. Introduction to Advanced Electronic Structure Theory. (1996)

38. Helgaker, T., Joregensen, P., and Olse, J. Molecular Eltectronic-Structure Theory (**2013**)

39. Atkins, P.W., and Friedman, R.S. Molecular Quantum Mechanics, 3rd edition (**1996**)

40. Bohr, N., and Oppenheimer, R. *Annalen der Physik* (in German) 389 (20), 457–484. (**1927**)

41. Parr, R.G., and Yang, W., Density-Functional Theory of Atoms and Molecules. (**1994**)

42. Perdew, J.P., Burke, K., and Ernzerhof, M., *Phys. Rev. Lett.* 77, 3865-3868 (1996)

43. Sholl, D.S., and Stackel, J.A. Density Functional Theory: A Practical Introduction (**2009**)

44. Hohenburg, P., and Kohn, W. *Phys. Rev.* 136, 864 (**1964**)

45. Kohn, W., and Sham, L.J. *Phys. Rev.* 140, 1133-1138 (**1965**)

46. Boys, S.F., *Pro. Royal Soc. A.* 200, 542-554 (**1950**)

47. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. *J. Chem. Phys.* 21(6), 1087 (**1953**)

48. Michel, J. The use of free energy simulations as scoring functions. PhD thesis, (**2006**).

49. Zwanzig, R. W. *J. Chem. Phys.* 22(8), 1420–1426 (**1954**)

50. Mezei, M. *J. Chem. Phys.* 86, 7084 (**1987**)

51. Woods, C. J., Essex, J. W., and King, M. A. *J. Phys. Chem. B* 107(49), 13711–13718 (**2003**)

52. Woods, C. J., Essex, J. W., and King, M. A. *J. Phys. Chem. B* 107(49), 13703–13710 (**2003**)

53. Guimaraes, C. R. W. and Mathiowetz, A. M. *J. Chem. Inf. Model.* 50(4), 547–559 (**2010**)

54. Graves, A. P., Shivakumar, D. M., Boyce, S. E., Jacobson, M. P., Case, D. A., and Shoichet, B. K. *J. Mol. Biol.* 377(3), 914–934 (**2008**)

55. Adams, D. *J. Mol. Phys.* 28(5), 1241–1252 (**1974**)

56. Adams, D. *J. Mol. Phys.* 29(1), 307–311 (**1975**)

57. Warshel, A., and Levitt, M. *J. Mol. Biol.* 103(2), 227-249 (**1976**)

58. Senn, H., and Thiel, W. *Curr. Opin. Chem. Biol.* 11(2), 182-187 (**2007**)

59. Torras, J., Deumens, E., and Trickey, S.B. *J. Comput. Aided Mat. Des.* 13, 201-212 (**2006**)

60. Senn, H.M, and Thiel, W. *Angew. Chem. Int. Ed. Engl.* 48(7), 1198-1229 (**2009**)

61. van der Kamp, M.W., and Mulholland, A. J. *Biochemistry* 52(16), 2708-2728 (**2013**)

62. Dapprich, S., Komaromi, I., Byun, K.S., Morokuma, K. and Frisch, M. J. *J. Mol. Struct. (Theochem)* 462, 1 (**1999**)

63. Vreven, T., Morokuma, K., Farkas, O., Schlegel, H.B., and Frisch, M.J. *J. Comp. Chem.* 24, 760 (**2003**)

64. Vreven, T., and Morokuma, K. *Annu. Rep. Comp. Chem.* 2, 35-51 (**2006**)

65. Vreven, T., Byun, K.S., Komaromi, I., Dapprich, S., Montgomery, J.A., Morokuma, K., and Frisch, M.J. *J. Chem. Theory Comput.* 2, 815-826 (**2006**)

66. Yan, Y., and Zhang, L., *Commun. Comput. Chem.* 1(2), 109-117 (**2001**)

67. Sprik, M., and Klein, M.L. *J. Chem. Phys.* 89, 7556 (**1988**)

68. Lamoureuz, G. MacKerell, A.D., and Roux, B. *J. Chem. Phys.* 119, 5185 (**2003**)

69. Martinez, J.M., Hernandez-Cobos, J. Saint-Martin, H., Pappalardo, R.R., Ortega-Blake, I., and Marcos, E.S. *J. Chem. Phys.* 112, 2339 (**2000**)

70. Rick, S.W., and Stuart, S.J. *Rev. Comput. Chem.* 18, 89 (**2003**)

71. Ponder, J.W., and Case, D.A. *Adv. Protein Chem.* 66, 27 (**2003**)

72. Thompson, M.A., and Schenter, G.K. *J. Phys. Chem.* 99, 6374 (**1995**)

73. Bakowies, D., and Thiel, W. *J. Phys. Chem.* 100, 10580 (**1996**)

74. Storer, J.W., Giesen, D.J., Hawkins, G.D., Lynch, G.C., Cramer, C.J., Truhlar, D.G., and Liotard, D.A. *ACS Symposium Series* 568, 24 (**1994**)

75. Rinaldi, D., and Rivali, J.L. *Theor. Chim. Acta.* 32, 57 (**1973**)

77. Maseras, F., and Morokuma, K. *J. Comput. Chem.* 16, 1170 (**1995**)

78. Field, M.J., Bash, P.A., and Karplus, M. *J. Comput. Chem* 11, 700 (**1990**)

79. Antes, I., and Thiel, W. *J. Phys Chem. A.* 103, 9290 (**1999**)

80. Byun, K., and Gao, J. *J. Mol. Grap. Mod.* 18, 50 (**2000**)

81. Dilabio, G.A., Hurley, M.M, and Christiansen, P.A. *J. Chem. Phys.* 116, 9578 (**2002**)

82. Monard, G., Loos, M., Thery, V., Baka, K., and Rivali, J-L. *Int. J. Quantum Chem.* 58, 153 (**1996**)

83. Ferre, N., Assfeld, X., Rivali, J-L. *J. Comput. Chem.* 23, 610 (**2002**)

84. Gao, J., Amara, P., Alhambra, C., and Field, M.J., *J. Phys. Chem. C.* 102, 4714 (**1998**)

85. Truhlar, D.G., Gao, J., Alhambra, C., Garcia-Viloca, M., and Vialla, J. *Acc. Chem. Res.* 35, 341 (**2002**)

86. Zhang, Y., Lee, T-S., and Wang, W. *J. Chem. Phys.* 110, 46 (**1999**)

87. Reuter, N., Dejaegere, A., Maigret, B., and Karplus, M. *J. Phys. Chem. A.* 104, 1720 (**2000**)

88. Hu, H., Lu, Z., and Yang, W. *J. Chem. Theory. Comput.* 3(2), 390-406 (**2007**)

89. Rodriguez, Q., Olivia, C., and Gonzalez, M. *Phys. Chem. Chem. Phys.* 12, 8001-8015 (**2010**)

90. Jun, X., Zhang, J.Z.H., and Xiang, Y. *J. Am. Chem. Soc.* 134(39), 16424-16429 (**2012**)

91. Lonsdale, R., Houghton, K.T., Zurek, J., Bathelt, C.M., Folopp, N., de Groot, M.J., Harvey, J.N. and Mulholland A.J. *J. Am. Chem. Soc.* 135, 8001-8015 (**2013**)

92. Lonsdale, R., Harvey, N.J., and Mulholland, A.J. *J. Chem. Theory. Comput.* 8, 4637-4645 (**2012**)

93. Lonsdale, R., Harvey, N.J., and Mulholland, A.J. *J. Phys. Chem. Lett.* 1, 3232-3237 (**2010**) 94. Lodola, A., Sirirak, J., Fey, N., Rivara, S., Mor, M., and Mulholland, A.J., *J. Chem. Theor. Comput.* 6, 2948-2960 (**2010**)

94. Bowman, A.L., Grant, I.M., and Mulholland, A.J. *Chem. Commun.* 37, 4425-4427 (**2008**)

Chem. Commun. (2008) (37) pp. 4425-4427

95. Hu, H., and Yang, W. *Annu. Rev. Phys. Chem.* 59, 573-601 (**2008**)

96. Gao, J., Garner, D.S., and Jorgensen, W.L. *J. Am. Chem. Soc.* 108, 4784-4790 (**1986**)

97. Singh, U.C., and Kollman, P.A. *J. Comp. Chem.* 7(6), 718-730 (**1986**)

98. Weiner, S.J., Seibel, G.L., and Kollman, P.A. *Proc. Natl. Acad. Sci. USA.* 83(3), 649-653 (**1986**)

99. Kollman, P.A., Kuhn, B., and Peraklya, M. *J. Phys. Chem. B.* 106(7), 1537-1542 (**2002**)

100. Liu, H., Zhang, Y., and Yang, W. *J. Am. Chem. Soc.* 122, 6560-6570 (2000)

101. Wu, P. Andres-Cisneros. G., Hu, H., Chaudret, R., Hu, X., and Yang, W. *J. Chem. Phys. B.* 116(23), 6889-6897 (**2012**)

102. Ishida, T., and Kato, S. *J. Am. Chem. Soc.* 125(39), 12035-12048 (**2003**)

103. Plotnikov, N.V., Kamerlin, S.C., and Warshel, A. *J. Phys. Chem. B.* 115(24), 7590-7562 (**2011**)

104. Warshel, A., and Weiss, R.M., *J. Am. Chem. Soc.* 102(20), 6218-6226 (**1980**)

105. Bogdan, L., Roux, B., and Noskov, S, Y. *J. Chem. Theory Comput.* 9(9), 4165-4175 (**2013**)

106. Iftimie, R., Salahub, D., Wei, D., and Schofield, J. *J. Chem. Phys.* 113, 4852 (**2000**)

107. Senn, H.M., and Thiel, W. *Curr. Opinion Mol. Biol.* 11(2), 182-187 (**2007**)

108. Rathore, R.S., Sumakanth, M., Reddy, M.S., Reddanna, P., Rao, A.A., Erion, M.D., and Reddy, M.R. *Curr. Pharm. Des.* 19(26), 4674-4686 (**2013**)

109. Reddy, M.R., and Erion, M.D. *J. Comput. Aided Mol. Des.* 23(12), 837-843 (**2009**)

110. Woods, C.J., Manby, F.R., and Mulholland A.J., *J. Chem. Phys.* 128(1), 14109 (**2008**)

111. Beierlein, F.R., Michel, J., and Essex, J.W. *J. Phys. Chem. B.* 115(17), 4911-4926 (**2011**)

112. Sitkoff, D., Sharp, K.A., and Honig, B. *J. Phys. Chem.* 98(7), 1978-1988 (**1994**)

113. Shirts, M.R., Mobley, D.L., and Chodera, J.D. *Annu. Rev. Comp. Chem.* 3, 41-59 (**2007**)

114. Chipot, C., and Pohorille, P. Free Energy Calculations: Theory and Applications in Chemistry and Biology. (2007)

115. Kuntz, I.D., Blaney, J.M, Oatley, S.J., Langridge, R., and Ferrin, T.E. *J. Mol. Biol.* 161(2), 269-288 (1982)

116. Jones, G., Willett, P., Glen, R.C., Leach, A.R., Taylor, R. *J. Mol. Biol.* 267(3), 727-748 (**1997**)

117. Grosdidier, A., Zoete, V., and Michielin, O. *J. Comput. Chem.* 30(13), 2021-2030 (**2009**)

118. Irwin, J.J., Schoichet, B.K., Mysinger, M.M., Huang, N., Colizzi, F., Wassam, P., and Cao, Y. *J. Med. Chem.* 52(18), 5712-5720 (**2009**)

119. Hummer, G., Pratt, L.R., and Garcia, A.E. *J. Phys. Chem.* 100, 1206-1215 (**1996**)

120. Rizzo, R.C., Aynechi, T., Case, D.A., and Kuntz, I.D. *J. Chem. Theory Comput.* 2, 128-139 (**2006**)

121. Huggins, D.J., and Payne, M.C. *J. Phys. Chem. B.* 117(27), 8232-8244 (**2013**)

122. Mobley, D.L., Bayly, C.I., Cooper, M.D., Shirts, M.R., and Dill, K.A. *J. Chem. Theory Comput.* 5(2), 350-358 (**2009**)

123. Woods, C. and Michel, J. (**2007**).

124. Jakalian, A., Bush, B. L., Jack, D. B., and Bayly, C. I. *J. Comp. Chem.* 21(2), 132–146 (**2000**).

125. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R.W., and Klein, M. L. J. Chem. Phys. 79(2), 926–935 (**1983**)

126. Gaussian 09, Revision **D.01**, Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng,

G., Sonnenberg, J. L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery, J. A., Jr., Peralta, J. E., Ogliaro, F., Bearpark, M., Heyd, J. J., Brothers, E., Kudin, K. N., Staroverov, V. N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J. C., Iyengar, S. S., Tomasi, J., Cossi, M., Rega, N., Millam, N. J., Klene, M., Knox, J. E., Cross, J. B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R. E., Yazyev, O., Austin, A. J., Cammi, R., Pomelli, C., Ochterski, J. W., Martin, R. L., Morokuma, K., Zakrzewski, V. G., Voth, G. A., Salvador, P., Dannenberg, J. J., Dapprich, S., Daniels, A. D., Farkas, Ö., Foresman, J. B., Ortiz, J. V., Cioslowski, J., and Fox, D. J. Gaussian, Inc., Wallingford CT, **2009**.

127. J. P. Guthrie, review in preparation

128. Riccardi, D., Schaefer, P., and Cui, Q. *J. Phys. Chem. B.,* 109, 17715-17733 (**2005**)

129. Gascon, J. A., and Batista, V. S. *Biophys. J.* 87(5), 2931-2941 (**2004**)

130. Fox, S.J., Pittock, C., Fox, T., Tautermann, C.S., Christ, C., Malcolm, N.O.J., and Skylaris C-K. *J. Chem. Phys.* 135(22), 224107 (**2011**)

131. Fox, S.J., Pittock, C., Fox, T., Tautermann, C.S., Christ, C., Malcolm, N.O.J., Essex, J.W., and Skylaris C-K. *J. Phys. Chem. B.* 117(32), 9478-9485 (**2013**)

132. Zazharias, M., Straatsma, T.P., and McCammon, J.A. *J. Chem. Phys.* 100, 9025 (**1994**)

133. Cheng, X., Cui, G., Hornak, V., and Simmerling C. *J. Phys. Chem. B.* 109, 8220-8230 (**2005**)

134. Wang, B., and Truhlar, D.G., *J. Chem. Theory. Comput.* 8, 1989-1998 (**2012**)

135. Wang, B., and Truhlar, D.G. *J. Chem. Theory. Comput.* 9, 1036-1042 (**2013**)

136. Isegawa, M., Wang, B., and Truhlar, D.G. *J. Chem. Theory. Comput.* 9, 1381-1393 (**2013**)

137. Isegawa, M., Fieldler, L., Leverentz, H.R., Wang, Y., Nachimuthu, S., Gao, J., and Truhlar, D.G. *J. Chem. Theory. Comput.* 9, 33-45 (**2013**)

138. Sherwood, P., de Vries, A. H., Guest, M.F., Schreckenbach, G., Richard, C., Catlow, A., French, S.A., Sokol, A.A., Bromlev, S.T., Thiel, W., Turner, A.J., Billeter, S., Terstegen, F., Thiel, S., Kendrick, J., Rogers, S.C., Casci, J., Watson, M., King, F., Karlsen, E., Siovoll, M., Fahmi, A., Schafer, A., and Lennartz, C. *J. Mol. Struc. (Theochem)* 632, 1-28 (**2003**)

139. Crofford, L.J, Lipsky, P.E., Brooks, P., Abramson, S. B., Simon, L.S., and van de Putte, L.B. *Arthritis Rheum.* 43, 4 (**2000**).

140. Weggen, S., Eriksen, J.L., Das, P., Sagi, S.A., Wang, R., Pietrzik, C.U., Findlay, K.A., Smith, T.E., Murphy, M.P., Butler, T., Kang, D.E., Marquez-Sterling, N., Golde, T.E. and Koo, E.H. *Nature* 414, 212 (**2001**).

141. Zhou, Y., Su, Y., Li. B., Liu, F., Ryder, J.W., Wu, X., Gonzalez-DeWhitt, P.A., Gelfanova, V., Hale, J.E., May, P.C., Paul, S.M., and Ni, B. *Science* 302, 1215 (**2003**).

142. Schiff, S.J., Shivaprasad, P., and Santini, D.L. *Curr. Opin. Pharmacol.* 3, 352 (**2003**).

143. Selinksky, B. S., Gupta, K., Sharkey, C. T., and Loll, P. J. *Biochemsitry.* 40, 5172–5180 (**2001**).

144. Furse, K.E., Pratt, D.A., Porter, N.A., and Lybrand T.P. *Biochemistry* 14, 3189-3205 (**2006**)

145. Price, M.L., Jorgensen, W.L. *J. Am. Chem. Soc.* 122, 9455-9466 (**2000**)

146. Michel, J., Verdonk, M.L., Essex, J.W., *J. Med. Chem.* 49, 7427-7439 (**2006**)

147. 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* 384, 644-648 (**1996**)

148. Word, J.M., Simon C. Lovell, S.C., Richardson, J.S. and Richardson, D.C. *J. Mol. Bio.* 285, 1735-1747, (**1999**)

149. Pearlman, D.A., and Charifson, P.S. *J. Med. Chem.* 44, 3417-3423 (**2001**)

150. Kendall, M.G. *Biometrika* 30, 81-89 (**1938**)

151. Limongelli, V., Bonomi, M., and Parinello, M. *PNAS* 110, 6358-6363 (**2013**)

152. Woo, H.-J., Dinner, A. R., and Roux, B. *J. Chem. Phys.* 121(13), 6392–400 (**2004**)

153. Mezei, M. *Mol. Phys.* 40(4), 901–906 (**1980**)

154. Bodnarchuk, M.S. Predicting the Location and Binding Affinity of Small Molecules in Protein Binding Sites, PhD Thesis (**2012**), University of Southampton

155. Moscana, A. *N. Engl. J. Med.* 353, 1363-1373 (**2005**)

156. Masukawa, K.M., Kollman, P.A., and Kuntz, I.D. *J. Med. Chem.* 26, 5628-5637 (**2003**)

157. Ripoli, D.R., Khavrutskii, I.V., Chaudhury, S., Liu, J., Kuschner R.A., Wallqvist, A., and Reifman, J. *PLOS* 8(8), 1-10 (**2012**)

158. Smith, P.W., and Sollis, S.L. *J. Med. Chem.* 41, 787-797 (**1998**)

159. Taylor, N.R. *J. Med. Chem.* 41(6), 798-807 (**1998**)

160. Cao, J. Bjornsson, R., Buhl, M., Thiel, W., and van Mourik, T. *Chemisty* 18(1), 184-195 (**2012**)

161. Cao, J., Bjornsson, R., Buhl, M., Thiel, W., and van Mourik, T. *Chemisty* 18(1), 184-195 (**2012**)

162. Woods, C.J., King, M.A., Essex, J.W. *J. Comput. Aided Mol. Design* 15, 129-144 (**2001**)

163. Bonnet, P., Bryce, R.A., *Prot. Sci.* 13, 946-957 (**2004**)

164. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., and Sudarsanam, S. *Science* 298(5600), 1912–34 (**2002**)

165. Malumbres, M. and Barbacid, M. *Nat. Rev. Can.* 9(3), 153–166 (**2009**).

166. Robinson, D. D., Sherman, W., and Farid, R. *Chem. Med. Chem.* 5(4), 618–627 (**2010**).

167. Michel, J., Verdonk, M. L., and Essex, J. W. *J. Med. Chem.* 49(25), 7427–7439 (**2006**).

168. Pierce, A. C., Sandretto, K. L., and Bemis, G. W. *Proteins: Struct. Funct. Bioinf.* 49(4), 567–576 (**2002**).

169. Kontopidis, G., McInnes, C., Pandalaneni, S.R., McNae, I., Gibson, D., Mezna, M., Thomas, M., Wood, G., Wang, S., Walkinshaw, M.D., and Fischer, P.M., *Chem. Biol.* 13(2), 201-211 (**2006**)

170. Heady, L., Fernandez-Serra, M., Mancera, R.L., Joyce, S., Venkitaraman, A.R., Artacho, E., Skylaris, C-K., Colombi Giacchi, L., and Payne, M.C. *J. Med. Chem.* 49, 5141 (**2006**)

UNIVERSITY OF SOUTHAMPTON

Development and Application of a QM/MM Method for Free Energy Calculations

SUPPORTING INFORMATION 1

Small Molecule Hydration Free Energy Results

Michael Keith Carter

School of Chemistry
University of Southampton

September 2013

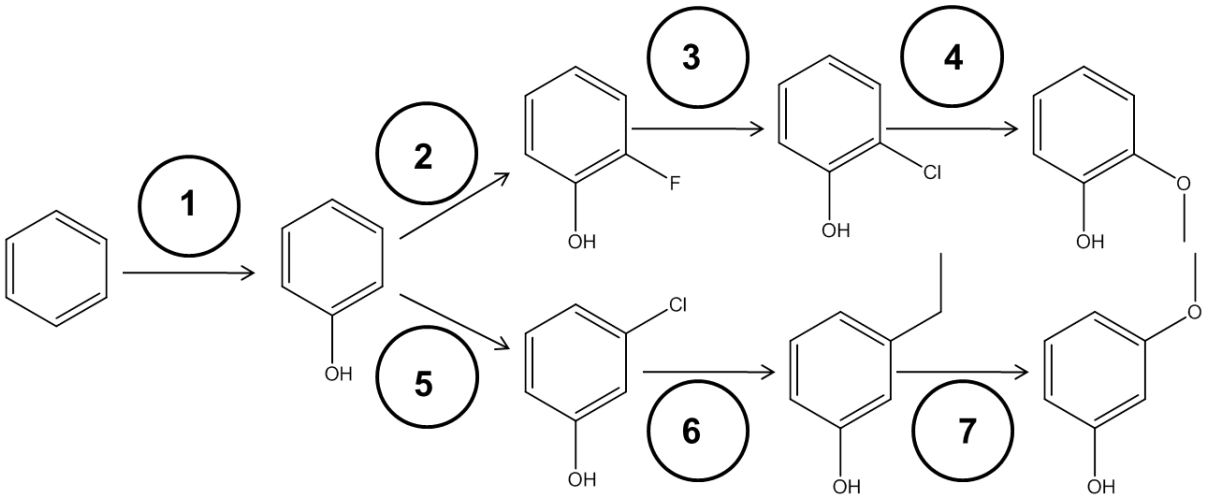


Figure 1.1: Phenol based perturbations

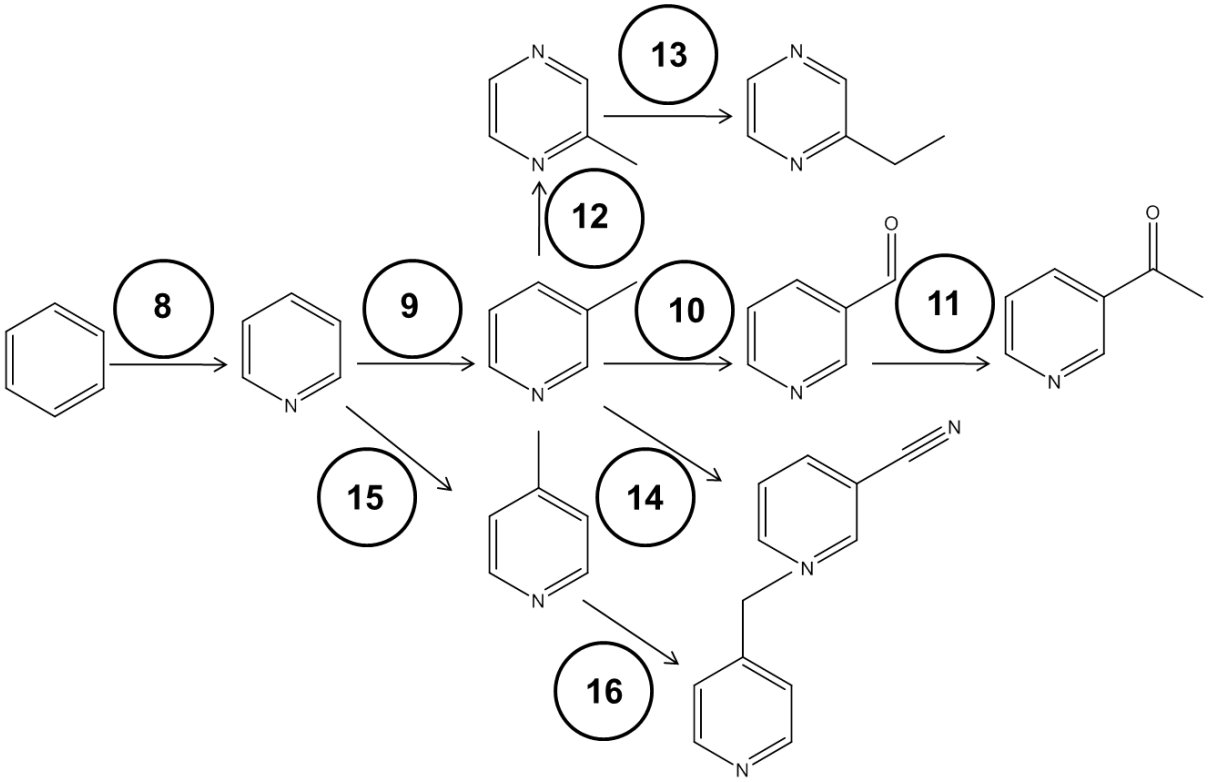


Figure 1.2: Pyridine based perturbations

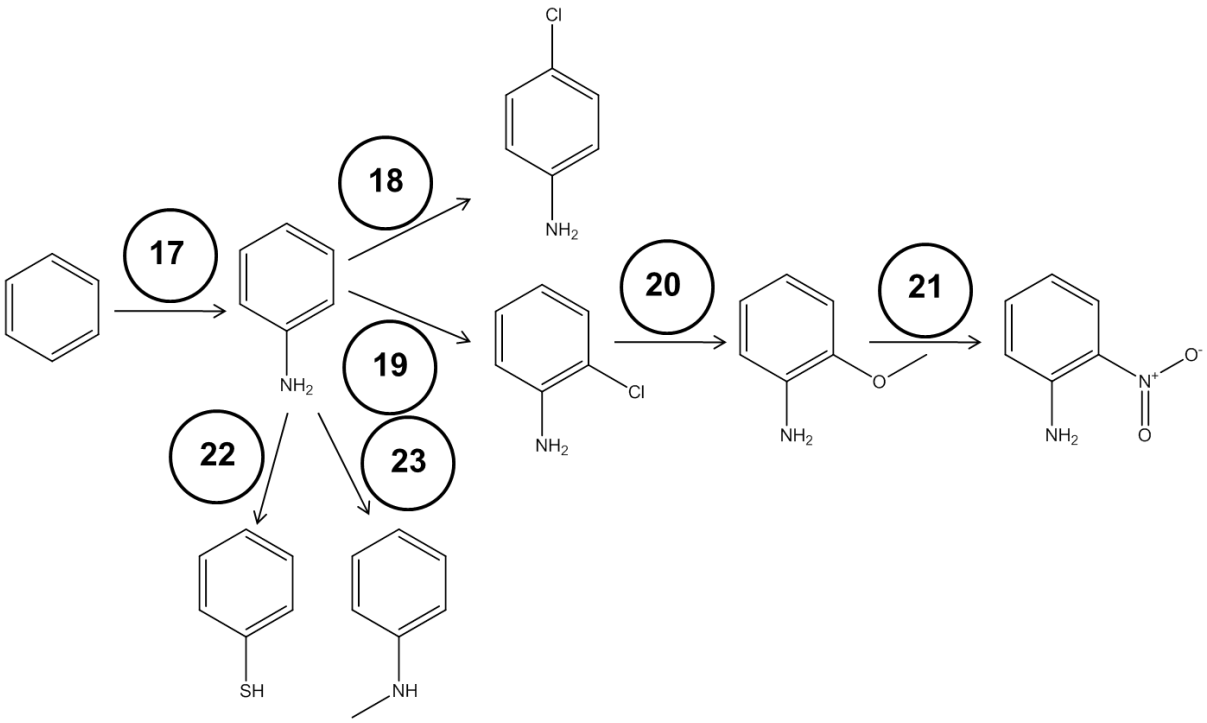


Figure 1.3: Aniline based perturbations

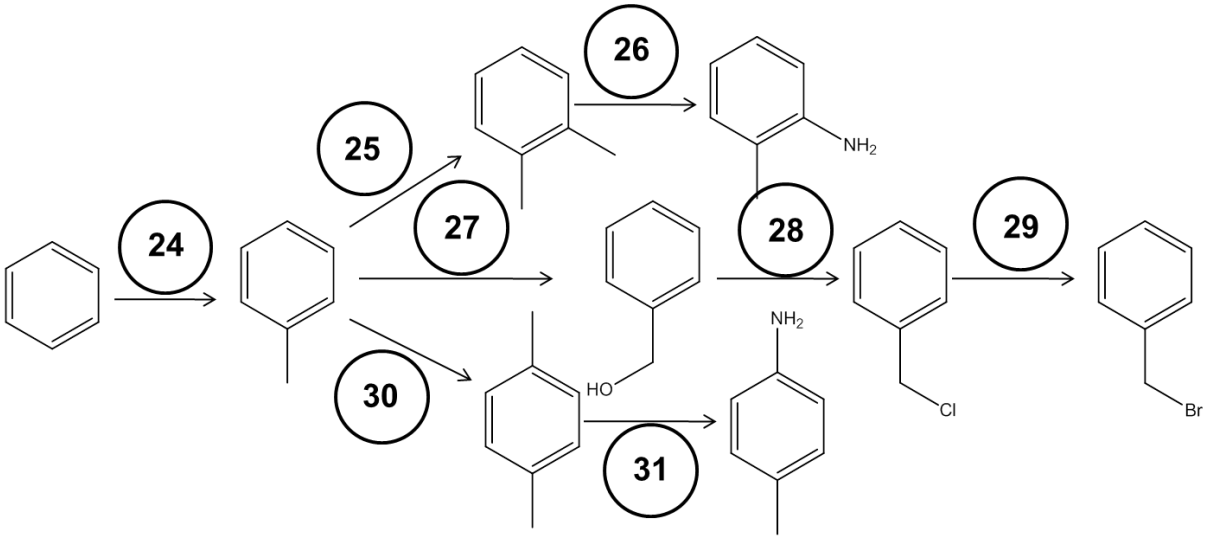


Figure 1.4: Toluene based perturbations

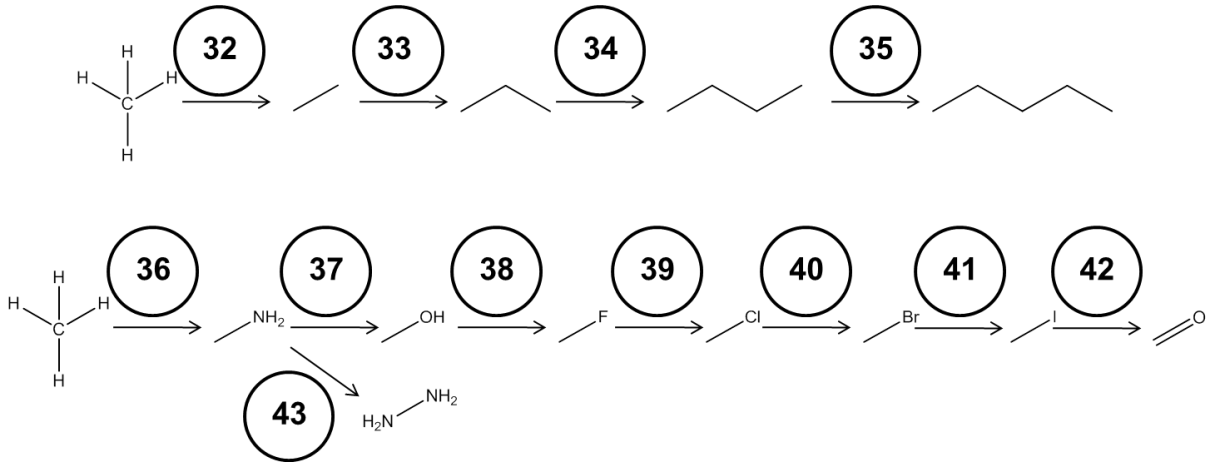


Figure 1.5: Methane based perturbations

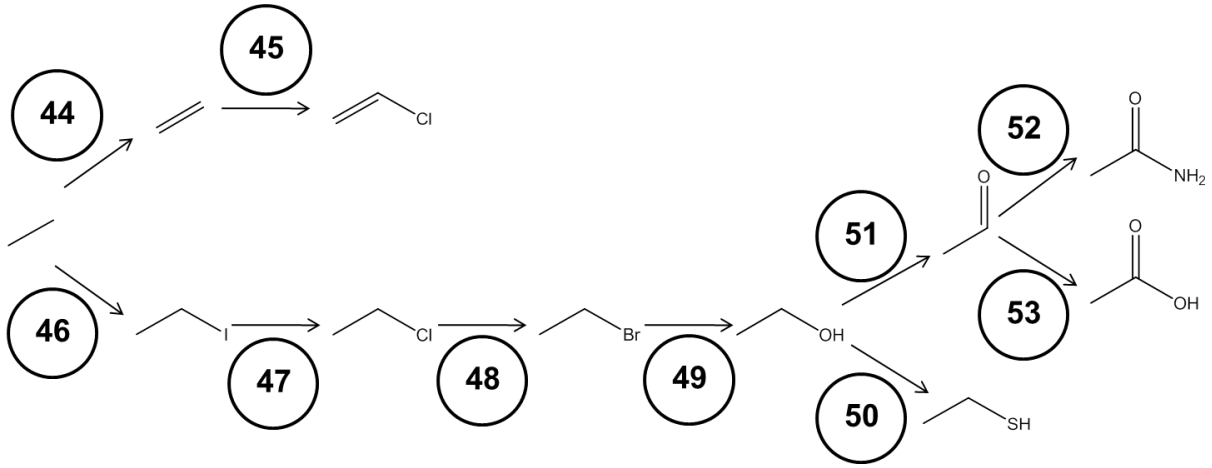


Figure 1.6: Ethane based perturbations

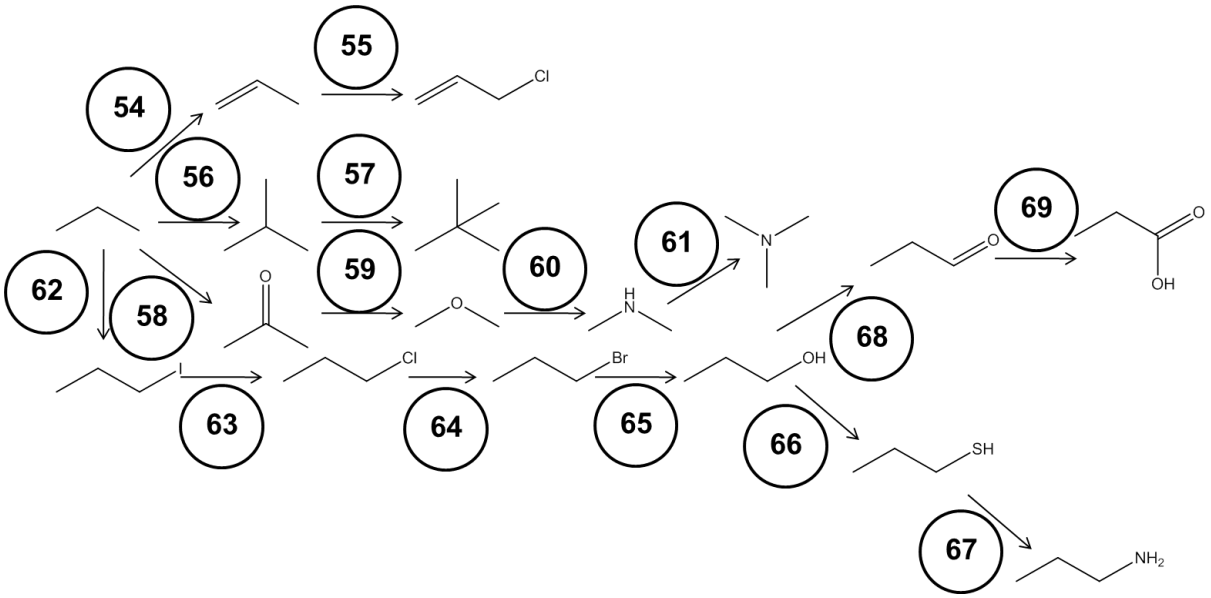


Figure 1.7: Propane based perturbations

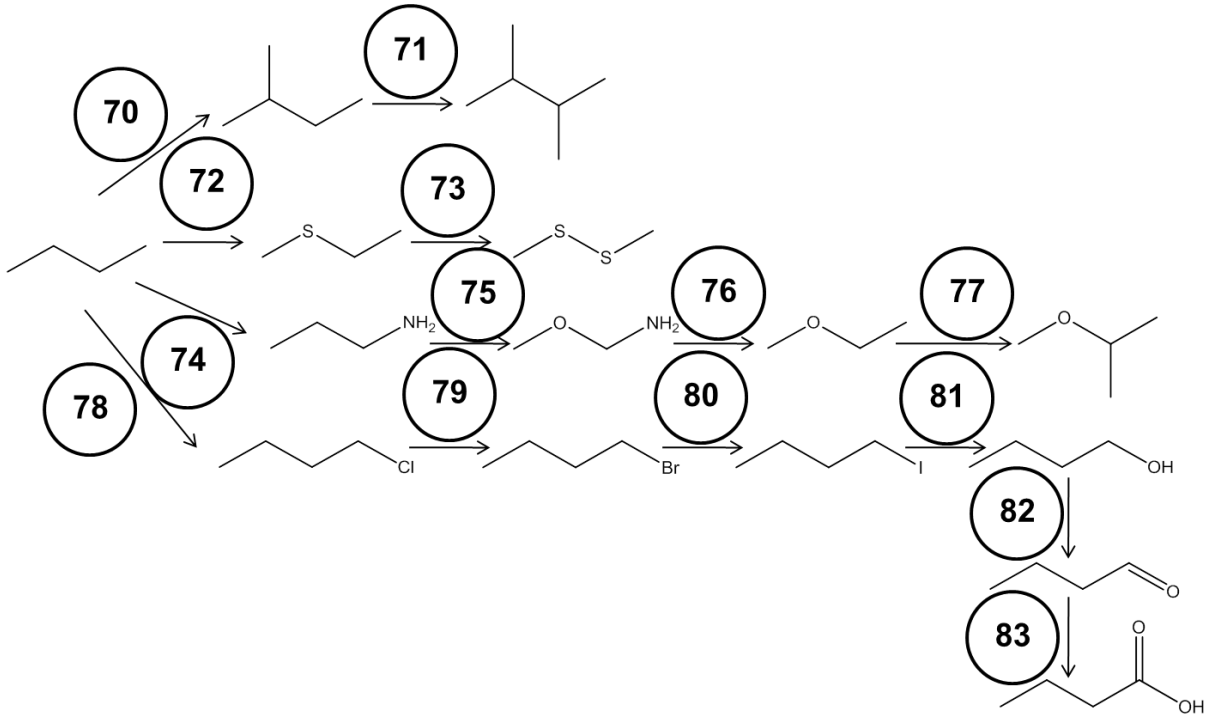


Figure 1.8: Butane based perturbations

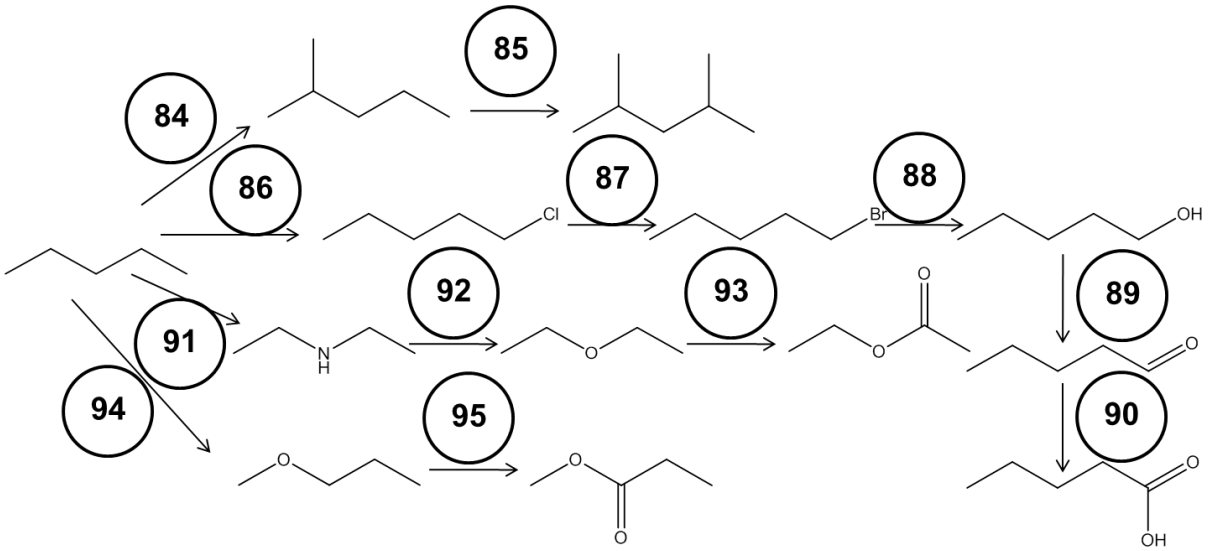


Figure 1.9: Pentane based perturbations

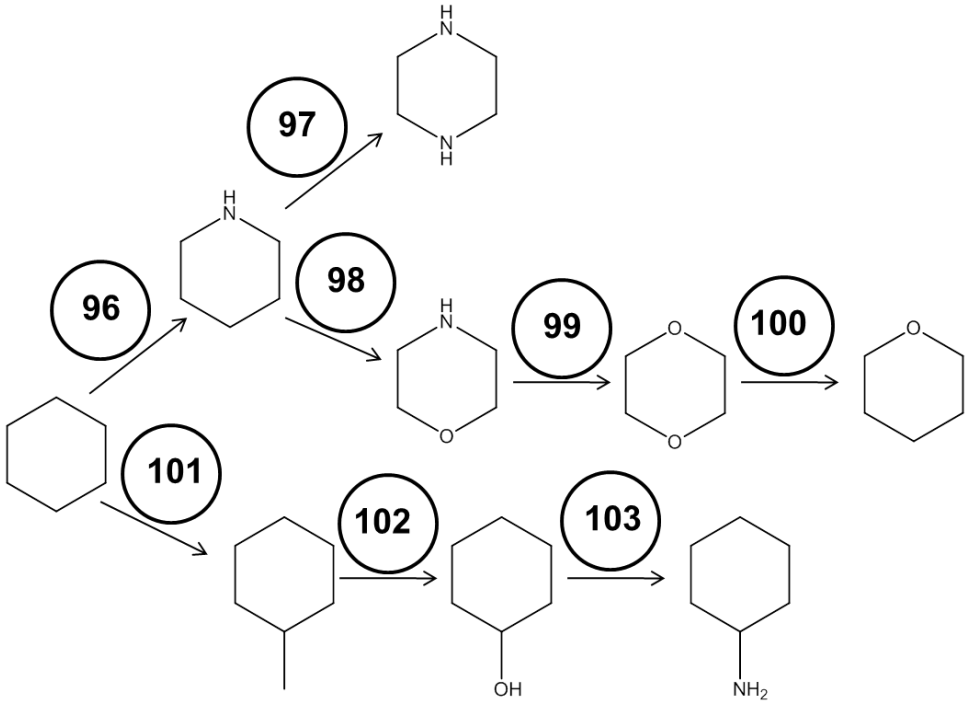


Figure 1.10: Cyclohexane based perturbations

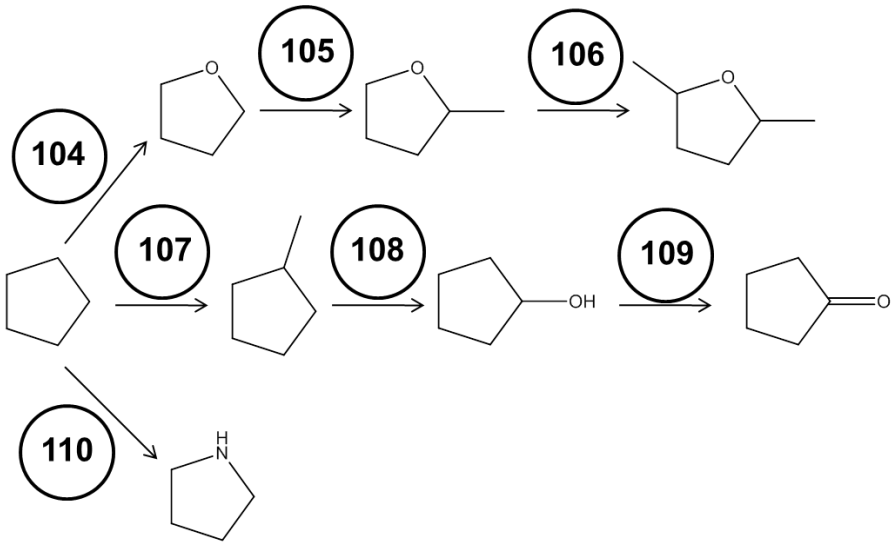


Figure 1.11: Cyclopentane based perturbations

Small Molecule Study – MM-RETI Results

Pert	$\Delta\Delta G_{hyd}$ (kcal.mol^{-1})	Error	Exp (Guthrie) (kcal.mol^{-1})	Exp (Mobley) (kcal.mol^{-1})
1	-4.63	0.12	-5.75	-5.76
2	2.15	0.16	1.32	2.43
3	0.54	0.18	0.74	0.01
4	-1.81	0.21	-1.27	-1.36
5	0.54	0.18	-0.01	0.39
6	-0.13	0.22	0.37	0.26
7	-2.14	0.28	-1.41	-2.01
8	-2.79	0.19	-1.85	-1.91
9	-0.10	0.09	0.09	-0.26

10	-4.12	0.28	-1.16	-4.45
11	-3.27	0.18	-1.16	-2.41
12	-2.53	0.14	-2.95	-3.24
13	0.44	0.13	0.06	0.42
14	-1.80	0.15	-4.04	-3.96
15	-0.09	0.06	-0.08	0.26
16	0.30	0.15	0.15	0.32
17	-4.75	0.23	-4.63	-5.22
18	0.41	0.11	-0.33	0.55
19	0.88	0.11	0.58	0.96
20	-1.81	0.19	-1.21	-1.57
21	-1.58	0.14	-1.47	-1.72
22	4.33	0.18	2.94	4.49
23	0.16	0.15	0.80	0.18
24	0.19	0.09	0.03	-0.01
25	-0.42	0.14	-0.01	-0.19
26	-3.26	0.19	-4.63	-4.81
27	-4.90	0.12	-5.81	-5.24
28	4.18	0.16	4.69	3.85
29	0.36	0.13	-0.45	0.42

30	0.13	0.07	0.09	0.04
31	-4.74	0.25	-4.77	-4.85
32	-0.12	0.04	-0.15	-0.09
33	-0.11	0.05	0.07	0.10
34	0.02	0.05	0.11	0.07
35	0.04	0.06	0.27	0.14
36	-6.59	0.16	-6.54	-6.21
37	-0.54	0.12	-1.97	-1.24
38	4.46	0.09	5.31	4.92
39	-0.13	0.10	-0.85	-0.24
40	-0.24	0.09	-0.71	-0.69
41	-0.33	0.14	-0.92	-0.74
42	-4.11	0.14	-5.66	-5.02
43	-4.14	0.16	-6.28	-4.74
44	-0.13	0.07	-0.60	-0.21
45	-1.20	0.10	-1.23	-0.92
46	-1.43	0.11	-2.82	-1.78
47	-0.31	0.09	0.56	0.24
48	-0.59	0.11	-0.44	-0.28
49	-2.82	0.08	-4.15	-3.31

50	3.40	0.14	4.35	3.92
51	-0.10	0.12	1.50	0.06
52	-6.49	0.21	-7.07	-6.96
53	-2.37	0.14	-3.19	-2.56
54	-0.17	0.11	-0.64	-0.16
55	-1.40	0.11	-1.89	-1.56
56	0.01	0.06	0.54	0.41
57	0.05	0.04	0.72	0.64
58	-6.52	0.14	-4.68	-5.96
59	2.64	0.17	1.25	1.94
60	-2.48	0.14	-1.94	-2.11
61	0.41	0.09	0.32	0.62
62	-0.72	0.12	-2.65	-1.34
63	-0.88	0.11	0.22	0.34
64	0.30	0.17	-0.55	-0.14
65	-4.36	0.11	-4.06	-3.98
66	3.16	0.15	3.92	3.44
67	-2.98	0.19	-3.42	-3.77
68	0.09	0.14	1.42	0.04
69	-3.84	0.12	-1.89	-3.33

70	0.11	0.04	0.34	0.18
71	0.05	0.03	0.00	0.21
72	-2.43	0.21	-3.55	-2.74
73	-0.18	0.13	0.41	0.72
74	-2.41	0.18	-3.11	-3.34
75	-1.83	0.11	-2.66	-3.13
76	4.83	0.19	4.45	4.80
77	0.01	0.09	0.09	0.07
78	-1.63	0.09	-2.28	-1.93
79	0.25	0.06	-0.45	-0.17
80	-0.48	0.17	0.26	-0.11
81	-2.91	0.18	-4.12	-3.24
82	0.92	0.12	1.54	0.18
83	-3.67	0.14	-3.13	-2.01
84	0.16	0.04	-0.23	-0.09
85	-0.11	0.09	-1.29	-0.66
86	-1.44	0.17	-2.35	-1.74
87	0.32	0.12	-0.09	-0.24
88	-4.40	0.19	-3.98	-4.46
89	0.01	0.12	1.82	0.16

90	-3.09	0.17	-4.16	-2.82
91	-6.19	0.21	-6.38	-6.44
92	2.61	0.18	2.16	2.31
93	-2.97	0.22	-0.94	-1.52
94	-2.41	0.23	-0.78	-1.91
95	-2.11	0.18	-0.94	-1.52
96	-6.31	0.21	-6.34	-5.13
97	-5.04	0.17	-2.80	-2.94
98	-2.75	0.24	-2.06	-2.82
99	2.48	0.21	2.11	2.57
100	2.41	0.18	1.94	2.57
101	0.20	0.10	0.47	0.15
102	-6.12	0.11	-7.16	-6.08
103	-6.89	0.14	-7.71	-6.93
104	-3.77	0.10	-4.68	-4.12
105	0.06	0.09	0.19	0.31
106	0.10	0.04	0.37	0.24
107	0.26	0.06	0.40	0.33
108	-6.43	0.14	-6.31	-5.98
109	0.49	0.11	0.79	0.44

110	-5.97	0.18	-6.67	-6.41

Table 1.1: MM-RETI relative hydration free energy results. $\Delta\Delta G_{hyd}$ is the free energy difference between the two end states of each perturbation. Two experimental datasets (Guthrie and Mobley) are also reported. The errors shown were calculated from four independent simulations using standard error.

Small Molecule Study – MM→QM/MM-FEP Results

Pert	ΔG_1 (kcal.mol^{-1})	Error	ΔG_2 (kcal.mol^{-1})	Error	ΔG_3 (kcal.mol^{-1})	Error	ΔΔGhyd MM→QM/MM (kcal.mol^{-1})	Error	Exp (Guthrie) (kcal.mol^{-1})
1	0.29	0.09	-4.63	0.12	-1.23	0.19	-5.56	0.23	-5.75
2	-1.27	0.18	2.15	0.16	-1.52	0.22	1.16	0.28	1.32
3	-1.36	0.11	0.54	0.18	-0.77	0.09	0.6	0.16	0.74
4	-0.86	0.09	-1.81	0.21	-1.96	0.19	-2.91	0.31	-1.27
5	-1.29	0.22	0.54	0.18	-1.74	0.28	0.27	0.34	-0.01
6	-1.54	0.18	-0.13	0.22	-1.63	0.11	-0.22	0.28	0.37
7	-1.60	0.27	-2.14	0.28	-2.41	0.21	-2.86	0.35	-1.41
8	0.31	0.06	-2.79	0.19	-1.59	0.16	-4.69	0.20	-1.85
9	-1.57	0.13	-0.10	0.09	-2.07	0.14	-0.60	0.17	0.09
10	-1.73	0.24	-4.12	0.28	-1.89	0.18	-4.28	0.21	-1.16
11	-1.85	0.22	-3.27	0.18	-2.36	0.23	-0.15	0.25	-1.16
12	-1.48	0.20	-2.53	0.14	-1.29	0.21	-2.53	0.18	-2.95
13	-1.25	0.13	0.44	0.13	-0.91	0.16	0.78	0.11	0.06
14	-1.82	0.17	-1.80	0.15	-2.82	0.22	-2.80	0.16	-4.04

15	-1.62	0.16	-0.09	0.06	-1.73	0.13	-0.20	0.10	-0.08
16	-1.98	0.25	0.30	0.15	-2.11	0.24	0.17	0.19	0.15
17	0.28	0.03	-4.75	0.23	-0.74	0.08	-5.77	0.09	-4.63
18	-0.73	0.11	0.41	0.11	-1.49	0.21	-0.35	0.17	-0.33
19	-0.68	0.14	0.88	0.11	-0.69	0.07	0.87	0.09	0.58
20	-0.58	0.13	-1.81	0.19	-1.52	0.19	-2.75	0.14	-1.21
21	-1.61	0.17	-1.58	0.14	-1.71	0.22	-1.68	0.16	-1.47
22	-0.93	0.18	4.33	0.18	-1.88	0.28	3.64	0.23	2.94
23	-0.86	0.12	0.16	0.15	-0.50	0.08	0.52	0.09	0.80
24	0.30	0.07	0.19	0.09	0.31	0.06	0.20	0.05	0.03
25	0.29	0.04	-0.42	0.14	0.25	0.03	-0.46	0.04	-0.01
26	0.27	0.07	-3.26	0.19	-0.80	0.17	-4.33	0.11	-4.63
27	0.16	0.02	-4.90	0.12	-1.63	0.26	-6.69	0.14	-5.81
28	-1.52	0.27	4.18	0.16	0.58	0.07	5.12	0.17	4.69
29	0.52	0.10	0.36	0.13	0.64	0.11	0.48	0.10	-0.45
30	0.28	0.07	0.13	0.07	0.17	0.06	0.02	0.04	0.09
31	0.18	0.05	-4.74	0.25	-0.89	0.14	-5.81	0.17	-4.77
32	-0.20	0.04	-0.12	0.04	-0.29	0.07	-0.21	0.04	-0.15
33	-0.29	0.09	-0.11	0.05	-0.39	0.08	-0.21	0.06	0.07
34	-0.39	0.11	0.02	0.05	-0.48	0.05	-0.07	0.06	0.11
35	-0.47	0.12	0.04	0.06	-0.57	0.15	-0.06	0.09	0.27
36	-0.18	0.06	-6.59	0.16	-1.20	0.24	-7.61	0.14	-6.54
37	-1.09	0.19	-0.54	0.12	-1.43	0.21	-0.88	0.16	-1.97

38	-1.65	0.27	4.46	0.09	-0.06	0.07	6.05	0.11	5.31
39	-0.12	0.02	-0.13	0.10	-0.33	0.12	-0.34	0.08	-0.85
40	-0.33	0.09	-0.24	0.09	-0.72	0.13	-0.63	0.08	-0.71
41	-0.69	0.16	-0.33	0.14	-0.88	0.09	-0.53	0.11	-0.92
42	-0.91	0.19	-4.11	0.14	-1.23	0.23	-4.43	0.17	-5.66
43	-1.24	0.21	-4.14	0.16	-2.64	0.27	-5.54	0.21	-6.28
44	-0.27	0.04	-0.13	0.07	-1.14	0.17	-1.00	0.07	-0.60
45	-1.03	0.18	-1.20	0.10	-1.16	0.13	-1.33	0.11	-1.23
46	-0.26	0.01	-1.43	0.11	-0.89	0.11	-2.06	0.07	-2.82
47	-0.93	0.19	-0.31	0.09	-0.62	0.13	0.00	0.09	0.56
48	-0.71	0.21	-0.59	0.11	-1.12	0.24	-1.00	0.21	-0.44
49	-1.09	0.28	-2.82	0.08	-1.31	0.19	-3.04	0.18	-4.15
50	-1.34	0.18	3.40	0.14	-1.54	0.14	3.20	0.15	4.35
51	-1.46	0.15	-0.10	0.12	0.07	0.09	1.43	0.11	1.50
52	0.09	0.01	-6.49	0.21	-2.28	0.21	-8.86	0.16	-7.07
53	0.06	0.04	-2.37	0.14	-1.41	0.23	-3.84	0.13	-3.19
54	-0.33	0.11	-0.17	0.11	-0.98	0.21	-0.82	0.17	-0.64
55	-1.13	0.22	-1.40	0.11	-1.50	0.13	-1.77	0.14	-1.89
56	-0.36	0.05	0.01	0.06	-0.42	0.14	-0.05	0.07	0.54
57	-0.45	0.14	0.05	0.04	-0.33	0.09	0.17	0.08	0.72
58	-0.36	0.11	-6.52	0.14	-0.83	0.10	-6.99	0.10	-4.68
59	-0.87	0.18	2.64	0.17	-1.39	0.13	2.12	0.14	1.25
60	-1.41	0.24	-2.48	0.14	-1.77	0.25	-2.87	0.23	-1.94

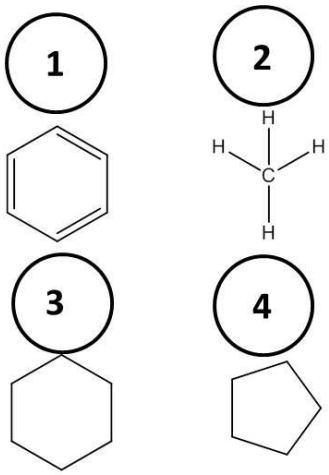
61	-1.69	0.29	0.41	0.09	-0.97	0.13	1.13	0.15	0.32
62	-0.32	0.06	-0.72	0.12	-1.17	0.15	-1.57	0.09	-2.65
63	-1.21	0.21	-0.88	0.11	-0.69	0.13	-0.36	0.14	0.22
64	-0.72	0.08	0.30	0.17	-0.51	0.05	0.51	0.07	-0.55
65	-0.55	0.13	-4.36	0.11	-1.12	0.21	-4.93	0.15	-4.06
66	-1.15	0.25	3.16	0.15	-1.67	0.28	2.64	0.24	3.92
67	-1.72	0.29	-2.98	0.19	-1.91	0.27	-3.17	0.26	-3.42
68	-1.68	0.23	0.09	0.14	0.13	0.05	1.90	0.11	1.42
69	0.15	0.04	-3.84	0.12	-0.98	0.09	-4.97	0.07	-1.89
70	-0.45	0.14	0.11	0.04	-0.58	0.08	-0.02	0.06	0.34
71	-0.61	0.13	0.05	0.03	-0.73	0.09	-0.07	0.05	0.00
72	-0.52	0.11	-2.43	0.21	-1.21	0.22	-3.12	0.17	-3.55
73	-1.23	0.23	-0.18	0.13	-2.43	0.27	-1.38	0.24	0.41
74	-0.50	0.08	-2.41	0.18	-2.10	0.25	-4.01	0.16	-3.11
75	-2.12	0.23	-1.83	0.11	-3.28	0.31	-2.99	0.29	-2.66
76	-3.28	0.29	4.83	0.19	-1.29	0.19	6.82	0.19	4.45
77	-1.35	0.17	0.01	0.09	-1.18	0.13	0.18	0.10	0.09
78	-0.51	0.16	-1.63	0.09	-0.85	0.10	-1.97	0.09	-2.28
79	-0.91	0.19	0.25	0.06	-1.17	0.17	-0.01	0.11	-0.45
80	-1.21	0.27	-0.48	0.17	-1.52	0.25	-0.79	0.24	0.26
81	-1.52	0.28	-2.91	0.18	-1.81	0.29	-3.20	0.27	-4.12
82	-1.85	0.22	0.92	0.12	-0.22	0.04	2.55	0.10	1.54
83	-0.25	0.04	-3.67	0.14	-2.12	0.23	-5.54	0.08	-3.13

84	-0.58	0.09	0.16	0.04	-0.57	0.08	0.17	0.05	-0.23
85	-0.67	0.11	-0.11	0.09	-0.77	0.14	-0.21	0.09	-1.29
86	-0.51	0.07	-1.44	0.17	-1.56	0.27	-2.49	0.15	-2.35
87	-1.58	0.13	0.32	0.12	-1.67	0.33	0.23	0.26	-0.09
88	-1.71	0.27	-4.40	0.19	-1.81	0.37	-4.50	0.34	-3.98
89	-1.83	0.23	0.01	0.12	0.49	0.11	2.33	0.13	1.82
90	0.51	0.07	-3.09	0.17	-2.21	0.27	-5.81	0.14	-4.16
91	-0.57	0.21	-6.19	0.21	-1.57	0.18	-7.19	0.17	-6.38
92	-1.59	0.28	2.61	0.18	-2.03	0.33	2.17	0.25	2.16
93	-2.05	0.21	-2.97	0.22	-1.03	0.15	-1.95	0.19	-0.94
94	-0.56	0.08	-2.41	0.23	-1.21	0.18	-3.06	0.13	-0.78
95	-1.21	0.17	-2.11	0.18	-1.71	0.21	-2.61	0.18	-0.94
96	-0.57	0.11	-6.31	0.21	-1.95	0.22	-7.69	0.16	-6.34
97	-1.90	0.22	-5.04	0.17	-4.09	0.34	-7.23	0.21	-2.80
98	-1.87	0.19	-2.75	0.24	-3.94	0.33	-4.80	0.26	-2.06
99	-3.76	0.26	2.48	0.21	-3.68	0.31	2.56	0.29	2.11
100	-3.54	0.28	2.41	0.18	-2.10	0.15	3.85	0.14	1.94
101	-0.57	0.05	0.20	0.10	-0.72	0.08	0.05	0.06	0.47
102	-0.68	0.13	-6.12	0.11	-2.27	0.23	-7.71	0.11	-7.16
103	-2.22	0.24	-6.89	0.14	-2.94	0.27	-7.61	0.18	-7.71
104	-0.55	0.06	-3.77	0.10	-2.27	0.32	-5.98	0.13	-4.68
105	-2.21	0.25	0.06	0.09	-2.10	0.23	0.17	0.11	0.19
106	-2.11	0.24	0.10	0.04	-2.28	0.27	-0.07	0.16	0.37

107	-0.59	0.04	0.26	0.06	-0.66	0.14	0.19	0.09	0.40
108	-0.66	0.10	-6.43	0.14	-1.97	0.19	-7.74	0.11	-6.31
109	-1.92	0.21	0.49	0.11	-1.77	0.17	0.64	0.14	0.79
110	-0.55	0.08	-5.97	0.18	-2.19	0.16	-7.74	0.12	-6.67

Table 1.2: MM→QM/MM relative hydration free energy results. ΔG_1 is the QM/MM correction for λ=0 for each perturbation. ΔG_3 is the QM/MM correction for λ=1 for each perturbation. ΔG_2 is the MM relative free energy change between the two endstates of each perturbation. $\Delta\Delta G_{hyd}$ is the MM→QM/MM free energy difference between the two end states of each perturbation. One experimental dataset (Guthrie) is also reported. The errors shown were calculated from four independent simulations using standard error.

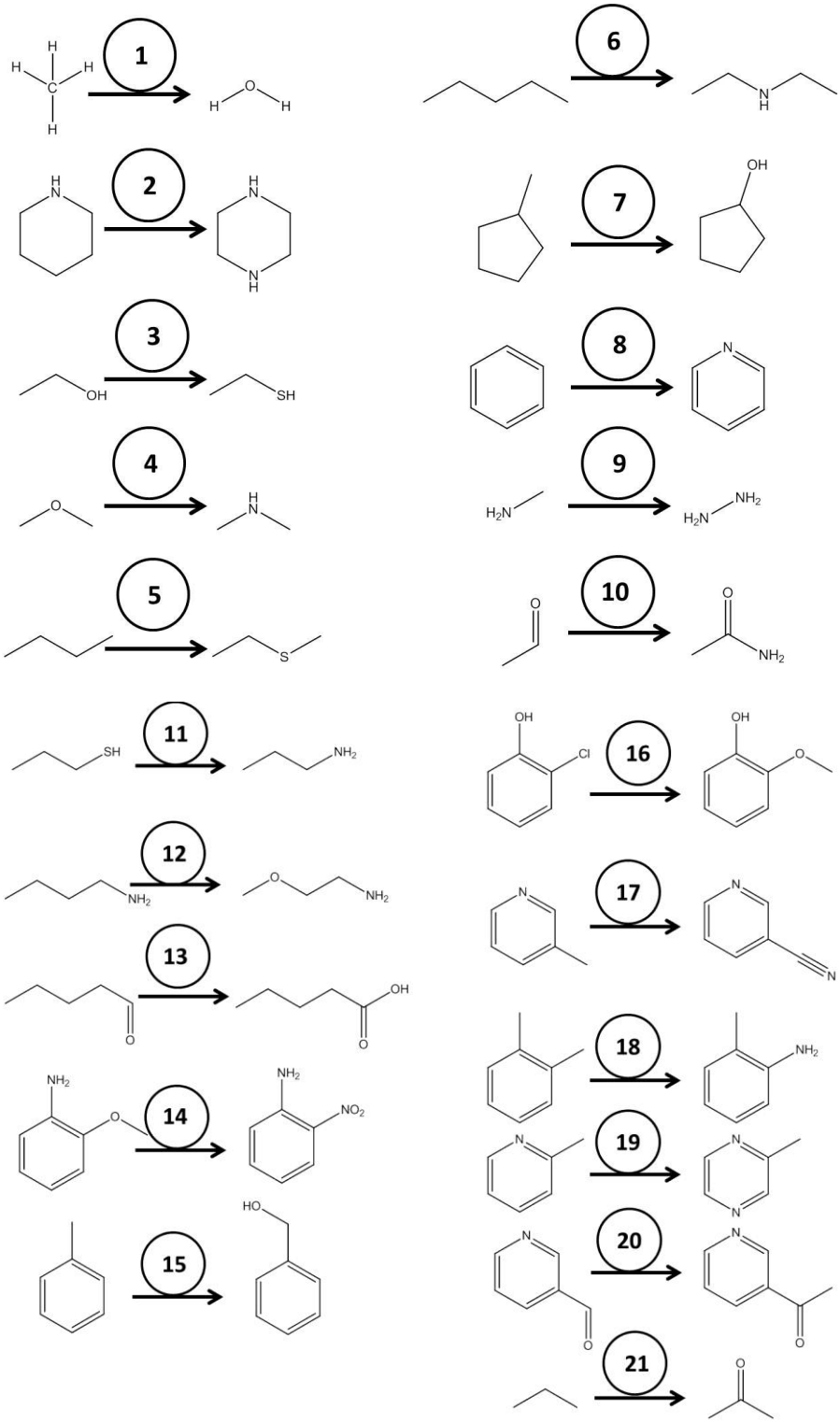
Small Molecule Study – Dual Topology Annihilations – MM-RETI Results

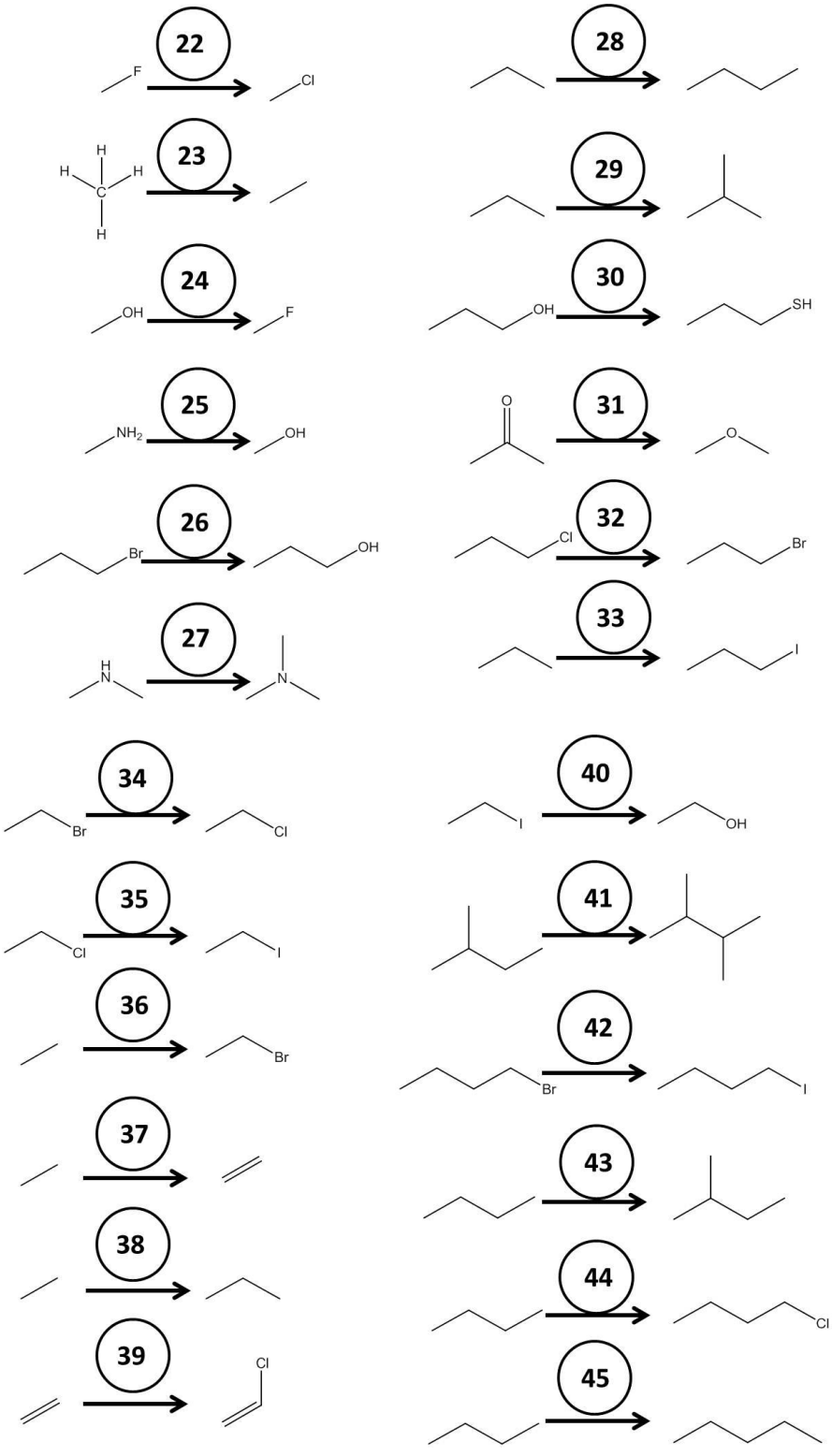


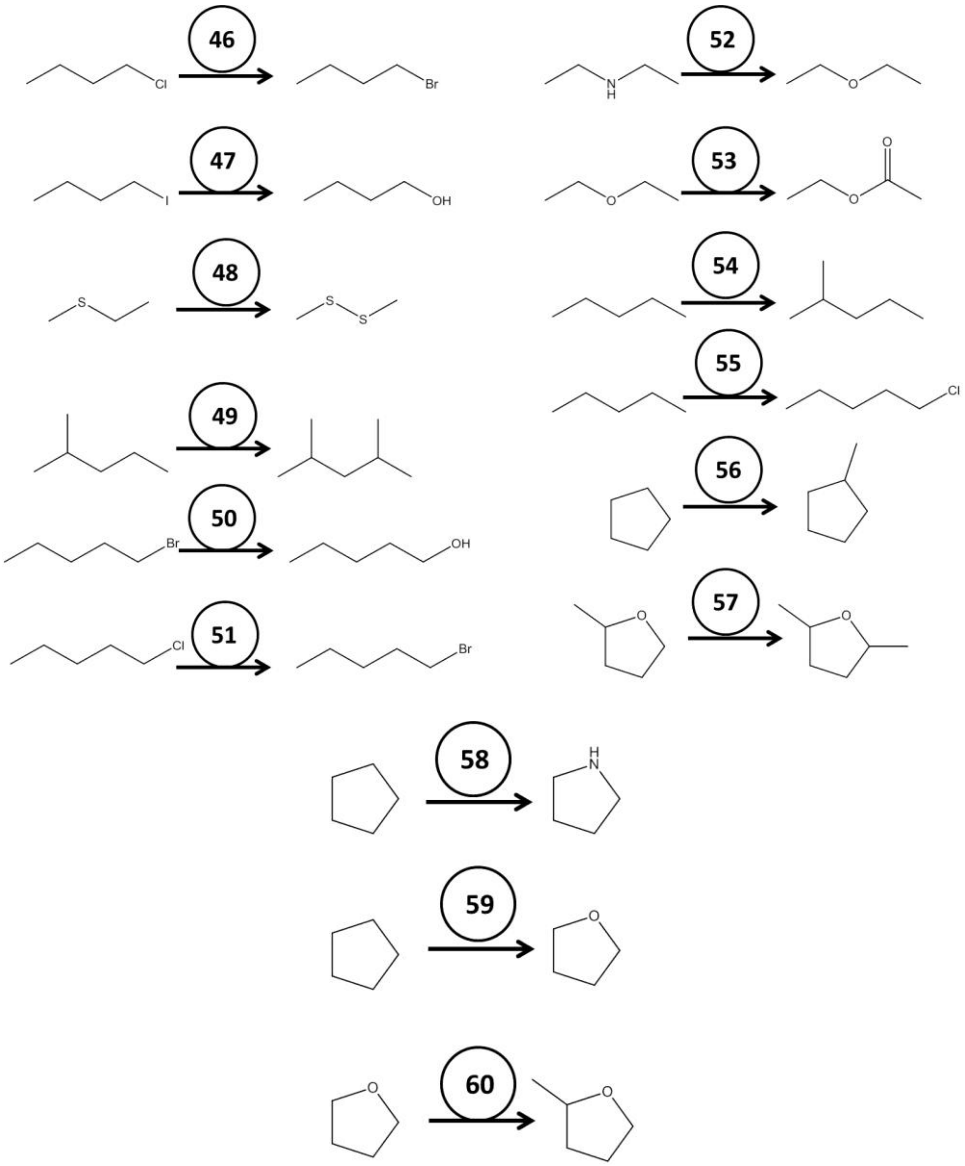
Pert	δ	σ	$\Delta\Delta G_{hyd}$ (kcal.mol^{-1})	Error	Exp (Guthrie) (kcal.mol^{-1})
1	1.5	1.0	-0.74	0.25	-0.86
2	1.0	2.0	1.45	0.29	1.23
3	1.5	1.0	1.83	0.21	1.20
4	1.5	1.5	2.17	0.32	1.99

Table 1.3: MM-RETI absolute hydration free energies for our perturbation web starting points. δ, and σ are the soft-core parameters utilised for each annihilation. $\Delta\Delta G_{hyd}$ is the free energy difference between the two end states of each perturbation. One experimental dataset (Guthrie) is also reported. The errors shown were calculated from four independent simulations using standard error.

Gaussian Blurring Dataset




22
F
Cl
23
H
H
H
C
H
24
OH
F
25
NH2
OH
26
Br
OH
27
H
N
N
28
29
30
OH
SH
31
O
O
32
Cl
Br
33
I
34
Br
Cl
35
Cl
I
36
Br
37
38
39
Cl
40
I
OH
41
42
Br
I
43
44
Cl
45


46
Cl
Br
47
I
OH
48
S
S
S
49
50
Br
OH
51
Cl
Br
52
N
H
O
53
O
O
O
54
55
Cl
56
57
O
O
58
H
N
59
O
60
O
O

Small Molecule Study – MM→QM/MM – 'Gaussian Blur' = 1.0

Pert	ΔG_1 (kcal.mol^{-1})	Error	ΔG_2 (kcal.mol^{-1})	Error	ΔG_3 (kcal.mol^{-1})	Error	$\Delta\Delta G_{hyd}$ (MM→QM/MM) 'Gaussian Blur' 1.0 (kcal.mol^{-1})	Error	Exp (Guthrie) (kcal.mol^{-1})
1	-0.03	0.15	-8.88	0.04	0.70	0.35	-9.17	0.28	-8.32
2	1.89	0.21	-5.04	0.08	2.76	0.38	-4.17	0.32	-2.80
3	0.02	0.13	3.4	0.12	0.66	0.19	4.04	0.18	4.35
4	0.07	0.09	-2.48	0.03	2.14	0.34	-0.41	0.27	-1.94
5	0.03	0.10	-2.43	0.05	0.04	0.06	-2.42	0.05	-3.55
6	0.01	0.05	-6.19	0.09	2.50	0.26	-3.70	0.19	-6.38
7	-0.06	0.03	-6.43	0.15	3.78	0.33	-2.59	0.19	-6.31
8	2.15	0.37	-2.79	0.11	1.65	0.22	-3.29	0.21	-1.85
9	1.01	0.21	-4.14	0.05	2.72	0.33	-2.43	0.29	-6.28
10	2.28	0.28	-6.49	0.06	2.54	0.38	-6.23	0.22	-7.07
11	0.49	0.19	-2.98	0.08	0.42	0.13	-3.05	0.11	-3.42
12	0.82	0.23	-1.83	0.07	0.41	0.18	-2.24	0.11	-2.66
13	2.21	0.28	-3.09	0.03	4.39	0.37	-0.91	0.27	-4.16
14	0.00	0.01	-1.58	0.02	0.50	0.13	-1.08	0.09	-1.47
15	2.02	0.29	-4.9	0.01	2.69	0.35	-5.57	0.25	-5.81
16	1.58	0.33	-1.81	0.13	2.46	0.41	-0.93	0.32	-1.27
17	3.21	0.31	-1.8	0.15	2.89	0.28	-2.12	0.22	-4.04
18	1.76	0.25	-3.26	0.17	2.03	0.31	-2.99	0.26	-4.63
19	2.42	0.22	-2.53	0.09	3.53	0.36	-1.42	0.27	-2.95

20	3.21	0.31	0.36	0.21	2.30	0.18	-0.55	0.22	-1.16
21	-0.04	0.29	-6.52	0.21	1.62	0.31	-4.86	0.26	-4.68
22	0.23	0.17	-0.13	0.01	-0.24	0.14	-0.60	0.09	-0.85
23	0.00	0.05	-0.12	0.02	-0.04	0.05	-0.16	0.03	-0.15
24	-0.03	0.09	4.46	0.02	0.26	0.19	4.75	0.10	5.31
25	0.95	0.22	0.54	0.04	-0.01	0.05	-0.42	0.09	-1.97
26	-0.56	0.27	-4.36	0.05	-0.08	0.08	-3.88	0.11	-4.06
27	2.36	0.11	0.41	0.03	2.64	0.38	0.69	0.17	0.32
28	-0.03	0.15	0.02	0.01	-0.03	0.08	0.02	0.09	0.11
29	-0.02	0.09	0.01	0.01	-0.03	0.09	0.00	0.07	0.54
30	-0.05	0.12	3.16	0.03	0.64	0.22	3.85	0.13	3.92
31	1.57	0.27	2.64	0.09	0.22	0.12	1.29	0.14	1.25
32	-0.22	0.29	0.3	0.14	-0.39	0.29	0.13	0.21	-0.55
33	-0.02	0.10	-1.82	0.18	-0.78	0.17	-2.62	0.13	-2.65
34	-0.29	0.18	-0.31	0.02	-0.13	0.09	-0.15	0.10	0.56
35	-0.27	0.16	-0.59	0.05	1.11	0.26	0.79	0.15	-0.44
36	0.00	0.06	-1.43	0.04	-0.19	0.09	-1.62	0.07	-2.82
37	-0.02	0.08	-0.13	0.01	-0.39	0.18	-0.50	0.09	-0.60
38	-0.02	0.09	-0.11	0.01	-0.03	0.06	-0.12	0.05	0.07
39	-0.30	0.21	-1.2	0.12	0.12	0.22	-0.78	0.18	-1.23
40	1.09	0.28	-2.82	0.15	-0.43	0.27	-4.34	0.19	-4.15
41	0.05	0.15	0.05	0.01	0.09	0.14	0.09	0.09	0.00
42	-0.21	0.18	-0.48	0.04	-0.11	0.19	-0.38	0.12	0.26

43	0.02	0.09	0.11	0.03	-0.10	0.09	-0.01	0.07	0.34
44	0.00	0.06	-1.63	0.11	2.63	0.36	-1.83	0.23	-2.28
45	-0.03	0.11	0.04	0.01	-0.07	0.15	0.00	0.09	0.27
46	2.56	0.21	0.25	0.02	-0.52	0.27	-2.83	0.16	-0.45
47	-0.31	0.12	-2.91	0.22	-0.21	0.19	-2.81	0.16	-4.12
48	-0.09	0.07	-0.18	0.01	-0.91	0.27	-1.00	0.15	0.41
49	-0.10	0.09	-0.11	0.01	-0.09	0.09	-0.10	0.06	-1.29
50	-0.60	0.23	-4.4	0.1	-0.02	0.03	-3.82	0.14	-3.98
51	-0.36	0.16	0.32	0.06	-0.46	0.22	0.22	0.13	-0.09
52	2.69	0.31	2.61	0.07	0.54	0.12	0.46	0.18	2.16
53	0.69	0.21	-2.97	0.04	2.12	0.31	-1.54	0.18	-0.19
54	-0.01	0.09	0.16	0.03	-0.05	0.04	0.12	0.07	-0.23
55	0.01	0.04	-1.44	0.11	-0.24	0.04	-1.69	0.08	-2.35
56	-0.10	0.17	0.26	0.11	0.00	0.10	0.36	0.12	0.40
57	0.05	0.09	0.1	0.05	0.18	0.12	0.23	0.09	0.37
58	-0.06	0.07	-5.97	0.18	1.71	0.27	-4.20	0.12	-6.67
59	-0.07	0.09	-3.77	0.16	-0.09	0.07	-3.79	0.09	-4.68
60	-0.08	0.10	0.06	0.01	0.24	0.16	0.38	0.08	0.19

Table 1.3: MM→QM/MM relative hydration free energy results. ΔG_1 is the QM/MM correction for λ=0 for each perturbation. ΔG_3 is the QM/MM correction for λ=1 for each perturbation. ΔG_2 is the MM relative free energy change between the two endstates of each perturbation. $\Delta\Delta G_{hyd}$ is the MM→QM/MM free energy difference between the two end states of each perturbation. One experimental dataset (Guthrie) is also reported. The errors shown were calculated from four independent simulations using standard error.

Small Molecule Study – MM→QM/MM – 'Gaussian Blur' = 0.95

Pert	ΔG_1 (kcal.mol^{-1})	Error	ΔG_2 (kcal.mol^{-1})	Error	ΔG_3 (kcal.mol^{-1})	Error	$\Delta\Delta G_{hyd}$ (MM→QM/MM) 'Gaussian Blur' 0.95 (kcal.mol^{-1})	Error	Exp (Guthrie) (kcal.mol^{-1})
1	0.01	0.15	-8.88	0.04	0.91	0.35	-7.98	0.28	-8.32
2	2.11	0.21	-5.04	0.08	3.17	0.38	-3.98	0.32	-2.8
3	0.16	0.13	3.4	0.12	0.74	0.19	3.98	0.18	4.35
4	0.19	0.09	-2.48	0.03	2.35	0.34	-0.32	0.27	-1.94
5	0.05	0.10	-2.43	0.05	0.12	0.06	-2.36	0.05	-3.55
6	0.04	0.05	-6.19	0.09	2.73	0.26	-3.5	0.19	-6.38
7	-0.03	0.03	-6.43	0.15	4.00	0.33	-2.4	0.19	-6.31
8	2.26	0.37	-2.79	0.11	1.82	0.22	-3.23	0.21	-1.85
9	1.22	0.21	-4.14	0.05	3.01	0.33	-2.35	0.29	-6.28
10	2.44	0.28	-6.49	0.06	2.83	0.38	-6.1	0.22	-7.07
11	0.59	0.19	-2.98	0.08	0.42	0.13	-3.15	0.11	-3.42
12	1.05	0.23	-1.83	0.07	0.73	0.18	-2.15	0.11	-2.66
13	2.39	0.28	-3.09	0.03	4.65	0.37	-0.83	0.27	-4.16
14	0.00	0.01	-1.58	0.02	0.52	0.13	-1.06	0.09	-1.47
15	2.12	0.29	-4.9	0.01	2.92	0.35	-5.7	0.25	-5.81
16	1.69	0.33	-1.81	0.13	2.55	0.41	-0.95	0.32	-1.27
17	3.49	0.31	-1.8	0.15	2.91	0.28	-2.38	0.22	-4.04
18	1.87	0.25	-3.26	0.17	2.16	0.31	-2.91	0.26	-4.63

19	2.62	0.22	-2.53	0.09	3.79	0.36	-1.36	0.27	-2.95
20	3.55	0.31	0.36	0.21	-2.68	0.18	-0.51	0.22	-1.16
21	-0.02	0.29	-6.52	0.21	1.80	0.31	-4.7	0.26	-4.68
22	0.26	0.17	-0.13	0.01	-0.21	0.14	-0.6	0.09	-0.85
23	0.02	0.05	-0.12	0.02	-0.02	0.05	-0.16	0.03	-0.15
24	0.09	0.09	4.46	0.02	0.28	0.19	4.65	0.10	5.31
25	1.16	0.22	0.54	0.04	0.12	0.05	-0.5	0.09	-1.97
26	-0.52	0.27	-4.36	0.05	0.05	0.08	-3.79	0.11	-4.06
27	2.58	0.11	0.41	0.03	2.62	0.38	0.45	0.17	0.32
28	-0.01	0.15	0.02	0.01	-0.01	0.08	0.02	0.09	0.11
29	-0.01	0.09	0.01	0.01	-0.01	0.09	0.01	0.07	0.54
30	0.09	0.12	3.16	0.03	0.72	0.22	3.79	0.13	3.92
31	1.74	0.27	2.64	0.09	0.33	0.12	1.23	0.14	1.25
32	-0.17	0.29	0.3	0.14	-0.37	0.29	0.1	0.21	-0.55
33	-0.01	0.10	-1.82	0.18	-0.77	0.17	-2.58	0.13	-2.65
34	-0.26	0.18	-0.31	0.02	-0.09	0.09	-0.14	0.10	0.56
35	-0.24	0.16	-0.59	0.05	1.11	0.26	0.76	0.15	-0.44
36	0.02	0.06	-1.43	0.04	-0.17	0.09	-1.62	0.07	-2.82
37	0.00	0.08	-0.13	0.01	-0.37	0.18	-0.5	0.09	-0.6
38	-0.01	0.09	-0.11	0.01	-0.01	0.06	-0.11	0.05	0.07
39	-0.28	0.21	-1.2	0.12	0.15	0.22	-0.77	0.18	-1.23
40	1.14	0.28	-2.82	0.15	-0.30	0.27	-4.26	0.19	-4.15
41	0.02	0.15	0.05	0.01	0.07	0.14	0.1	0.09	0

42	-0.18	0.18	-0.48	0.04	-0.05	0.19	-0.35	0.12	0.26
43	0.04	0.09	0.11	0.03	-0.08	0.09	-0.01	0.07	0.34
44	0.02	0.06	-1.63	0.11	2.62	0.36	-1.63	0.23	-2.28
45	0.00	0.11	0.04	0.01	-0.04	0.15	0	0.09	0.27
46	2.56	0.21	0.25	0.02	-0.48	0.27	-2.79	0.16	-0.45
47	-0.28	0.12	-2.91	0.22	-0.11	0.19	-2.74	0.16	-4.12
48	0.00	0.07	-0.18	0.01	-0.97	0.27	-1.15	0.15	0.41
49	-0.07	0.09	-0.11	0.01	-0.06	0.09	-0.1	0.06	-1.29
50	-0.60	0.23	-4.4	0.1	-0.02	0.03	-3.73	0.14	-3.98
51	-0.32	0.16	0.32	0.06	-0.42	0.22	0.22	0.13	-0.09
52	2.91	0.31	2.61	0.07	0.68	0.12	0.38	0.18	2.16
53	0.78	0.21	-2.97	0.04	2.32	0.31	-1.43	0.18	-0.19
54	0.02	0.09	0.16	0.03	-0.02	0.04	0.16	0.07	-0.23
55	0.04	0.04	-1.44	0.11	-0.19	0.04	-1.67	0.08	-2.35
56	-0.08	0.17	0.26	0.11	0.03	0.10	0.37	0.12	0.4
57	0.21	0.09	0.1	0.05	0.35	0.12	0.24	0.09	0.37
58	-0.04	0.07	-5.97	0.18	1.93	0.27	-4	0.12	-6.67
59	-0.05	0.09	-3.77	0.16	0.02	0.07	-3.7	0.09	-4.68
60	0.06	0.10	0.06	0.01	0.40	0.16	0.4	0.08	0.19

Table 1.3: MM→QM/MM relative hydration free energy results. ΔG_1 is the QM/MM correction for λ=0 for each perturbation. ΔG_3 is the QM/MM correction for λ=1 for each perturbation. ΔG_2 is the MM relative free energy change between the two endstates of each perturbation. $\Delta\Delta G_{hyd}$ is the MM→QM/MM free energy difference between the two end states of each perturbation. One experimental dataset (Guthrie) is also reported. The errors shown were calculated from four independent simulations using standard error.

UNIVERSITY OF SOUTHAMPTON

Development and Application of a QM/MM Method for Free Energy Calculations

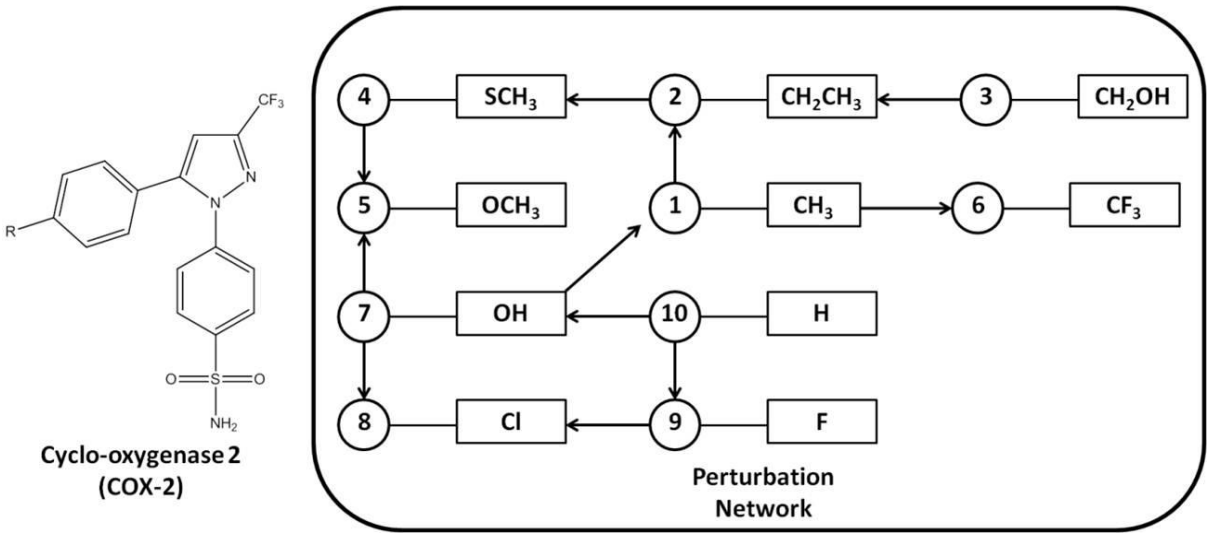
SUPPORTING INFORMATION 2

COX-2 Binding Free Energy Results

Michael Keith Carter

School of Chemistry
University of Southampton

September 2013


CF3
R
N
N
O=S=O
NH2
Cyclo-oxygenase 2
(COX-2)
4
SCH3
2
CH2CH3
3
CH2OH
5
OCH3
1
CH3
6
CF3
7
OH
10
H
8
Cl
9
F
Perturbation
Network

COX-2 – MM-RETI Results

Pert	Exp ($kcal.mol^{-1}$)	$\Delta\Delta G_{bind}$ ($kcal.mol^{-1}$)	Error	ΔG_{prot} ($kcal.mol^{-1}$)	Error	ΔG_{wat} ($kcal.mol^{-1}$)	Error	ΔG_{vac} ($kcal.mol^{-1}$)	Error
1t2	1.82	2.25	0.36	3.56	0.27	1.32	0.24	0.74	0.09
3t2	-2.78	-2.17	0.45	0.8	0.26	2.98	0.37	-3.95	0.09
1t6	3.16	3.9	0.25	20.71	0.11	16.81	0.23	16.42	0.07
7t1	-4.64	-2.9	0.37	15.03	0.15	17.93	0.34	13.54	0.08
7t8	-5.46	-3.49	0.31	17.12	0.11	20.61	0.29	15.41	0.02
9t8	-0.84	-1.33	0.18	-0.19	0.08	1.14	0.16	1.22	0.01
10t9	0.15	0.01	0.17	-3.47	0.05	-3.48	0.18	-4.43	0.02
10t7	4.77	1.66	0.29	-21.47	0.1	-23.13	0.27	-18.64	0.02
2t4	-2.7	-1.7	0.44	-3.19	0.36	-1.49	0.26	-1.03	0.22
4t5	-0.07	-1.75	0.52	-6.56	0.38	-4.81	0.36	-3.63	0.53
7t5	-5.59	-3.68	0.75	9.35	0.39	13.03	0.64	9	0.35

Table 2.1: MM-RETI relative binding free energies for a set of COX-2 perturbations. $\Delta\Delta G_{bind}$ is the overall binding free energy change between our endstates. ΔG_{prot} is the free energy change for the bound leg, ΔG_{wat} is the free energy change in aqueous solution and ΔG_{vac} is the free energy change within a vacuum. The errors shown were calculated from four independent simulations using standard error.

COX-2 – MM→QM/MM-FEP Results

Pert	ΔG_1	Error	ΔG_2	Error	ΔG_3	Error	ΔG_5	Error	ΔG_6	Error	ΔG_7	Error
1t2	-2.23	0.36	3.56	0.27	-2.47	0.35	-0.77	0.33	1.32	0.24	-0.71	0.21
3t2	-1.95	0.24	0.8	0.26	-2.47	0.31	-1.64	0.36	2.98	0.37	-1.89	0.24
1t6	-2.36	0.28	20.71	0.11	-3.66	0.44	-1.17	0.26	16.81	0.23	-0.79	0.28
7t1	-1.63	0.2	15.03	0.15	-2.12	0.21	-0.62	0.24	17.93	0.34	-3.13	0.42
7t8	-1.61	0.48	17.12	0.11	-2.4	0.2	-2.12	0.39	20.61	0.29	-3.05	0.32
9t8	-1.12	0.39	-0.19	0.08	-2.35	0.23	-2.12	0.31	1.14	0.16	-1.35	0.26
10t9	-0.77	0.42	-3.47	0.05	-1.03	0.26	-1.08	0.41	-1.48	0.18	-1.62	0.25
10t7	-0.69	0.22	-21.47	0.1	-1.71	0.25	-1.75	0.35	-23.13	0.27	-1.53	0.38
2t4	-1.65	0.34	-3.19	0.36	-1.95	0.32	-1.74	0.25	-1.49	0.26	-1.52	0.25
4t5	-2.97	0.25	-6.56	0.38	-2.01	0.25	-3.56	0.34	-4.81	0.36	-1.45	0.31
7t5	-1.56	0.31	9.35	0.39	-1.87	0.3	-2.14	0.22	13.03	0.64	-1.93	0.39

Table 2.2: MM→QM/MM relative binding free energies for a set of COX-2 perturbations. ΔG_1 is the QM/MM correction for λ=0 in the bound state. ΔG_2 is the MM free energy changes between our two endstates in our bound free energy simulations. ΔG_3 is the QM/MM correction for λ=1 in the bound state. ΔG_5 is the QM/MM correction for λ=1 in the free state. ΔG_6 is the MM free energy changes between our two endstates in our aqueous free energy simulations. ΔG_7 is the QM/MM correction for λ=0 in the free state. The errors shown were calculated from four independent simulations using standard error.

Compound	Perturbation Pathway	$\Delta\Delta G_{bind}$ (MM) (kcal.mol^{-1})	Error	$\Delta\Delta G_{bind}$ (MM→QM/MM) (kcal.mol^{-1})	Error	Exp (kcal.mol^{-1})
5 (OCH3)	[1t7+7t5] ; [1t2+2t4+4t5]	-1	0.71	0.77	0.86	-0.95
4 (SCH3)	[1t2+2t4]	0.73	0.55	2.04	0.8	-0.88
8 (Cl)	[1t7+7t8]	-0.67	0.21	0.55	0.58	-0.82
10 (H)	[1t7+7t8+8t9+9t10];[1t7+7t10]	1.89	0.22	4.23	0.59	-0.13
1 (CH3)		0	0	0	0	0
9 (F)	[1t7+7t8+8t9];[1t7+7t10+10t9]	0.74	0.22	2.29	0.62	0.01
2 (CH2CH3)	1t2	2	0.32	1.82	0.71	1.82
6 (CF3)	1t6	2.34	0.19	1.42	0.67	3.15
3 (CH2OH)	[1t2+2t3]	4.96	0.43	5.65	0.76	4.59
7 (OH)	1t7	3.07	0.2	6.07	0.6	4.63

Table 2.3: MM-RETI and MM→QM/MM relative binding free energies for a set of COX-2 perturbations in reference to compound 1. $\Delta\Delta G_{bind}$ (MM) is the overall binding free energy change between our endstates in MM. $\Delta\Delta G_{bind}$ (MM→QM/MM) is the overall binding free energy change between our endstates in QM/MM. The errors shown were calculated from four independent simulations using standard error.

COX-2 – Charge Perturbation Results

Ligand 10

Free

CF	ΔG_{71} (MM) (kcal.mol^{-1})	Error	ΔG_{72} (QM/MM) (kcal.mol^{-1})	Error	ΔG_7 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_7 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.49	0.01	-0.39	0.14	-0.88	0.14	-1.58
1.05	-2.49	0.05	1.88	0.16	-0.61	0.17	
1.07	-3.55	0.03	4.02	0.21	0.47	0.21	
1.10	-5.23	0.01	5.82	0.19	0.59	0.19	
1.15	-8.13	0.06	7.01	0.14	-1.12	0.15	
1.20	-11.23	0.09	9.37	0.2	-1.86	0.22	

Table 2.4: Charge Perturbation free energies for the free leg of ligand 10. ΔG_{71} is the MM free energy change at a specific scaling factor. ΔG_{72} is the QM/MM free energy change. ΔG_7 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{12} (MM) (kcal.mol^{-1})	Error	ΔG_{13} (QM/MM) (kcal.mol^{-1})	Error	ΔG_1 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_1 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.27	0.05	-0.58	0.13	-0.85	0.14	-0.73
1.05	-1.39	0.11	0.82	0.07	-0.57	0.13	
1.07	-2.06	0.14	1.46	0.31	-0.6	0.34	
1.10	-2.83	0.17	2.25	0.15	-0.58	0.23	
1.15	-4.43	0.16	3.45	0.11	-0.98	0.19	
1.20	-6.18	0.09	6.09	0.18	-0.09	0.2	

Table 2.5: Charge Perturbation free energies for the bound leg of ligand 10. ΔG_{12} is the MM free energy change at a specific scaling factor. ΔG_{13} is the QM/MM free energy change. ΔG_1 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Ligand 9

Free

CF	ΔG_{51} (MM) (kcal.mol^{-1})	Error	ΔG_{52} (QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.47	0.09	-1.43	0.09	-1.9	0.13	-1.38
1.05	-2.4	0.1	1.75	0.2	-0.65	0.22	
1.07	-3.41	0.07	3.05	0.18	-0.36	0.19	
1.10	-5.05	0.08	5.84	0.29	0.79	0.3	
1.15	-7.81	0.11	6.61	0.31	-1.2	0.33	
1.20	-10.92	0.12	9.43	0.23	-1.49	0.26	

Table 2.6: Charge Perturbation free energies for the free leg of ligand 9. ΔG_{51} is the MM free energy change at a specific scaling factor. ΔG_{52} is the QM/MM free energy change. ΔG_5 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{32} (MM) (kcal.mol^{-1})	Error	ΔG_{33} (QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.26	0.07	-1.35	0.19	-1.61	0.20	-1.07
1.05	-1.36	0.09	-0.21	0.14	-1.57	0.17	
1.07	-1.94	0.16	0.8	0.23	-1.14	0.28	
1.10	-2.71	0.17	1.3	0.18	-1.41	0.25	
1.15	-4.24	0.12	2.35	0.24	-1.89	0.27	
1.20	-5.89	0.15	5.18	0.25	-0.71	0.29	

Table 2.7: Charge Perturbation free energies for the bound leg of ligand 9. ΔG_{32} is the MM free energy change at a specific scaling factor. ΔG_{33} is the QM/MM free energy change. ΔG_3 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Ligand 7

Free

CF	ΔG_{51} (MM) (kcal.mol^{-1})	Error	ΔG_{52} (QM/MM) (kcal.mol^{-1})	Error	ΔG_{5} (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_{5} (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.58	0.002	-2.31	0.18	-2.89	0.18	-3.08
1.05	-2.98	0.01	0.82	0.13	-2.16	0.13	
1.07	-4.26	0.02	2.81	0.17	-1.45	0.17	
1.10	-6.27	0.02	3.23	0.13	-3.04	0.13	
1.15	-9.61	0.06	6.61	0.24	-3	0.25	
1.20	-13.52	0.09	12.38	0.29	-1.14	0.30	

Table 2.8: Charge Perturbation free energies for the free leg of ligand 7. ΔG_{51} is the MM free energy change at a specific scaling factor. ΔG_{52} is the QM/MM free energy change. ΔG_{5} is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{32} (MM) (kcal.mol^{-1})	Error	ΔG_{33} (QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.3	0.09	-0.73	0.17	-1.03	0.19	-1.62
1.05	-1.51	0.12	0.48	0.18	-1.03	0.22	
1.07	-2.13	0.08	1.16	0.21	-0.97	0.22	
1.10	-3.1	0.11	2.14	0.2	-0.96	0.23	
1.15	-4.87	0.15	4.15	0.22	-0.72	0.27	
1.20	-6.6	0.13	6.02	0.24	-0.58	0.27	

Table 2.9: Charge Perturbation free energies for the bound leg of ligand 7. ΔG_{32} is the MM free energy change at a specific scaling factor. ΔG_{33} is the QM/MM free energy change. ΔG_3 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

UNIVERSITY OF SOUTHAMPTON

Development and Application of a QM/MM Method for Free Energy Calculations

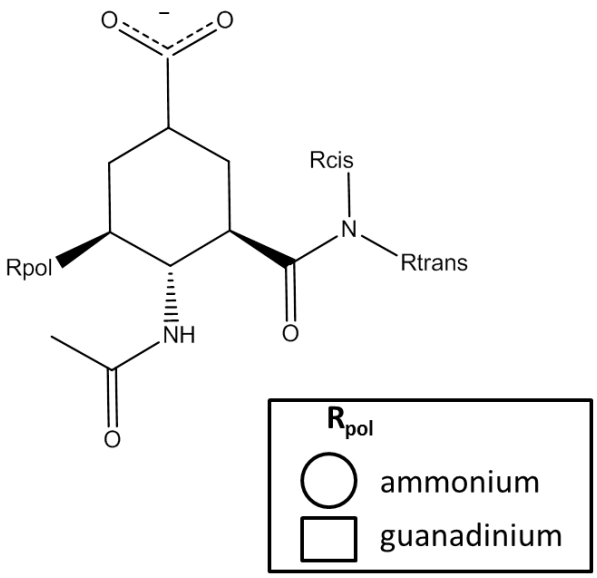
SUPPORTING INFORMATION 3

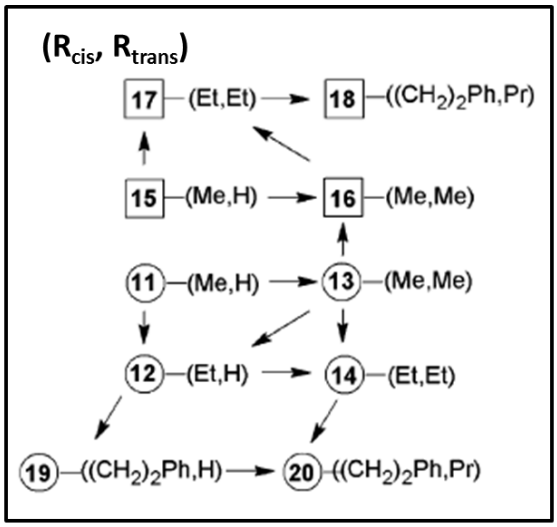
Neuraminidase Binding Free Energy Results

Michael Keith Carter

School of Chemistry
University of Southampton

September 2013


O
O
Rcis
N
Rpol
Rtrans
NH
O
O
Rpol
ammonium
guanadinium


(Rcis, Rtrans)
17 (Et,Et) 18 ((CH2)2Ph,Pr)
15 (Me,H) 16 (Me,Me)
11 (Me,H) 13 (Me,Me)
12 (Et,H) 14 (Et,Et)
19 ((CH2)2Ph,H) 20 ((CH2)2Ph,Pr)

Neuraminidase – MM-RETI Results

Pert	Exp ($kcal.mol^{-1}$)	$\Delta\Delta G_{bind}$ ($kcal.mol^{-1}$)	Error	ΔG_{prot} ($kcal.mol^{-1}$)	Error	ΔG_{wat} ($kcal.mol^{-1}$)	Error	ΔG_{vac} ($kcal.mol^{-1}$)	Error
11t13	-2.67	-5.25	0.62	21.87	0.34	27.12	0.52	28.7	0.43
11t12	-1.63	1.44	0.41	1.87	0.29	0.43	0.29	-0.81	0.14
12t13	-1.04	-7.19	0.77	19.74	0.43	26.93	0.64	29.45	0.45
13t14	-4.09	-4.14	0.69	-8.78	0.51	-4.64	0.47	-6.1	0.27
12t14	-5.13	-9.32	1.06	12.04	0.65	21.36	0.84	23.89	0.68
12t19	0.08	-2.56	1.21	-7.3	0.8	-4.74	0.91	-5.58	0.2
14t20	0.25	-5.46	1.38	-3.12	0.9	2.34	1.04	0	0.31
19t20	-4.8	-11.86	1.33	13.33	0.81	25.19	1.06	26.68	0.78
13t16	-2.78	-5.14	1.3	-19.82	1.04	-14.64	1.23	-8.05	0.89
15t16	-3.45	-5.61	0.58	20.69	0.27	26.3	0.51	27.92	0.35
15t17	-5.15	-7.34	1.18	15.12	0.78	22.46	0.88	22.36	0.56
16t17	-1.7	-1.04	0.69	-4.3	0.48	-3.26	0.5	-4.78	0.29
17t18	0.65	-3.97	1.39	-4.62	0.99	-0.65	0.97	-3.72	0.37

Table 3.1: MM-RETI relative binding free energies for a set of Neuraminidase perturbations. $\Delta\Delta G_{bind}$ is the overall binding free energy change between our endstates. ΔG_{prot} is the free energy change for the bound leg, ΔG_{wat} is the free energy change in aqueous solution and ΔG_{vac} is the free energy change within a vacuum. The errors shown were calculated from four independent simulations using standard error.

Neuraminidase – MM→QM/MM-FEP Results

Pert	ΔG_1	Error	ΔG_2	Error	ΔG_3	Error	ΔG_5	Error	ΔG_6	Error	ΔG_7	Error
11t13	-7.24	0.44	21.87	0.34	-7.65	0.29	-9.84	0.44	27.12	0.52	-9.77	0.3
11t12	-7.11	0.23	1.87	0.29	-6.01	0.34	-9.12	0.35	0.43	0.29	-9.62	0.35
12t13	-7.72	0.35	19.74	0.43	-7.61	0.46	-8.42	0.36	26.93	0.64	-8.91	0.33
13t14	-7.41	0.34	-8.78	0.51	-7.61	0.35	-9.12	0.39	-4.64	0.47	-9.42	0.41
12t14	-7.71	0.31	12.04	0.65	-7.34	0.32	-9.11	0.44	21.36	0.84	-9.41	0.35
12t19	-7.99	0.29	-7.3	0.8	-10.94	0.3	-13.24	0.45	-4.74	0.91	-9.81	0.39
14t20	-7.52	0.45	-3.12	0.9	-10.95	0.2	-13.11	0.49	2.34	1.04	-9.17	0.42
19t20	-11.45	0.23	13.33	0.81	-11.62	0.42	-10.98	0.5	25.19	1.06	-13.52	0.26
13t16	-7.63	0.36	-19.82	1.04	-8.41	0.29	-9.05	0.3	-14.64	1.23	-10.21	0.31
15t16	-8.62	0.44	20.69	0.27	-8.42	0.38	-10.32	0.4	26.3	0.51	-10.56	0.4
15t17	-8.91	0.29	15.12	0.78	-9.16	0.24	-10.42	0.41	22.46	0.88	-10.89	0.35
16t17	-8.85	0.35	-4.3	0.48	-9.08	0.25	-10.66	0.39	-3.26	0.5	-10.94	0.52
17t18	-9.21	0.38	-4.62	0.99	-11.91	0.4	-10.94	0.33	-0.65	0.97	-12.81	0.42

Table 3.2: Table 2.2: MM→QM/MM relative binding free energies for a set of Neuraminidase perturbations. ΔG_1 is the QM/MM correction for λ=0 in the bound state. ΔG_2 is the MM free energy changes between our two endstates in our bound free energy simulations. ΔG_3 is the QM/MM correction for λ=1 in the bound state. ΔG_5 is the QM/MM correction for λ=1 in the free state. ΔG_6 is the MM free energy changes between our two endstates in our aqueous free energy simulations. ΔG_7 is the QM/MM correction for λ=0 in the free state. The errors shown were calculated from four independent simulations using standard error.

Compound	Perturbation Pathway	$\Delta\Delta G_{bind}$ (MM) (kcal.mol^{-1})	Error	$\Delta\Delta G_{bind}$ (MM→QM/MM) (kcal.mol^{-1})	Error	Exp
17 (□ (Et, Et))	[11t13+13t16+16t17]	-9.69	1.43	-10.64	1.6	-7.15
14 (○ (Et, Et))	[11t13+13t14];[11t12+12t14]	-8.64	1.03	-9.66	1.27	-6.76
18 (□ ((CH2)2Ph, Pr))	[11t13+13t16+16t17+17t18]	-13.66	2	-15.06	2.13	-6.51
20 (○ ((CH2)2Ph, Pr))	[11t12+12t19+19t20]	-13.92	1.75	-15.07	1.91	-6.51
16 (□ (Me, Me))	[11t13+13t16]	-8.65	1.26	-10.54	1.43	-5.45
13 (○ (Me, Me))	[11t13]	-5.25	0.62	-5.59	0.97	-2.67
15 (□ (Me, H))	[11t13+13t16+16t15]	-3.04	1.39	-3.63	1.59	-2
19 (○ ((CH2)2Ph, H))	[11t12+12t19]	-1.12	1.28	-1.33	1.45	-1.71
12 (○ (Et, H))	[11t12];[11t13+13t12]	1.69	0.7	2.29	0.95	-1.63
11 (○ (Me, H))		0	0	0	0	0

Table 3.3: MM-RETI and MM→QM/MM relative binding free energies for a set of Neuraminidase perturbations in reference to compound 11. $\Delta\Delta G_{bind}$ (MM) is the overall binding free energy change between our endstates in MM. $\Delta\Delta G_{bind}$ (MM→QM/MM) is the overall binding free energy change between our endstates in QM/MM. The errors shown were calculated from four independent simulations using standard error.

Neuraminidase – Charge Perturbation Results

Ligand 11

Free

CF	ΔG_{71} (MM) (kcal.mol^{-1})	Error	ΔG_{72} (QM/MM) (kcal.mol^{-1})	Error	ΔG_{7} (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_{7} (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.23	0.01	-7.52	0.13	-7.75	0.13	-9.68
1.05	-1.18	0.05	-6.89	0.19	-8.07	0.19	
1.07	-1.65	0.03	-6.03	0.21	-7.68	0.21	
1.10	-2.39	0.01	-5.28	0.22	-7.67	0.27	
1.15	-3.55	0.06	-4.41	0.34	-7.96	0.35	
1.20	-4.75	0.09	-4.05	0.35	-8.8	0.41	

Table 3.4: Charge Perturbation free energies for the free leg of ligand 11. ΔG_{71} is the MM free energy change at a specific scaling factor. ΔG_{72} is the QM/MM free energy change. ΔG_{7} is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{12} (MM) (kcal.mol^{-1})	Error	ΔG_{13} (QM/MM) (kcal.mol^{-1})	Error	ΔG_{1} (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_{1} (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.23	0.05	-7.49	0.1	-7.72	0.21	-7.18
1.05	-1.17	0.11	-6.92	0.2	-8.09	0.22	
1.07	-1.62	0.14	-5.52	0.19	-7.14	0.22	
1.10	-2.32	0.17	-4.27	0.21	-6.59	0.31	
1.15	-3.49	0.16	-3.84	0.19	-7.33	0.32	
1.20	-4.69	0.09	-1.98	0.34	-6.67	0.47	

Table 3.5: Charge Perturbation free energies for the bound leg of ligand 11. ΔG_{12} is the MM free energy change at a specific scaling factor. ΔG_{13} is the QM/MM free energy change. ΔG_{3} is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Ligand 12

Free

CF	ΔG_{51} (MM) (kcal.mol^{-1})	Error	ΔG_{52} (QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.23	0.09	-9.72	0.19	-9.95	0.21	-9.41
1.05	-1.15	0.12	-9.62	0.21	-10.77	0.24	
1.07	-1.62	0.16	-7.78	0.28	-9.4	0.32	
1.10	-2.34	0.23	-7.66	0.35	-10	0.42	
1.15	-3.5	0.22	-6.6	0.36	-10.1	0.42	
1.20	-4.7	0.25	-5.74	0.43	-10.44	0.5	

Table 3.6: Charge Perturbation free energies for the free leg of ligand 12. ΔG_{51} is the MM free energy change at a specific scaling factor. ΔG_{52} is the QM/MM free energy change. ΔG_5 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{32} (MM) (kcal.mol^{-1})	Error	ΔG_{33} (QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.23	0.08	-7.98	0.19	-8.21	0.21	-7.82
1.05	-1.14	0.11	-7.34	0.27	-8.48	0.29	
1.07	-1.61	0.2	-5.86	0.32	-7.47	0.38	
1.10	-2.29	0.35	-5.36	0.41	-7.65	0.54	
1.15	-3.43	0.32	-5.21	0.44	-8.64	0.54	
1.20	-4.62	0.3	-4.42	0.46	-9.04	0.55	

Table 3.7: Charge Perturbation free energies for the bound leg of ligand 12. ΔG_{32} is the MM free energy change at a specific scaling factor. ΔG_{33} is the QM/MM free energy change. ΔG_3 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Ligand 13

Free

CF	ΔG_{51} (MM) (kcal.mol^{-1})	Error	ΔG_{52} (QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.23	0.11	-9.25	0.19	-9.48	0.22	-9.42
1.05	-1.14	0.18	-8.87	0.24	-10.01	0.3	
1.07	-1.59	0.19	-7.41	0.33	-9	0.38	
1.10	-2.27	0.2	-7.78	0.35	-10.05	0.4	
1.15	-3.41	0.31	-5.82	0.48	-9.23	0.57	
1.20	-4.55	0.43	-4.43	0.41	-8.98	0.59	

Table 3.8: Charge Perturbation free energies for the free leg of ligand 13. ΔG_{51} is the MM free energy change at a specific scaling factor. ΔG_{52} is the QM/MM free energy change. ΔG_5 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{32} (MM) (kcal.mol^{-1})	Error	ΔG_{33} (QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.23	0.18	-8.75	0.22	-8.98	0.28	-7.72
1.05	-1.14	0.24	-8.01	0.35	-9.15	0.42	
1.07	-1.6	0.28	-6.99	0.28	-8.59	0.4	
1.10	-2.26	0.26	-6.44	0.39	-8.7	0.47	
1.15	-3.44	0.22	-5.15	0.35	-8.59	0.41	
1.20	-4.57	0.36	-4.73	0.31	-9.3	0.48	

Table 3.9: Charge Perturbation free energies for the bound leg of ligand 13. ΔG_{32} is the MM free energy change at a specific scaling factor. ΔG_{33} is the QM/MM free energy change. ΔG_3 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

UNIVERSITY OF SOUTHAMPTON

Development and Application of a QM/MM Method for Free Energy Calculations

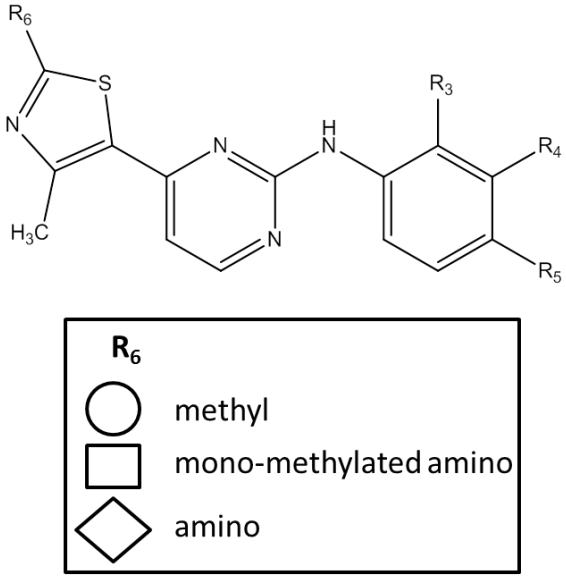
SUPPORTING INFORMATION 4

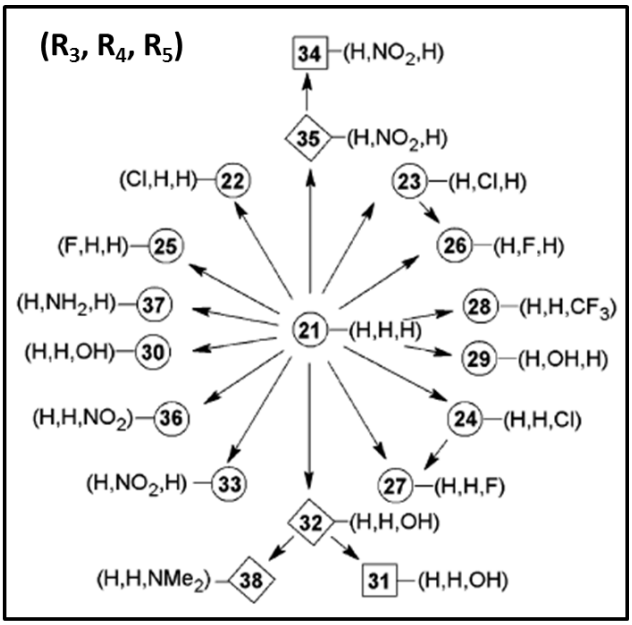
CDK2 Binding Free Energy Results

Michael Keith Carter

School of Chemistry
University of Southampton

September 2013





CDK2 – MM-RETI Results

Pert	Exp (kcal.mol^{-1})	$\Delta\Delta G_{bind}$ (kcal.mol^{-1})	Error	ΔG_{prot} (kcal.mol^{-1})	Error	ΔG_{wat} (kcal.mol^{-1})	Error	ΔG_{vac} (kcal.mol^{-1})	Error
21t25	1.61	6.86	0.24	37.67	0.18	30.81	0.16	26.79	0.04
21t22	3.27	-1.22	0.6	16.8	0.55	18.03	0.25	15.48	0.04
21t24	2.04	-3.1	0.37	4.06	0.27	7.16	0.25	5.67	0.01
21t33	0.19	1.21	0.58	17.61	0.4	16.4	0.41	18.71	0.15
21t26	0.13	-2.4	0.19	-6.67	0.13	-4.27	0.14	-6.45	0.02
21t27	-0.41	-1.87	0.19	8.94	0.14	10.81	0.13	8.96	0.02
35t34	3.55	-2.06	0.54	37.72	0.37	39.78	0.39	37.4	0.15
32t38	1.87	9.47	0.94	19.09	0.6	9.62	0.73	9.24	0.16
21t29	-0.17	-0.66	0.36	-43.84	0.26	-43.18	0.27	-37.67	0.1
21t23	1.26	-1.45	0.58	-12.39	0.4	-10.93	0.42	-12.74	0.02
23t26	-1.13	-0.52	0.35	5.76	0.24	6.28	0.25	6.26	0.01
21t36	2.33	-5.18	0.59	-7.49	0.4	-2.31	0.44	1.48	0.07

32t31	0.5	-1.86	0.52	19.48	0.29	21.34	0.37	19.7	0.15
21t35	-2.18	-3.36	0.79	-26.14	0.55	-22.78	0.57	-13.8	0.31
21t28	0.77	-4.49	0.63	7.2	0.42	11.69	0.47	9.24	0.06
24t27	2.45	1.07	0.21	4.49	0.14	3.42	0.16	3.28	0.01
21t32	-0.58	-10.21	0.72	-36.72	0.52	-26.51	0.5	-19.42	0.53
21t30	0.33	-4.7	0.35	-18.32	0.26	-13.63	0.26	-9.99	0.04
21t37	0.96	0.05	0.29	-33.75	0.21	-33.8	0.24	-30.61	0.04

Table 4.1: MM-RETI relative binding free energies for a set of CDK2 perturbations. $\Delta\Delta G_{bind}$ is the overall binding free energy change between our endstates. ΔG_{prot} is the free energy change for the bound leg, ΔG_{wat} is the free energy change in aqueous solution and ΔG_{vac} is the free energy change within a vacuum. The errors shown were calculated from four independent simulations using standard error.

CDK2 – MM→QM/MM-FEP Results

Pert	**ΔG_1**	**Error**	**ΔG_2**	**Error**	**ΔG_3**	**Error**	**ΔG_5**	**Error**	**ΔG_6**	**Error**	**ΔG_7**	**Error**
21t25	-0.35	0.08	37.67	0.18	-0.62	0.17	0.32	0.16	30.81	0.16	-0.44	0.14
21t22	-0.29	0.18	16.8	0.55	-0.78	0.21	0.22	0.21	18.03	0.25	-0.65	0.12
21t24	-0.26	0.15	4.06	0.27	-0.69	0.19	0.17	0.12	7.16	0.25	-0.72	0.19
21t33	-0.24	0.09	17.61	0.4	-2.34	0.43	-1.81	0.1	16.4	0.41	-0.52	0.34
21t26	-0.21	0.12	-6.67	0.13	-0.31	0.16	0.33	0.09	-4.27	0.14	-0.32	0.12
21t27	-0.32	0.14	8.94	0.14	-0.55	0.14	0.19	0.14	10.81	0.13	-0.41	0.14
35t34	-2.23	0.36	37.72	0.37	-3.21	0.35	-2.51	0.41	39.78	0.39	-2.78	0.31
32t38	-2.85	0.32	19.09	0.6	-1.2	0.23	-1.33	0.39	9.62	0.73	-1.79	0.21

21t29	-0.34	0.15	-43.84	0.26	-2.41	0.38	-1.05	0.19	-43.18	0.27	-0.51	0.32
21t23	-0.41	0.12	-12.39	0.4	-1.06	0.16	-0.33	0.17	-10.93	0.42	-0.41	0.14
23t26	-1.01	0.23	5.76	0.24	-0.44	0.18	-0.87	0.18	6.28	0.25	-0.41	0.1
21t36	-0.23	0.17	-7.49	0.4	-2.69	0.32	-2.88	0.21	-2.31	0.44	-0.35	0.41
32t31	-2.99	0.28	19.48	0.29	-3.34	0.27	-1.25	0.32	21.34	0.37	-2.15	0.33
21t35	-0.35	0.18	-26.14	0.55	-2.52	0.41	-2.29	0.1	-22.78	0.57	-0.45	0.39
21t28	-0.29	0.11	7.2	0.42	-0.88	0.13	0.27	0.09	11.69	0.47	-0.35	0.16
24t27	-0.75	0.32	4.49	0.14	-0.61	0.1	-0.77	0.1	3.42	0.16	-0.43	0.08
21t32	-0.32	0.21	-36.72	0.52	-3.05	0.2	-1.78	0.12	-26.51	0.5	-0.33	0.3
21t30	-0.3	0.22	-18.32	0.26	-3.21	0.34	-1.82	0.14	-13.63	0.26	-0.41	0.34
21t37	-0.2	0.15	-33.75	0.21	-2.85	0.3	-2.12	0.15	-33.8	0.24	-0.38	0.45

Table 4.2: Table 2.2: MM→QM/MM relative binding free energies for a set of Neuraminidase perturbations. ΔG_1 is the QM/MM correction for λ=0 in the bound state. ΔG_2 is the MM free energy changes between our two endstates in our bound free energy simulations. ΔG_3 is the QM/MM correction for λ=1 in the bound state. ΔG_5 is the QM/MM correction for λ=1 in the free state. ΔG_6 is the MM free energy changes between our two endstates in our aqueous free energy simulations. ΔG_7 is the QM/MM correction for λ=0 in the free state. The errors shown were calculated from four independent simulations using standard error.

Compound	Perturbation Pathway	$\Delta\Delta G_{bind}$ (MM) (kcal.mol^{-1})	Error	$\Delta\Delta G_{bind}$ (MM→QM/MM) (kcal.mol^{-1})	Error	Exp (kcal.mol^{-1})
35 (◊ (H, NO2, H))	[21t35]	-3.36	0.89	-3.69	1.05	-2.18
32 (◊ (H, H, OH))	[21t32]	-10.21	0.72	-11.49	0.84	-0.58
27 (○ (H, H, F))	[21t27]	-3.91	0.19	-4.74	0.34	-0.41
29 (○ (H, OH, H))	[21t29]	-0.66	0.36	-2.19	0.66	-0.17
31 (□ (H, H, OH))	[21t32+32t31]	-12.07	0.86	-14.25	0.98	-0.08
21 (○ (H, H, H))		0	0	0	0	0
26 (○ (H, F, H))	[21t26]	-4.37	0.31	-5.12	0.4	0.13
33 (○ (H, NO2, H))	[21t33]	1.21	0.41	0.4	0.7	0.19
30 (○ (H, H, OH))	[21t30]	-4.7	0.35	-6.2	0.65	0.33
28 (○ (H, H, CF3))	[21t28]	-4.49	0.63	-5.7	0.68	0.76
37 (○ (H, NH2, H))	[27t37]	0.05	0.29	-0.86	0.65	0.95
23 (○ (H, Cl, H))	[21t23];[21t26+26t23]	-1.45	0.75	-2.18	0.81	1.26
38 (◊ (H, H, NMe2))	[21t32+32t38]	-0.74	1.18	-2.02	1.26	1.28
34 (□ (H, NO2, H))	[21t35+35t34]	-5.42	0.96	-6.22	1.18	1.36
25 (○ (F, H, H))	[21t25]	6.86	0.24	5.83	0.37	1.6
24 (○ (H, H, Cl))	[21t24];[21t27+27t24]	-3.1	0.47	-4.42	0.57	2.04
36 (○ (H, H, NO2))	[21t36]	-5.18	0.59	-5.11	0.83	2.33
22 (○ (Cl, H, H))	[21t22]	-1.22	0.6	-2.58	0.7	3.27
35 (◊ (H, NO2, H))	[21t35]	-3.36	0.89	-3.69	1.05	-2.18

Table 4.3: MM-RETI and MM→QM/MM relative binding free energies for a set of CDK2 perturbations in reference to compound 21. $\Delta\Delta G_{bind}$ (MM) is the overall binding free energy

change between our endstates in MM. $\Delta\Delta G_{bind}$ (MM→QM/MM) is the overall binding free energy change between our endstates in QM/MM. The errors shown were calculated from four independent simulations using standard error.

CDK2 – Charge Perturbation Results

Ligand 21

Free

CF	ΔG_{72} (MM) (kcal.mol^{-1})	Error	ΔG_{73} (QM/MM) (kcal.mol^{-1})	Error	ΔG_{7} (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_{7} (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.37	0	0.68	0.04	0.31	0.04	-0.35
1.05	-1.91	0.01	1.53	0.12	-0.38	0.12	
1.07	-2.75	0.02	3.56	0.05	0.81	0.05	
1.10	-3.98	0.01	4.54	0.11	0.56	0.11	
1.15	-6.21	0.04	6.97	0.05	0.76	0.06	
1.20	-8.83	0.06	8.42	0.17	-0.41	0.18	

Table 4.4: Charge Perturbation free energies for the free leg of ligand 21. ΔG_{72} is the MM free energy change at a specific scaling factor. ΔG_{73} is the QM/MM free energy change. ΔG_{7} is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{12} (MM) (kcal.mol^{-1})	Error	ΔG_{13} (QM/MM) (kcal.mol^{-1})	Error	ΔG_1 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_1 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.18	0.002	-0.16	0.1	-0.34	0.012	-0.25
1.05	-0.97	0.01	0.49	0.08	-0.48	0.02	
1.07	-1.37	0.01	0.92	0.12	-0.45	0.02	
1.10	-2	0.04	1.32	0.12	-0.68	0.05	
1.15	-3.13	0.06	2.67	0.08	-0.46	0.07	
1.20	-4.32	0.08	4.67	0.03	0.35	0.08	

Table 4.5: Charge Perturbation free energies for the bound leg of ligand 21. ΔG_{12} is the MM free energy change at a specific scaling factor. ΔG_{13} is the QM/MM free energy change. ΔG_1 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Ligand 22

Free

CF	ΔG_{52} (MM) (kcal.mol^{-1})	Error	ΔG_{53} (QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.31	0.002	-0.13	0.04	-0.44	0.04	-0.72
1.05	-1.57	0.01	0.76	0.19	-0.81	0.19	
1.07	-2.22	0.02	1.4	0.15	-0.82	0.15	
1.10	-3.28	0.02	2.67	0.16	-0.61	0.16	
1.15	-5.1	0.02	4.23	0.11	-0.87	0.11	
1.20	-7.1	0.03	6.66	0.24	-0.44	0.24	

Table 4.6: Charge Perturbation free energies for the free leg of ligand 22. ΔG_{52} is the MM free energy change at a specific scaling factor. ΔG_{53} is the QM/MM free energy change. ΔG_5 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{32} (MM) (kcal.mol^{-1})	Error	ΔG_{33} (QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.18	0.004	0.39	0.03	0.21	0.03	0.22
1.05	-0.93	0.02	0.86	0.05	-0.07	0.05	
1.07	-1.29	0.04	1.11	0.07	-0.18	0.08	
1.10	-1.91	0.03	1.78	0.07	-0.13	0.08	
1.15	-2.93	0.04	3.42	0.07	0.49	0.08	
1.20	-4.05	0.03	4.67	0.16	0.62	0.16	

Table 4.7: Charge Perturbation free energies for the bound leg of ligand 22. ΔG_{32} is the MM free energy change at a specific scaling factor. ΔG_{33} is the QM/MM free energy change. ΔG_3 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Ligand 23

Free

CF	ΔG_{52} (MM) (kcal.mol^{-1})	Error	ΔG_{53} (QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_5 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.34	0.004	-0.46	0.06	-0.8	0.06	-0.33
1.05	-1.77	0.01	1.27	0.25	-0.5	0.25	
1.07	-2.49	0.01	1.69	0.1	-0.8	0.1	
1.10	-3.65	0.03	2.99	0.4	-0.66	0.4	
1.15	-5.71	0.06	5.67	0.28	-0.04	0.29	
1.20	-7.78	0.03	8.92	0.19	1.14	0.19	

Table 4.8: Charge Perturbation free energies for the free leg of ligand 23. ΔG_{52} is the MM free energy change at a specific scaling factor. ΔG_{53} is the QM/MM free energy change. ΔG_5 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.

Bound

CF	ΔG_{32} (MM) (kcal.mol^{-1})	Error	ΔG_{33} (QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (kcal.mol^{-1})	Error	ΔG_3 (MM→QM/MM) (Original) (kcal.mol^{-1})
1.01	-0.21	0.07	-0.93	0.02	-1.14	0.07	-1.06
1.05	-1.09	0.04	0.15	0.08	-0.94	0.09	
1.07	-1.51	0.05	1.17	0.04	-0.34	0.06	
1.10	-2.26	0.03	1.98	0.09	-0.28	0.1	
1.15	-3.36	0.08	3.91	0.23	0.55	0.24	
1.20	-4.75	0.06	5.48	0.19	0.73	0.2	

Table 4.9: Charge Perturbation free energies for the bound leg of ligand 23. ΔG_{32} is the MM free energy change at a specific scaling factor. ΔG_{33} is the QM/MM free energy change. ΔG_3 is the combined MM→QM/MM free energy change for each perturbation. This should be comparable to the original MM→QM/MM value. The errors shown were calculated from four independent simulations using standard error.