Development and Implementation of Algorithms to Determine the Reaction Kinetics and Stability of Biomolecules

A dissertation submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

May 2021

Hafiz Saqib Ali

Department of Chemistry

“The underlying laws necessary for the mathematical theory of large parts of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble.”

Paul Dirac, 1929

Contents

List of Figures

List of Tables

List of Reaction Schemes

List of Abbreviations

MCC	Multiscale Cell Correlation
DFT	Density Functional Theory
QM/MM	Quantum Mechanics / Molecular Mechanics
EE	Energy Entropy
MD	Molecular Dynamics
MC	Monte Carlo
GAFF	General Amber Force Field
OPLS	Optimized Potentials for Liquids Simulation
GAFF2	Second Generation General Amber Force Field
HF-SCF	Hartee-Fock Self-Consistent Field
SE	Semiempirical
SCC-DFTB	Self-Consistent Charge Density Function Tight Binding
B3LYP	Becke, 3-parameter, Lee–Yang–Par
LSD	Local-Spin Density Correlation
FEP	Free Energy Perturbation
TI	Thermodynamic Integration
BAR	Bennett Acceptance Ratio
WHAM	Weighted Histogram Analysis Method
PMF	Potential of Mean Force
MM/PBSA	Molecular Mechanics Poisson-Boltzmann Surface Area
MM/GBSA	Molecular Mechanics Generalised Born Surface Area
NMA	Normal Mode Analysis
QHA	Quasi-Harmonic Analysis
2PT	2-Phase Thermodynamic
CB8	Cucurbit[8]Uril
SAMPL	Statistical Assessment of the Modeling of Proteins and Ligands
TraPPE	Transferable Potentials for Phase Equilibria
DMFA	Dimethylformamide
DMSO	Dimethylsulfoxide
TBA	Tert-Butyl Alcohol
TFE	Tetrafluoroethylene
PCM	Polarizable Continuum Model

TST	Transition State Theory
REC	Reactant Encounter Complex
TS	Transition State
US	Umbrella Sampling
αKG	α-Ketoglutarate
PDB	Protein Data Bank
PMEMD	Particle Mesh Ewald Molecular Dynamics
MCPB	Metal Centre Parameter Builder
ECP	Electron Core Potential
BS	Basis Set
ZPE	Zero-Point Energies
KIE	Kinetic Isotope Effect
BFE	Binding Free Energy

Abstract

Accurately calculating the Gibbs free energies of biomolecules in aqueous phase solution is an important challenge and future goal because most processes take place in aqueous solution. Gibbs free energies give the information about the stability and kinetics of biomolecules which will be helpful in understanding the structure and functions of biomolecules. We have developed a new energy-entropy (EE) method based on multiscale cell correlation (MCC) correction theory which is used to calculated the Gibbs free energy values. Firstly, we applied our MCC theory to calculate the entropy for the range of important industrial liquids modeled with GAFF and OPLS force fields. The calculated entropy values are in good agreement with the experimental values having unsigned errors are 8.7 J K^{-1} mol^{-1} and 9.8 J K^{-1} mol^{-1} for GAFF and OPLS force fields respectively. Later we combined our MCC theory with density functional theory (DFT) in quantum mechanics/molecular mechanics (QM/MM) formalism to develop a new EE-MCC method to calculate the Gibbs free energy barriers. We applied our EE-MCC method to calculate the reaction kinetics for the series of nucleophilic substitution reactions where one halogen atom is replaced by a hydroxyl ion in aqueous solution. The calculated Gibbs free energy barriers agree well with experimental and potentials of mean force (PMF) values and with previous computational methods. Furthermore, we applied our EE-MCC method to calculate the binding free energies directly for molecular dynamics (MD) simulations for the series of seven host-guest complexes in SAMPL8 challenge. The EE-MCC binding free energies are found in good agreement with the experiment values giving average unassigned error of 0.9 kcal mol^{-1}.

We also studied the chemical reactions which are catalyzed by various non-heme iron enzymes and cytochrome P450 enzymes. To understand the activities of various enzymes we have used either density function theory (DFT), full QM/MM simulations or both of them. For example, the activations of L-arginine (L-Arg) by OrfP and VioC have been studied with active site cluster model techniques. The activations of syringol by the GcoA enzyme have been investigated with the help of computational modeling.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements and Dedications

All praises are attributed to creature of the universe "**ALMIGHTY ALLAH**" merciful, the source of all knowledge and wisdom invoked Allah's blessing and peace for "**HAZRAT MUHAMMAD (PUBH)**, the most perfect whose moral and spiritual teachings enlightened my heart and mind. I do not have words to express my feelings by which I can thank my supervisors. I can hardly express my deep gratitude to the kind, extremely sincere and highly cooperative supervisors **Dr. Richard Henchman**, Senior Lecturer, Department of Chemistry, and **Dr. Samuel de Visser**, Senior Lecturer, Department of Chemical Engineering and Analytical Sciences, The University of Manchester for their able guidance, keen interest, skilled advices and ever encouraging attitudes throughout the research work. Both of them are the real source of knowledge and I feel proud to be their PhD student.

I convey my sincere gratitude and love to my family, especially my parents, for being there in times of need and for their encouragement for fulfillment of my research, without them, I would not been able to achieve this goal. A special thanks to my brother and sisters for bearing me and helping me throughout my carrier. In the last, I offer my heartiest gratitude to my friend Jas Kalayan for her cooperation and sincere attitude. I am really grateful for PhD funding Chief Minister Merit Scholarship (CMMS) from the Punjab Education Endowment Fund (PEEF), Pakistan and the School of Chemistry, The University of Manchester. Support is acknowledged from IT Services at the University of Manchester in providing the Computational Shared Facility. There are so many other names in my mind to say a word thanks but space is limited. So, my special thanks are also for those who helped me in achieving this goal, Thank you!

Chapter 1. Introduction

Both stability and kinetics are important parameters for understanding the structures and functions of biomolecules, and are helpful in understanding basic biological phenomena, such as enzyme mechanisms, protein-protein interactions, ligand-protein binding, host-guest binding and so on. The theoretical modelling of biomolecules to study the thermodynamics and the kinetics has gained great attention in the last few decades and has progressed substantially but is still a challenging objective in the field of computational chemistry. In fact, the direct estimation of kinetics and entropy associated properties needs sampling on prolonged time scales, which is quite expensive and limiting in computational viability. Furthermore, various biological processes are governed by entropy, for example, protein folding, protein-ligand binding, and adsorption and desorption processes [1-5]. In these processes the entropy values are calculated indirectly from the free energy calculations [6]. Also, most of the biological processes occur in an aqueous environment and so solvent plays an important role in these processes, which makes it more complicated [7-9]. Enthalpy and energy values have been calculated by means of energy-entropy decomposition methods [10,11] but determination of entropy has been long-known as a key challenge [11,12], not only for computational scientists but also for experimentalists [13].

1.1 Aims of This Work

There are some computational models which are used to directly calculate entropy values for biomolecules and the surrounding solvent but still there is a trade-off between correctness and efficiency. The aim of this project is to develop a general computational method that directly calculates entropy for a wide range of systems including liquids, solutions, chemical reactions etc. In the first step we developed a new method, Multiscale Cell Correlation (MCC) to calculate the entropy values for the range of liquids including the ones that are used as a solvent in biological processes such as water, octanol as well as many other important industrial liquids [14]. Later we combined MCC theory with density functional theory (DFT) in a quantum mechanics/molecular mechanics (QM/MM) formulism to develop a new energy-entropy method called as EE-MCC method to study the kinetics of chemical reactions in aqueous environment [15]. Furthermore, we investigated the metabolic mechanism and reactions kinetics of various enzymes, particularly cytochrome P450 enzymes and non-heme iron or alpha-ketoglutarate enzymes [16-22]. Lastly, we applied our EE-MCC method to investigate the binding Gibbs free energy values of host-guest systems.

1.2 Molecular Dynamics Simulations

In the past half century, simulation methods have been developed with high performance and accuracy [23,24]. One of the most commonly and widely used computational methods is molecular dynamics (MD) simulation. MD simulation is a computer-based method used to determine the physical movements of atoms and molecules [25]. In the system the atoms and molecules are allowed to interact with each other at successive periods of time, producing a trajectory that gives an understanding of the dynamics of a system. The trajectories of these interacting atoms and molecules are determined with the help of Newton's laws of motion and forces are determined from the molecular mechanic force field. Newton's second law of motion for each atom can be written as

$$F_i = m_i a_i \tag{1.1}$$

$$a_i = \frac{d^2 x_i}{dt^2} = -\frac{1}{m_i}\frac{dU}{dx_i} = \frac{F_{x_i}}{m_i} \tag{1.2}$$

where a_i is the acceleration of an atom of mass m_i is calculated by the second derivative of motion along one coordinate x_i with respect to time *t*. The negative gradient of potential energy *U* with respect to particle position is used to calculate the force acting on the atom at a particular position. A continuous potential considers for forces acting on each particle due to change in particle position and positional changes of particles around it. The positions, velocities and accelerations for each particle can be determined by various algorithms. In the AMBER software package, the velocity Verlet algorithm [26] is used to determine the position and velocities of a system at a given interval of time. This algorithm is derived from Taylor's expansion as:

$$r(t + \delta t) = r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) \tag{1.3}$$

$$v(t + \delta t) = v(t) + \frac{1}{2}\delta t[a(t) + a(t + \delta t)] \tag{1.4}$$

where t is the time, r is the position, v is the velocity and a is the acceleration. The implementation of this takes place in two steps: the position is calculated using equation 1.3 and the velocity is calculated at time t from acceleration using the equation

$$v\left(t+\frac{1}{2}\delta t\right)=v(t)+\frac{1}{2}\delta t a(t) \tag{1.5}$$

The acceleration is calculated at time $t + \delta t$ and velocity cycle is completed using

$$v(t+\delta t)=v\left(t+\frac{1}{2}\delta t\right)+\frac{1}{2}\delta t a(t+\frac{1}{2}\delta t) \tag{1.6}$$

When these quantities are evaluated for each atom, the structure of the system is able to be visualised at any time. The molecular averaged properties are computed from the trajectory of the system.

1.3 Force Fields

A force field is a set of functions and parameters that are used to calculate the potential energy of atoms or particles of a system in molecular mechanics (MM) or molecular dynamics (MD) or Monte Carlo (MC) simulations. These functional parameters are calculated experimentally or theoretically or using both together. All-atom force fields consider every atom and calculate parameters of each atom counting hydrogen atoms while united-atom force fields consider hydrogen and carbon as a group and calculate parameters of a molecule i.e. in a methyl group it considers hydrogen and carbon as a one. In this work, we used the Assisted Model Building with Energy Refinement force field. The total potential energy of a system is calculated with the help of force field in terms of bond stretching, bending, torsions and nonbonding interactions [27]. The functional form of potential energy is the sum of bonded interactions of atoms that are linked by covalent bond and non-bonded interactions or non-covalent interactions that are described by electrostatic and van der Waals forces [28]. This decomposition depends on the force field but the general form of potential energy is written as

$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{non-bonded}} \tag{1.7}$$

The covalently bonded bond, angle and dihedral terms are represented as

$$E_{\text{bonded}} = \sum_{\text{bond}} k_{\text{b}}(r - r_{\text{eq}})^2 + \sum_{\text{angle}} k_{\varphi}(\varphi - \varphi_{eq})^2 + \sum_{\text{dihedrals}} \frac{C_{\text{n}}}{2}[1 + \cos(n\phi - \vartheta)] \tag{1.8}$$

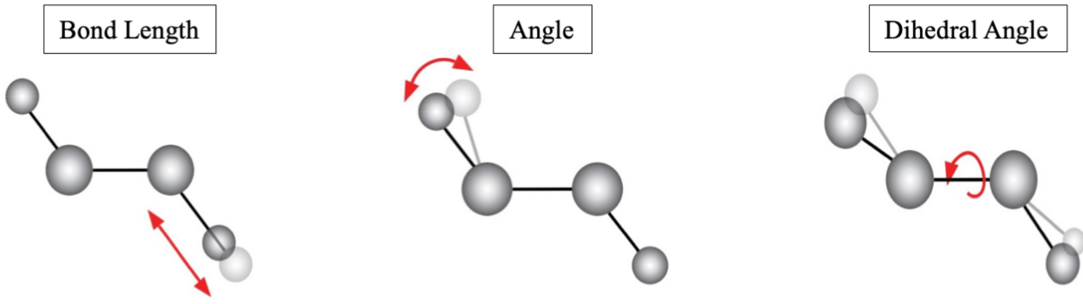


Figure 1.1. The three types of bonded interaction.

where k_{b}, k_{φ} and C_{n} represent the force constants for each motion while r, θ and ϕ represent the bond length, bond angle and dihedral angle respectively. On the other hand, the non-bonded term is computationally intensive and so is limited to pairwise energies typically up to a long-range cutoff and is represented as

$$E_{\text{non-bonded}} = E_{\text{electrostatic}} + E_{\text{van der Waals}} \tag{1.9}$$

The van der Waals interaction is calculated by the Lennard-Jones potential and the electrostatic interaction is calculated by Coulomb's law

$$E_{\text{non-bonded}} = \sum_{ij}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] + \sum_{ij}\left[\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right] \tag{1.10}$$

where r_{ij} is the distance between non-bonded atoms i and j, q_i and q_j are the charges on these atoms and ε_0 is the free space permittivity. The attractive van der Waals interaction between atoms are represented by $B_{ij} = 4\varepsilon\sigma_0^6$ and van der Waals repulsion is represented by $A_{ij} = 4\varepsilon\sigma_0^{12}$ term, where ε is the bonding energy of two atoms and σ_0 is the distance at which the potential energy term between the two particles is zero. These pair-wise calculations depend on the square of the number of particles N^2 that makes the calculations computationally expensive. To solve this problem, a cut-off distance is introduced that limits these interactions within this cut-off range which is about 8 Å to 10 Å and beyond this range these pairwise interactions are approximated by an analytical correction for van der Waals and Particle Mesh Ewald (PME) for electrostatics.

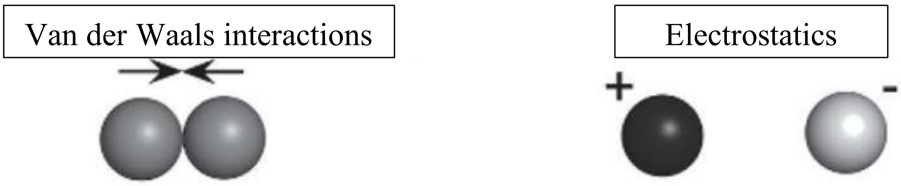


Figure 1.2. The two types of non-bonded interactions.

The General Amber force field (GAFF) [29] and the Optimized Potentials for Liquid Simulations (OPLS) [30] force field are used here. In the GAFF force field, the HF/6-31G* method is used as a charge method while the bonds lengths are derived by X-ray and neutron diffraction as well by as ab-initio calculation at the MP2/6-31G* level of theory. GAFF applies the same empirical rules to calculate the missing parameters of bond-length force constants as the Merck molecular force field (MMFF 94) [31],

$$K_r = K_{ij}^{ref}\left(\frac{r_{ij}^{ref}}{r_{ij}}\right)^6 \tag{1.11}$$

The functional form of GAFF is different than MMFF 94 and it is represented with a more general power law

$$K_r = K_{ij}\left(\frac{1}{r_{ij}}\right)^m \tag{1.12}$$

$$\log K_r = -m\log r_{ij} + \log K_{ij} \tag{1.13}$$

where m and K_{ij} are calculated with the help of bond-length parameters. The K_q force constant can be estimated using the following functional form

$$K_{ijk}^{\theta} = 143.9 \times Z_i C_j \times Z_k\left(r_{ij}^{eq} + r_{jk}^{eq}\right)^{-1} \theta_{ijk}^{eq^{-2}} \exp(-2D) \tag{1.14}$$

$$D = \frac{\left(r_{ij}^{eq} - r_{jk}^{eq}\right)^2}{\left(r_{ij}^{eq} + r_{jk}^{eq}\right)^2} \tag{1.15}$$

In the case of the OPLS force field these parameters are very similar to that of AMBER. The functional form of OPLS force field is only different in the non-bonded term which is similar to Equation 1.8 with the addition of combining rules. The combining rules are $A_{ij} = \sqrt{A_{ii}A_{jj}}$ and $C_{ij} = \sqrt{C_{ii}C_{jj}}$ and the intermolecular non-bonded interaction is counted only for the atoms separated by three or more covalent bonds. There are two charge methods which are mostly used for the derivation of changes in AMBER simulation package: the restrained electrostatic potential (RESP) fitting procedure [32] and Austin model 1-bond charge corrections (AM1–BCC) [33]. The RESP fitting method uses *ab initio* calculation with 6-31G* basis set to generate electrostatic charges (ESP) which are then fitted by different conformations of the molecules, while the AM1–BCC method is a fast and efficient method to generate high quality atomic charges which are used in simulations. It works in two steps: first it captures the electronic features of a molecule by AM1 population charges and then bond charge corrections are applied to AM1 to produce AM1-BCC charges [33]. The RESP method is originally used in the derivation of AMBER force fields while AM1-BCC is used with antechamber tool in the AMBER simulation package to generate these charges.

There are many force fields with good performance and accuracy that have been developed in the literature to study different systems including liquids, solutions, biomolecules etc. by means of MD simulations. In this work we have used GAFF [29], GAFF2 (second-generation GAFF) [34] OPLS-AA (All Atom Optimized Potentials for Liquid Simulations) force field with 1.14*CM1A charges [30] and TraPPE (Transferable Potentials for Phase Equilibria) force field [35] for pure liquids while for protein systems we have used ff14SB [36] and for water TIP3P force field [37] was used. These force fields use different ways of being parameterized but generally their outcome is similar.

1.3.1 Limitations of Classical Mechanics Force Field

The use of classical molecular dynamics simulations still has a number of limitations. The molecular mechanics (MM) force fields requires further improvement in parameters and high computational demands [38,39]. Furthermore, MM is not suitable to study chemical reaction processes since the harmonic potential used for bond stretching and bending which is not suitable for bond cleaving reactions. It also ignores electronic effect. Therefore, to study the chemical reactions we need quantum mechanics which is used to describe the behavior of a system at electronic level [40]. It is used to determine the properties of molecules, atoms and their nuclei [41].

1.4 Quantum Mechanical Calculations

Quantum mechanical (QM) calculations are used to study the behavior of chemical systems and calculate the electronic properties of molecules, atoms and their constituent particles of electrons and nuclei. QM methods are divided into three types: *ab initio*, semiempirical, and density functional theory. Each of these are explained next.

1.4.1 Ab Initio Method

The ab initio or first-principles method is based on the principles of quantum mechanics and is used to solve the electronic Schrödinger equation by utilizing various approximations. Ab initio method do not need any pre-defined potentials as used in MM methods rather than it uses empirical data or independent electronic calculations to define the properties of different complex processes including folding of proteins [42-44]. However, it is an expensive method as compared to the classical MM methods when used in MD simulations because in the ab initio method the forces on each particle printed at each step take much longer to calculate using electronic-structure quantum mechanical methods. The most important ab initio approximations are the Born-Oppenheimer approximation and Hartee-Fock self-consistent field (HF-SCF) or mean-field approximation which are used to solve polyelectronic structural problems with good accuracy [45-47]. In HF method the wave function can be expressed with a Slater determinant (SD) or configuration state functions which are the linear sum of SDs with fixed coefficients [48]. The electronic Schrödinger's equation can be written as

$$\hat{\mathrm{H}}\psi = E\psi \tag{1.16}$$

where $\hat{H}$ is the Hamiltonian operator and it has four electronic energy terms namely kinetic energy of electron $\hat{\mathrm{T}}_{\mathrm{e}}$, electron-nuclear attraction $\hat{\mathrm{V}}_{\mathrm{n-e}}$, electron-electron $\hat{\mathrm{V}}_{\mathrm{e-e}}$ and nuclear-nuclear $\hat{\mathrm{V}}_{\mathrm{n-n}}$ repulsion, written as [45]:

$$\widehat{\mathrm{H}} = \widehat{\mathrm{T}}_{\mathrm{e}} + \widehat{\mathrm{V}}_{\mathrm{n-e}} + \widehat{\mathrm{V}}_{\mathrm{e-e}} + \widehat{\mathrm{V}}_{\mathrm{n-n}} \tag{1.17}$$

The Schrödinger is solved using the variational principle which states that the approximate wave function will always be higher in energy than the exact wave function so the best wave function can be calculated by minimizing the energy [45].

1.4.2 Semiempirical Method

Semiempirical quantum methods (SE-QM) are in between QM and MM methods because they treat electrons quantum mechanically but only up to valence electrons to speed up QM calculations at reduced computational cost. This makes SE-QM methods fast and applicable for understating metabolic process like kinetics study of enzymes and proteins and with good accuracy [49]. There are many SE-QM methods available in the literature including AM1 [50], PM3 [51], OMx [52] and MNDO/d [53]. Another approach that has gained attention in the last few decades is the density functional tight binding (DFTB) method, also known as Self-Consistent Charge DFTB (SCC-DFTB) [54-56]. SCC-DFTB [56,57] is an approximate semi-empirical method derived from the DFT method by considering the 2nd order expansion of total energy functional of the DFT with respect to the fluctuations of charge density $\delta\rho$ around a reference density ρ_o.

$$E^{\mathrm{SCC-DFTB}} = \sum_i^{\mathrm{occ}} \langle \psi_i | \widehat{\mathrm{H}}_0 | \psi_i \rangle - \frac{1}{2} \iint' \frac{\rho_o \rho_o'}{|\vec{r}-\vec{r}'|} + E_{xc}[\rho_o] - \int V_{xc}[\rho_o]\rho_o + E_{ii} + \frac{1}{2} \iint' \left(\left| \frac{1}{|\vec{r}-\vec{r}'|} + \frac{\delta^2 E_{xc}}{\delta\rho\delta\rho'} \right| \right) \delta\rho\delta\rho' \tag{1.18}$$

Here $\delta\rho = \rho - \rho_o$ is the charge fluctuation which is the second order term represented by atomic components $\delta\rho = \sum_\alpha \delta\rho_\alpha$ and the approximation of $\delta\rho_\alpha$ is done by atomic charge fluctuations $\Delta q_\alpha = q_\alpha - q_\alpha^0$. These atomic fluctuations are calculated by Mulliken charge analysis. The exchange term E_{xc} is approximated by a function γ which depends on the chemical hardness η_α, leading to the second order term which is

$$E^{\mathrm{2nd}} = \frac{1}{2} \sum_{\alpha\beta} \gamma_{\alpha\beta} \Delta q_\alpha \Delta q_\beta \tag{1.19}$$

where $\gamma_{\alpha\beta} = \gamma_{\alpha\beta}(U_\alpha, U_\beta, U_{\alpha\beta})$ and $U_\alpha = \frac{1}{2}\frac{\partial^2 E_{\mathrm{at}}}{\partial q_{\mathrm{at}}^2}$ is the second derivative of energy of an atom α with respect to its total charge, and ψ_i are the Kohn-Sham orbitals which are expanded in an optimized linear combination of atomic orbital (LCAO) basis set ϕ_μ as suggested by Eschrig and Bergert [58], $\psi_i = \sum_\mu c_\mu^i \phi_\mu$ and $\widehat{\mathrm{H}}_0$ is Hamiltonian matrix solved by Kohn-Sham theory.

1.4.3 Density Functional Theory

Density Functional Theory (DFT) [59-61] is used to determine the properties of many-electron system with functionals. It is based on two theorems: the Hohenberg-Kohn [60] and the Kohn-Sham formalism [61]. According to the Hohenberg-Kohn theorem, the total energy of a system composed of fixed nuclei and interacting electrons at ground state is a functional $E\ (\rho)$ of the electron density function $\rho(r)$ the variational principle to any trial density of a system will result in energy higher than or equal to the exact ground state energy

$$E\ [\rho'(r)]\ \geq E\ [\rho(r)] \tag{1.20}$$

Form the Born-Oppenheimer approximation [62] of the Schrödinger wave equation, the total energy of a system is expressed as the combination of three different terms, which are the kinetic energy of electrons (T), nuclear electron attraction (V_{ne}) and the electron- electron repulsion (V_{ee}). It is represented as

$$\widehat{H}\Psi(r_1, r_2, \ldots r_N) = E\Psi(r_1, r_2, \ldots r_N) \tag{1.21}$$

$$E[\rho(r)] = T[\rho(r)] + V_{\mathrm{ne}}[\rho(r)] + V_{\mathrm{ee}}[\rho(r)] \tag{1.22}$$

The electron- electron repulsion consist of two components: one is the classical Coulomb repulsion while the other is composed of all non-classical contributions to the electron – electron repulsion.

According to the Kohn-Sham formalism, the kinetic energy and electron density comes from the orbitals (ϕ_i) as shown in Equation. 1.23 and Equation 1.24.

$$T_s[\rho] = -\tfrac{1}{2}\textstyle\sum_i^N \langle \phi_i | \nabla^2 | \phi_i \rangle \tag{1.23}$$

Here T_s is not a true kinetic energy but the energy of a system of non-interacting electrons, which reproduce the density of a system at ground state.

$$\rho(r) = \sum_i^N |\phi_i|^2 \tag{1.24}$$

The total energy of a system at ground state according to DFT method is

$$E[\rho] = T_\mathrm{s}[\rho] + V_\mathrm{ext}[\rho] + V_H[\rho] + E_\mathrm{xc}[\rho] \tag{1.25}$$

where $V_\mathrm{ext}[\rho]$ represents the electromagnetic interaction of the electron density with the external potential, $V_H[\rho]$ as the classical Coulomb interaction, $T_\mathrm{s}[\rho]$ approximates the kinetic energy of electrons and $V_\mathrm{xc}[\rho]$ is the exchange correlation functional which corrects the first three terms in Equation 1.25. Though DFT can be an accurate method, only approximate solutions to the electronic energy can be attained due to the lack of exact expression for the E_xc term. For this purpose various exchange correlations have been developed [63].

The Becke, 3-parameter, Lee–Yang–Par (B3LYP) [64-66] is one of the most popular and most commonly used hybrid functionals [67], which we have employed during the quantum mechanical calculations presented in this work. The exact exchange energy E_{xc} play an important role to accurate the density functional theory and is calculated as.

$$E_{xc}^{B3LYP} = E_x^{LDA} + a_o(E_x^{HF} - E_x^{LDA}) + a_x(E_x^{GGA} - E_x^{LDA}) + E_c^{LDA} + a_c(E_c^{GGA} - E_c^{LDA}) \tag{1.26}$$

where a_o, a_x and a_c are semiemprical constants determined by a suitable fitting to experimental data and E_x^{LDA} is the exchange functional according to the local-spin density correlation (LSD) [68]. E_x^{HF} is the exact Hartree-Fock (HF) exchange and E_x^{GGA} is the gradient correlation proposed by Becke [64,69]. B3LYP is a hybrid functional of local exchange correlation functional which overestimate the atomization energy and Hartree-Fock method which underestimate this atomization energy. It sits in above the local density approximation (LDA) and generalised gradient approximation (GGA) in the Jacob's ladder interpretation of DFT method. The inclusion of HF exchange in hybrid functional increases the accuracy of this DFT method because it includes partial self-corrections and thus gives a description of static correlation [70]. This makes the hybrid functional more accurate to study the spin states of transition metals because bonding of transition metal complexes involves considerable static correction. However, there are some problems that affect the accuracy of DFT method. First of all, the pure dispersion interaction between unbound chemical species are not well reproduced by these functionals. Second is the poor elimination of the electron self-interaction and exchange energies. Third is the large energy errors for an important species even dispersion and self-interactions are not involved [71,72]. For example, in chapter 4 we have applied different hybrid functionals including B3LYP, PBE0 and B3LYP-D3 to study the dihydroxylation of the arginine substrate by a nonheme iron enzyme (OrfP) and we find that B3LYP with dispersion correction (B3LYP-D3) gives more accurate activation energy barriers close to experiment than other methods. In B3LYP hybrid functional with 25% of exact exchange contribution would be expected to favor higher spin ground state, for example during the dihydroxylation of arginine by a nonheme iron enzyme (OrfP) shows that quintet state spin state is the ground state and is well separated from the triplet spin state by $\Delta E + \text{ZPE} = 7$ kcal mol^{-1} which is in-line with experimentally characterized quintet spin state for nonheme iron enzymes [20-21].

1.4.4 Quantum Mechanical-Cluster Technique

QM calculation show slow convergence and are quite expensive in terms of computational cost with increase in system size. This limits the QM calculations to small systems study and it is hard to implement them on larger systems of thousands of atoms such as reactions in solvent, proteins and enzymes mechanisms. To study the QM calculation in solution, phase two type of solvent techniques have been used: implicit solvent models and explicit solvent models. In implicit solvent models the solvent is treated as a structureless continues medium with a specific dielectric constant and interfacial properties [73,74]. Thus, the number of interacting particles and degree of freedoms of a system is significantly reduces. Implicit solvent models include the polarizable continuum model (PCM) [75], the conductor like polarizable continuum model (CPCM) [76] the Onsager model [77] and the universal solvent model (SMD) [78]. The CPCM and PCM are the most common and widely used implicit solvent models. Both of these models define the cavities as envelopes of spheres around the atom and molecules. The dielectric constant inside these cavities remains the same as in the gas phase while outside it is different, depending on the desired solvent. Although the implicit solvent models are less expensive computationally and help to improve sampling, they are not giving the real picture of solvent molecules [79]. On the other hand, the explicit solvent models include individual solvent molecules, giving a more realistic picture where direct solvent-solvent interactions are observed [80,81]. The solvent molecules in explicit solvent models are defined by molecular mechanics force fields such as the TIP3P force field [33]. Inclusion of explicit solvent molecules makes the calculations quite expensive. However, this problem is partly addressed by using two different techniques, namely the QM-cluster model and a mixed quantum mechanics / molecular mechanics method, which are described below [82].

The quantum mechanical (QM) cluster or only-QM model is the first method used to study the reactivity and structure of fairly small systems and also more complex systems including enzyme catalytic processes. The QM-cluster approach is used to understand the enzymatic reaction mechanisms of the active site along with relevant surrounding amino-acid residues. The QM-cluster model is built from an already available X-ray crystal structural [83-85] but we can only consider a limited number of atoms about few hundred and not the whole enzyme which consists of thousands or hundred thousand atoms. The first QM-cluster model was used by Siegbahn and Crabtree to study methane hydroxylation by a non-heme iron monooxygenase enzyme [86]. This QM-cluster model consists of 20 atoms and was treated with DFT using B3LYP/6-311+G level. Later on, they studied heme peroxidases with 67 atoms in the QM-cluster model [87] and the mechanism of nitrogenase enzyme with QM-cluster model having 200 atoms treated at the same level of DFT. Recently, work on the mechanism of hydration and carboxylation activities of phenolic acid has been done which had up to 312 atoms in the QM-cluster model at the B3LYP/6-311+G (2d,2p) level of DFT [88,89]. Since the whole enzyme is not considered in QM-cluster model some other environmental effects like short and long-range electrostatics interactions, steric and polarization effects have been ignored.

1.4.5 Quantum Mechanics / Molecular Mechanics method

The quantum mechanical / molecular mechanical (QM/MM) method is a hybrid method used to study the chemical reaction in solutions and other complex reactions including enzymatic reaction [90]. In QM/MM the system is divided into two parts: one is the QM region, where the chemical reaction takes place and is treated with the more accurate quantum mechanical method the second is the MM region, the surrounding part of the QM region which is treated with the less expensive force field method [91]. The former region only consists of a small number of atoms while the latter region covers the larger area of the system. The schematic representation of hybrid QM/MM method is shown in Figure 1.3.

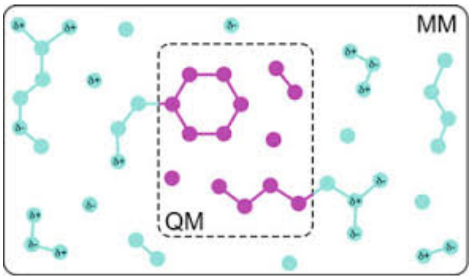


Figure 1.3. QM/MM regions of a system [92].

There are two approaches used to combine the QM and MM parts: one is the additive scheme [93] and the second is the subtractive scheme [94]. In the first approach, the total energy of a system is represented as the sum of the QM energy terms (E_{QM}), MM the energy terms (E_{MM}) and the energy of QM/MM coupling terms ($E_{\mathrm{QM/MM}}$).

$$E_{\mathrm{tot}} = E_{\mathrm{QM}} + E_{\mathrm{MM}} + E_{\mathrm{QM/MM}} \tag{1.27}$$

The second approach is the subtractive scheme which is also known as integrated molecular orbitals molecular mechanics. It divides the system into two layers and subtracts the double counted energy of the smaller layer as:

$$E_{\mathrm{tot}} = E_{\mathrm{real}}^{\mathrm{MM}} + E_{\mathrm{model}}^{\mathrm{QM}} - E_{\mathrm{model}}^{\mathrm{MM}} \tag{1.28}$$

$E_{\mathrm{real}}^{\mathrm{MM}}$ represents the energy of the whole real system, $E_{\mathrm{model}}^{\mathrm{QM}}$ represents the energy of the QM part while $E_{\mathrm{model}}^{\mathrm{MM}}$ represents the energy of MM part. The main advantage of the subtractive scheme is that no communication is required between quantum mechanics and molecular mechanics which makes its implementation relatively straightforward. The major disadvantage of this method is that the force filed parameters are required for the QM part which are not always available. Additionally, the force field should be flexible to describe the effect of chemical changes during the chemical reactions. Furthermore, this method also ignores the polarization effect of QM part by the MM environment. These problems are addressed in the additive scheme where the interactions in the MM environment are described by a force field and the polarization effect of MM environment on QM region by the electrostatic embedding method [92].

1.5 Free Energy Methods

The Gibbs free energy is the principal quantity in thermodynamics under ambient conditions which drives the vast majority of chemical processes in nature such as chemical reactions, protein folding and protein-ligand binding [95]. The calculation of free energy has been a main longstanding goal for computational and theoretical chemists because it enables one to model chemical reaction and design new materials and drugs against diseases efficiently [96]. The recent development in new methods, algorithms and advancement in technologies has improved the efficiency and accuracy of free energy calculations. There are many methods available in the literature which are used to calculate the free energy values including alchemical free energy methods such as free energy perturbation (FEP) [97], thermodynamic integration (TI) [98] and the Bennett acceptance ratio (BAR) [99], which calculate the free energy difference between two states along a reaction coordinate. In these methods the reaction pathway is divided into small intermediate steps which helps them to achieve the convergence between initial and final stages. However, these methods are quite expensive in terms of computational cost and time to produce the free energy difference [98]. While the BAR method uses the method.

The Weighted histogram analysis method (WHAM) [100] is another method used along with the umbrella sampling technique to calculate the free energy and potential of mean force (PMF) for bimolecular simulations. The PMF is calculated along the reaction coordinates by using umbrella sampling which divide the whole reaction profile into a number of windows. This method allows the multiple overlaps probability distribution and gives the better free energy value estimation. Furthermore, it accounts for all the simulations along the reaction pathway to calculate the free energy values which reduces the statistical error. However, it still has convergence problems which make it computationally expensive [100,101].

The other free energy methods are energy-entropy (EE) decomposition methods such as the Molecular Mechanics Poisson-Boltzmann and Generalised Born Surface Area continuum solvation models known as MM/PBSA [102] and MM/GBSA [103], respectively. These are popular methods to determine the binding free energies of small ligands with complex biomolecules such as proteins and help in structure-based drug design. These are based on MD simulations of ligand-protein structures so they are intermediate in term of accuracy and computational cost than FEP and TI methods [104]. These methods calculate the energy values directly from the force field in the MD simulation but for entropy calculation they need additional methods which are not so straightforward. To calculate the entropy, various methods have been proposed which are as follows.

1.6 Entropy Calculations

Calculating and understanding entropy values of liquids and solutions is an important challenge because many processes in nature take place in the liquid phase. There are several approaches available in the literature to calculate entropy values [105-113]. Amongthe most common methods are normal mode analysis (NMA) [114] and quasi-harmonic analysis (QHA) [115] which are based on the rigid-harmonic oscillator (RRHA) approximation [116]. Other entropy methods are 2-Phase thermodynamic (2PT) [117], the minimal spanning tree (MIST) variant [118] and mutual information [119].

1.6.1 Normal Mode Analysis

Normal Mode Analysis (NMA) [114] is one of the most commonly used methods to calculate the entropy of systems including liquids and biomolecules. In NMA the entropy is calculated from the vibrational frequencies which are derived from multidimensional Gaussian distribution using the eigenvalues of Hessian matrix at selected energy minima. The Hessian matrix may be articulated in term of normal coordinates using orthonormal transformation matrix by diagonalizing Hessian matrix [120]. The eigenvalues from these matrixes represents the force constant which is used to calculate the vibrational frequencies using

$$\nu_\mathrm{i} = \frac{1}{2\pi}\sqrt{\lambda_\mathrm{i}} \tag{1.29}$$

where ν_i are the vibration frequencies and λ_i are the eigenvalues of the Hessian. The vibrational frequency is used in quantum harmonic oscillator equation to calculate the vibration entropy as.

$$S_\mathrm{vib} = k_\mathrm{B} \sum_{\mathrm{i}=1}^{3\mathrm{N}} \left(\frac{h\upsilon_i/k_\mathrm{B}T}{\mathrm{e}^{h\upsilon_i/k_\mathrm{B}T}-1} - \ln\left(1 - \mathrm{e}^{-h\upsilon_i/k_\mathrm{B}T}\right) \right) \tag{1.30}$$

NMA somewhat underestimates the entropy values compared to experiment because it derives frequency values from the energy minimum of the system. Also, NMA in principle needs to be done at every minimum to calculate entropy, which is impractical for larger systems like biomolecules that make NMA limited to small molecules only [121,122].

1.6.2 Quasiharmonic Analysis

Quasiharmonic Analysis (QHA) [115] is another method which is widely used to calculate the vibrational frequency and absolute entropy values from multidimensional Gaussian probability distribution but makes use of the internal coordinates of atoms relative to their average positions in an MD simulation [123-125]. The entropy is calculated from the using

$$\nu_{ij} = k_{\mathrm{B}} T \sigma_{ij}^{-1} \tag{1.31}$$

where ν_{ij} are the generalised force constants k_{B} is Boltzmann's constant, and σ is the mass-weighted covariance matrix of coordinate fluctuations. These force constants are substituted into Equation 1.29 and the resulting frequencies are put into Equation 1.30 to calculate the absolute entropy. However, for larger system having many degrees of freedoms, such as biomolecules QHA leads to the overestimation of entropy values because of the anharmonicity of coordinates. Furthermore, the other disadvantage of QHA is that the system convergence is very slow due to the noise in the off-diagonal correlations [126,127] but this noise can be reduced by using internal coordinates instead of Cartesian coordinates [128] or a von Mises distribution [129].

1.6.3 Force Covariance Method

The overestimation of QHA method is addressed by the force covariance (FC) method which quantifies the shape of potential energy surface of a particle's harmonic motion from forces rather than coordinates by using the relationship $F = -k\Delta q$ [112]. The average potential energy of the system in the harmonic approximation is calculated using

$$\langle U \rangle = \frac{1}{2} k_{\mathrm{B}} T = \frac{1}{2} k \langle \Delta q^2 \rangle = \frac{1}{2} \frac{\langle F^2 \rangle}{k} \tag{1.32}$$

where $\langle U \rangle$ is the average potential of a harmonic oscillator, *k* is the force constant and $\Delta q = q - \langle q \rangle$ is the particle's position q minus the average position $\langle q \rangle$. By converting the mass weighted coordinates $q' = m^{1/2}q$ and forces $F' = m^{1/2}F$ and substituting the angular frequency ν for a harmonic oscillator Equation 1.30 becomes

$$\langle U \rangle = \frac{1}{2}\nu^2\langle \Delta q'^2 \rangle = \frac{1}{2}\frac{\langle F'^2 \rangle}{\nu^2} \qquad (1.33)$$

For a system of *N* atoms, a mass weighted forces covariance matric can be constructed as

$$\sigma_{ij}^F = \langle F_i' F_j' \rangle \qquad (1.34)$$

By diagonalizing this matrix, 3*N* eigenvalues are obtained which leads to the calculation of force derived angular frequency using

$$\nu_i^F = \frac{1}{2\pi}\sqrt{\frac{\lambda_i^F}{k_{\mathrm{B}}T}} \qquad (1.35)$$

where ν_i^F is the force derived frequency which is used in Equation 1.30 to calculate the vibration entropy values [130]. The FC method sits in between under and over estimation entropy values by NMA and QHA methods.

1.7 Overview of Chapters in the Thesis

The main results of this thesis are in five chapters which are summarized below.

Chapter 2. Entropy of Simulated Liquids Using Multiscale Cell Correlation, describes the multiscale cell correlation (MCC) method to calculate the entropy of pure liquids from molecular dynamics (MD) simulations. The method uses forces and torques from MD simulations at two level of length scales i.e. molecular level and united atom level and probability distributions of molecular coordinations and conformations. MCC is applied broader set of fifty-six (56) important industrial liquids modeled using Generalized AMBER Force Field (GAFF) and Optimized Potentials for Liquid Simulations (OPLS) force fields with 1.14*CM1A charges. The entropy values obtained from MCC are compared with experimental values and obtained unsigned errors which are 8.7 J K^{-1} mol^{-1} and 9.8 J K^{-1} mol^{-} for GAFF and OPLS respectively. This is significantly better than the 2-Phase Thermodynamics method for the subset of molecules in common, which is the only other method that has been applied to such systems. MCC makes clear why the entropy has the value it does by providing a decomposition in terms of translational and rotational vibrational entropy and topographical entropy at the molecular and united-atom levels.

Chapter 3. Comparison of Free-Energy Methods to Calculate the Barriers for the Nucleophilic Substitution of Alkyl Halides by Hydroxide, describes a new proposed energy-entropy (EE) method to calculate the Gibbs free energy of reactants and transition states of S_N2 type reaction in explicit solvent by combining with quantum mechanics/molecular mechanics (QM/MM) molecular dynamics simulations with MCC. The EE-MCC method is applied to six nucleophilic substitution reactions of the hydroxide transfer to methyl and ethyl halides in water, where the halides are F, Cl, and Br. The Gibbs free energy values are calculated using three methods i.e. EE-MCC, EE-NMA, PMF and with two Hamiltonian i.e. self-consistent charge density functional based tight-binding (SCC-DFTB), B3LYP/6-31+G* and M06/6-31+G* density functional theory (DFT), potential of mean force (PMF) in explicit water solvent. The EE-NMA values are also calculated using B3LYP/6-31+G* and M06/6-31+G* in implicit solvent water model while the experimental values are derived via transition state theory (TST). The barriers using SCC-DFTB are found to agree well with the PMF and experiment and previous computational studies, being slightly higher but improving on the lower values obtained for the implicit solvent.

Chapter 4. What determines the selectivity of arginine dihydroxylation by the nonheme iron enzyme OrfP? Describes the use of molecular dynamics simulation and quantum mechanical cluster model techniques with density functional theory (DFT) to study the dihydroxylation reaction mechanism of L-Arginine catalyzed by a nonheme iron enzyme OrfP. The OrfP reacts with L-Arg selectively to form the 3*R*,4*R*-dihydroxyarginine product, which in mammals can inhibit the nitric oxide synthase enzymes involved in blood pressure control. We show that substrate binding and positioning in the active site guides a highly selective reaction through C^3–H hydrogen atom abstraction. This happens despite the fact that the C^3–H and C^4–H bond strengths of L-Arg are very similar. Electronic differences in the two hydrogen atom abstraction pathways drive the reaction with an initial C^3–H activation to a low-energy $^5\sigma$-pathway, while substrate positioning destabilizes the C^4–H abstraction and sends it over the higher-lying $^5\pi$-pathway. We show that substrate and monohydroxylated products are strongly bound in the substrate binding pocket and hence product release is difficult and consequently its lifetime will be long enough to trigger a second oxygenation cycle.

Chapter 5. Lignin biodegradation by a cytochrome P450 enzyme: A computational study into syringol activation by GcoA, uses the large size quantum mechanical cluster model with density functional theory to study the lignin biodegradation reaction mechanism catalyzed by an isozyme of cytochrome P450 known as GcoA. DFT study of GcoA to investigate syringol activation by an iron(IV)-oxo heme cation radical oxidant (Compound I) leading to hemiacetal and acetal products. Several substrate-binding positions were tested and full energy landscapes calculated. The study shows that substrate positioning determines the product distributions. Thus, with the phenol group pointing away from the heme, an *O*-demethylation is predicted, whereas an initial hydrogen-atom abstraction of the weak phenolic O–H group would trigger a pathway leading to ring-closure to form acetal products. Predictions on how to engineer P450 GcoA to get more selective product distributions are also given.

Chapter 6. Energy-Entropy Method Using Multiscale Cell Correlation to Calculate Host-Guest Free Energies of Binding, uses the energy-entropy method with MCC to calculate the binding Gibbs free energies of host-guest complexes. The seven guest molecules "drug of abuse" and a host molecule cucurbit [8]uril (CB8), which can serve as a drug carrier are taken from the SAMPL8 challenge. The binding Gibbs free energy values for these host-guest systems were calculated by using our energy-entropy method, where energy and entropy are calculated directly from molecular dynamics simulations. The mean average error of calculated binding free energy versus experimental values is 0.15 kcal mol^{-1}. MCC yields a value of the entropy of the system and a decomposition over molecules, level, motion and minima.

1.8 References

[1] Karplus, M., Behind the folding funnel diagram. *Nat. Chem. Biol.* **2011,** *7*, 401–404.

[2] Zhou, H.X. and Gilson, M.K., Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* **2009,** *109*, 4092–4107.

[3] Wand, A.J., Moorman, V.R. and Harpole, K.W., A surprising role for conformational entropy in protein function. *Top. Curr. Chem.* **2013,** *337*, 69–94.

[4] Weaver, J. F., Chemistry. Entropies of adsorbed molecules exceed expectations. *Science* **2013,** *339*, 39–40.

[5] Campbell, C.T. and Sellers, J.R.V., The entropies of adsorbed molecules. *J. Am. Chem. Soc.* **2012,** *134*, 18109–18115.

[6] Chipot C, Pohorille A, eds. *Free Energy Calculations*, vol. 85. Berlin Heidelberg: Springer-Verlag; 2007.

[7] Simmerling, C., Strockbine, B. and Roitberg, A. E., All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002,** *124*, 11258–11259.

[8] Juraszek, J. and Bolhuis, P.G., Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Natl. Acad. Sci.* **2006,** *103*, 15859.

[9] Snow, C.D., Zagrovic, B. and Pande, V.S., The Trp Cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.* **2002,** *124*, 14548–14549.

[10] Andrés, J.; Ayers, P. W.; Boto, R. A.; Carbó-Dorca, R.; Chermette, H.; Cioslowski, J.; Contreras-García, J.; Cooper, D. L.; Frenking, G.; Gatti, C.; Heidar-Zadeh, F.; Joubert, L.; Martín Pendás, Á.; Matito, E.; Mayer, I.; Misquitta, A. J.; Mo, Y.; Pilmé, J.; Popelier, P. L. A.; Rahm, M.; Ramos-Cordoba, E.; Salvador, P.; Schwarz, W. H. E.; Shahbazian, S.; Silvi, B.; Solà, M.; Szalewicz, K.; Tognetti, V.; Weinhold, F.; Zins, É.-L., Nine questions on energy decomposition analysis. *J. Comput. Chem.* **2019,** *40*, 2248-2283.

[11] Brady, G.P. and Sharp, K. A., Entropy in protein folding and in protein-protein interactions. *Curr. Opin. Struct. Biol.* **1997,** *7*, 215–21.

[12] Suárez, D. and Díaz, N., Direct methods for computing single-molecule entropies from molecular simulations. *WIREs Comput. Mol. Sci.* **2015,** *5*, 1–26.

[13] Wand, A. J., The dark energy of proteins comes to light: conformational entropy and its role in protein function revealed by NMR relaxation. *Curr. Opin. Struct. Biol.* **2013,** *23*, 75–81.

[14] Ali, H.S., Higham, J. and Henchman, R.H., Entropy of simulated liquids using multiscale cell correlation. *Entropy* **2019,** *21*, 750.

[15] Ali, H.S., Higham, J., de Visser, S.P. and Henchman, R.H., Comparison of free-energy methods to calculate the barriers for the nucleophilic substitution of alkyl halides by hydroxide. *J. Phys. Chem. B* **2020,** *124*, 6835-6842.

[16] Ali, H.S., Henchman, R.H. and de Visser, S.P., Lignin biodegradation by a cytochrome P450 enzyme: A computational study into syringol activation by GcoA. *Chem. Eur. J.* **2020,** *26*, 13093-13102.

[17] Ali, H.S., Henchman, R.H. and de Visser, S.P., Cross-linking of aromatic phenolate groups by cytochrome P450 enzymes: A model for the biosynthesis of vancomycin by OxyB. *Org. Biomol. Chem.* **2020,** *18*, 4610-4618.

[18] Chowdhury, A.S., Ali, H.S., Faponle, A.S. and de Visser, S.P., How external perturbations affect the chemoselectivity of substrate activation by cytochrome P450 OleTJE. *Phys. Chem. Chem. Phys.* **2020,** *22*, 27178-27190.

[19] Louka, S., Barry, S.M., Heyes, D.J., Mubarak, M.Q.E., Ali, H.S., Alkhalaf, L.M., Munro, A.W., Scrutton, N.S., Challis, G.L. and de Visser, S.P., Catalytic mechanism of aromatic nitration by cytochrome P450 TxtE: Involvement of a ferric-peroxynitrite intermediate. *J. Am. Chem. Soc.* **2020,** *142*, 15764-15779.

[20] Ali, H.S., Henchman, R.H. and de Visser, S.P., What determines the selectivity of arginine dihydroxylation by the nonheme iron enzyme OrfP? *Chem. Eur. J.* **2021,** *27*, 1795-1809.

[21] Ali, H.S., Henchman, R.H., Warwicker, J. and de Visser, S.P., How do electrostatic perturbations of the protein affect the bifurcation pathways of substrate hydroxylation versus desaturation in the nonheme iron-dependent viomycin biosynthesis enzyme? *J. Phy. Chem. A* **2021,** *125*, 1720-1737.

[22] Han, S.B., Ali, H.S. and de Visser, S.P., Glutarate hydroxylation by the carbon starvation-induced protein D: A computational study into the stereo- and regioselectivities of the reaction. *Inorg. Chem.* **2021,** *60*, 4800-4815.

[23] Hünenberger, P.H. and van Gunsteren, W.F., Computer Simulation of Biomolecular systems: Theoretical and Experimental Applications, Eds. Springer Netherlands: Dordrecht, 1997, 3–82.

[24] van Gunsteren, W.F. and Mark, A.E., Validation of molecular dynamics simulation. *J. Chem. Phys.* **1998,** *108*, 6109–6116.

[25] Comstock, M.J., Molecular-based study of fluids, copyright, advances in chemistry Series, Forward, About the Editors, in Molecular-Based Study of Fluids, M.J. Comstock, Editor. **1983**, Am. Chem. Soc. 1–7.

[26] Verlet, L., Computer "Experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **1967**, 159, 98–103.

[27] Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., DeBolt, S., Ferguson, D., Seibel, G. and Kollman, P., AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995,** *91*, 1–41.

[28] Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P. A., A Second-generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995,** *117*, 5179–5197.

[29] Wang, J.M., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[30] Jorgensen, W.L., D.S. Maxwell, and Tirado-Rives, J., Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, 118, 11225−11236.

[31] Halgren, T.A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Compu. Chem*. **1996**, 17, 490−519.

[32] Cieplak, P., Cornell, W.D., Bayly, C. and Kollman, P.A., Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J. Compu. Chem.* **1995,** 16, 1357-1377.

[33] Jakalian, A., Jack, D.B. and Bayly, C.I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Compu. Chem.* **2002,** 23, 1623-1641.

[34] Träg, J. and Zahn, D., Improved GAFF2 parameters for fluorinated alkanes and mixed hydro- and fluorocarbons. *J. Mol. Model.* **2019,** *25*, 39.

[35] Potoff, J.J. and Siepmann, J.I., Vapor-liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AICHE J.* **2001**, *47*, 1676–1682.

[36] Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. Simmerling, C., ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015,** *11*, 3696−3713.

[37] Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

[38] Chodera, J.D., Mobley, D.L., Shirts, M.R., Dixon, R.W., Branson, K. and Pande, V.S., Alchemical free energy methods for drug discovery: progress and challenges. *Curr. Opin. Struct. Biol.* **2011,** *21*, 150−160.

[39] Durrant, J.D. and McCammon, J.A., Molecular dynamics simulations and drug discovery. *BMC Biology* **2011,** *9*, 71.

[40] Jaeger, G. What in the (quantum) world is macroscopic? *Am. J. Phys.* **2014,** *82*, 896−905.

[41] Dirac, P.A.M. The basis of statistical quantum mechanics. *Math. Proc. Cambridge Philos. Soc.* **1929,** *25*, 62−66.

[42] Skolnick, J., Putting the pathway back into protein folding. *Proc. Natl. Acad. Sci. USA.* **2005,** *102*, 2265−2266.

[43] Schueler-Furman, O., Wang, C., Bradley, P., Misura, K. and Baker, D., Progress in modeling of protein structures and interactions. *Science* **2005,** *310*, 638−42.

[44] Bonneau, R. and Baker, D., Ab initio protein structure prediction: progress and prospects. *Annu. Rev. Biophys Biomol. Struct.* **2001,** *30*, 173−89.

[45] Jensen, F., Chapter 1 An introduction to the state of the art in quantum chemistry. *Annu. Rep. Comput. Chem.* **2005**, *1*, 3−17.

[46] Morozov, A. V., Kortemme, T., Tsemekhman, K. and Baker, D., Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proc. Natl. Acad. Sci. USA.* **2004,** *101*, 6946−6951.

[47] Barden, C. J. and Schaffer III, H. F. Quantum chemistry in the 21st century. *Pure. Appl. Chem.* **2000**, 72, 1405–1423.

[48] Slater, J.C., The Theory of Complex Spectra. *Phys. Rev.* **1929,** *34*, 1293−1322.

[49] Pan, Y.K., Approximate molecular orbital theory. *J. Chem. Educ.* **1971,** *48*, A116.

[50] Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P., Development and use of quantum mechanical molecular models. AM1: A new general-purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985,** *107*, 3902−3909.

[51] Stewart, J.J.P., Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **1989,** *10*, 221−264.

[52] Thiel, W. and Voityuk, A.A., Extension of MNDO to d orbitals: Parameters and results for the second-row elements and for the zinc Group. *J. Phys. Chem.* **1996,** *100*, 616−626.

[53] Dral, P.O., Wu, X. and Thiel, W., Semiempirical quantum-chemical methods with orthogonalization and dispersion corrections. *J. Chem. Theory Comput.* **2019,** *15*, 1743−1760.

[54] Elstner, M., Frauenheim, T., Kaxiras, E., Seifert, G. and Suhai, S., A self-consistent charge density-functional based tight-binding scheme for large biomolecules. *Phys. Status Solidi* **2000,** *217*, 357−376.

[55] Elstner, M. and Seifert, G., Density functional tight binding. *Philos. Trans. R. Soc. A: Math. Phy. Engin. Sci.* **2014,** *372*, 20120483.

[56] Gaus, M., Cui, Q. and Elstner, M., Density functional tight binding: application to organic and biological molecules. *WIREs Comput. Mol. Sci.* **2014,** *4*, 49−61.

[57] Elstner, M., The SCC-DFTB method and its application to biological systems. *Theor. Chem. Acc.* **2006**, 116, 316−325.

[58] Eschrig, H. and I. Bergert, An optimized LCAO version for band structure calculations application to copper. *Phys. Status solidi B,* **1978**, 90, 621−628.

[59] Dirac, P.A.M. The basis of statistical quantum mechanics. *Math. Proc. Cambridge Philos. Soc.* **1929,** *25*, 62−66.

[60] Hohenberg, P. and W. Kohn, Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, 864−871.

[61] Kohn, W. and L.J. Sham, Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, 140, 1133−1138.

[62] Calais, J., Density-functional theory of atoms and molecules. R.G. Parr and W. Yang, Oxford University Press, New York, Oxford, **1989**.

[63] Gonis, A., The pursuit of fallacy in density functional theory: The quest for exchange and correlation, the rigorous treatment of exchange in the Kohn-Sham formalism and the continuing search for correlation. *J. Microbiol. Biotechnol.* **2014**, *04*, 26.

[64] Becke, A.D., Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993,** *98*, 5648-5652.

[65] Becke, A.D., Density-functional thermochemistry. I. The effect of the exchange-only gradient correction. *J. Chem. Phys.* **1992,** *96*, 2155-2160.

[66] Lee, C., W. Yang, and R.G. Parr, Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B.* **1988**, 37, 785−789.

[67] Tirado-Rives, J. and W.L. Jorgensen, Performance of B3LYP density functional methods for a large set of organic molecules. *J. Chem. Theory Comput.* **2008**, 4, 297−306.

[68] Vosko, S.H., L. Wilk, and M. Nusair, Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, 58, 1200−1211.

[69] Becke, A.D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, **1988**, 38, 3098−3100.

[70] Polo, V., Kraka, E. and Cremer, D., Electron correlation and the self-interaction error of density functional theory. *Mol. Phys.* **2002,** 100, 1771-1790.

[71] Staroverov, V.N., Scuseria, G.E., Tao, J. and Perdew, J. P., Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *The J. Chem. Phys.* **2003,** 119, 12129-12137.

[72] Harvey, J.N., On the accuracy of density functional theory in transition metal chemistry. *Annu. Rep. Sect. C: Phys. Chem.* **2006,** 102, 203-226.

[73] Ren, P., Chun, J., Thomas, D.G., Schnieders, M.J., Marucho, M., Zhang, J. and Baker, N.A., Biomolecular alectrostatics and solvation: A computational perspective. *Q. Rev. Biophys.* **2012,** 45, 427-491.

[74] Tomasi, J., Mennucci, B. and Cammi, R., Quantum mechanical continuum solvation models. *Chem. Rev.* **2005,** 105, 2999-3094.

[75] Cossi, M. Scalmani, G., Rega, N. and Barone, V., New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution. *The J. Chem. Phys.* **2002,** 117, 43-54.

[76] Andzelm, J.; Kölmel, C.; Klamt, A., Incorporation of solvent effects into density functional calculations of molecular energies and geometries. *J. Chem. Phys.* **1995,** 103, 9312-9320.

[77] Onsager, L., Electric moments of molecules in liquids. *J. Am. Chem. Soc.* **1936,** 58, 1486-1493.

[78] Marenich, A.V., Cramer, C.J. and Truhlar, D.G., Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009,** 113, 6378-6396.

[79] Onufriev, A. V., & Case, D. A. (2019). Generalized Born Implicit Solvent Models for Biomolecules. *Annual review of biophysics*, *48*, 275–296.

[80] Jensen, F. Introduction to computational chemistry. *John Wiley and Sons*, **2007**.

[81] Cramer, C. J. Essentials of Computational Chemistry: Theories and Models. *John Wiley and Sons*, 2013.

[82] Himo, F., Quantum chemical modeling of enzyme active sites and reaction mechanisms. *Theor. Chem. Acc.* **2006,** *116*, 232−240.

[83] Leopoldini, M., Marino, T., Michelini, M.D.C., Rivalta, I., Russo, N., Sicilia, E. and Toscano, M., The role of quantum chemistry in the elucidation of the elementary mechanisms of catalytic processes: from atoms, to surfaces, to enzymes. *Theor. Chem. Acc.* **2007,** *117*, 765−779.

[84] Siegbahn, P.E.M. and Himo, F., The quantum chemical cluster approach for modeling enzyme reactions. *WIREs Comput. Mol. Sci.* **2011,** *1*, 323−336.

[85] Ramos, M.J. and Fernandes, P.A., Computational enzymatic catalysis. *Acc. Chem. Res.* **2008,** *41*, 689-698.

[86] Siegbahn, P.E.M. and Crabtree, R.H., Mechanism of C−H activation by diiron methane monooxygenases: Quantum chemical studies. *J. Am. Chem. Soc.* **1997,** *119*, 3103−3113.

[87] Wirstam, M., M.R.A. Blomberg, and P.E.M. Siegbahn, Reaction mechanism of compound I formation in heme peroxidases: A density functional theory study. *J. Am. Chem. Soc.* **1999**, 121, 10178−10185.

[88] Sheng, X. and F. Himo, Theoretical study of enzyme promiscuity: Mechanisms of hydration and carboxylation activities of phenolic acid decarboxylase. *ACS Catal.* **2017**, 7, 1733−1741.

[89] Banás, P., Jurecka, P., Walter, N.G., Sponer, J. and Otyepka, M. Theoretical studies of RNA catalysis: hybrid QM/MM methods and their comparison with MD and QM. *Methods* **2009,** *49*, 202−16.

[90] Lin, H. and Truhlar, D.G., QM/MM: what have we learned, where are we, and where do we go from here? *Theor. Chem. Acc.* **2006**, 117, 185.

[91] Lu, X., Fang, D., Ito, S., Okamoto, Y., Ovchinnikov, V. and Cui, Q. QM/MM free energy simulations: recent progress and challenges. *Mol. Simul.* **2016,** *42*, 1056−1078.

[92] Groenhof, G., Introduction to QM/MM simulations. In *biomolecular simulations: methods andprotocols*, Monticelli, L., Salonen, E., Eds. Humana Press: Totowa, NJ, 2013; pp 43-66.

[93] Duarte, F., Amrein, B.A., Blaha-Nelson, D. and Kamerlin, S.C.L. Recent advances in QM/MM free energy calculations using reference potentials. *Biochim. Biophys. Acta Gen. Subj.* **2015,** *1850*, 954−965.

[94] Futera, Z. and J.V. Burda, Reaction mechanism of Ru(II) piano-stool complexes: Umbrella sampling QM/MM MD study. *J. Compu. Chem.* **2014**, 35, 1446−1456.

[95] Genheden, S. and Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **2015,** *10*, 449−461.

[96] Montalvo-Acosta, J.J. and Cecchini, M. Computational approaches to the chemical equilibrium constant in protein-ligand binding. *Mol. Inform.* **2016,** *35*, 555−567.

[97] Zwanzig, R. W., Kirkwood, J. G., Oppenheim, I. and Alder, B. J., Statistical mechanical theory of transport processes. VII. The coefficient of thermal conductivity of monatomic liquids. *J. Chem. Phys.* **1954,** *22*, 783−790.

[98] Kirkwood, J. G. and Monroe, E., Statistical Mechanics of Fusion. *J. Chem. Phys.* **1941,** *9*, 514−526.

[99] Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976,** *22*, 245−268.

[100] Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H. and Kollman, P.A., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* **1992,** *13*, 1011−1021.

[101] Souaille, M. and Roux, B.t., Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comput. Phys. Commu.* **2001,** *135*, 40-57.

[102] Miller, B.R., 3rd, McGee, T.D., Jr., Swails, J.M., Homeyer, N., Gohlke, H. and Roitberg, A.E., MMPBSA.py: An efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* **2012,** *8*, 3314−21.

[103] Ylilauri, M. and Pentikäinen, O.T., MMGBSA As a tool to understand the binding affinities of filamin–peptide interactions. *J. Chem. Info. Model.* **2013,** *53*, 2626−2633.

[104] Genheden, S. and Ryde, U., The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin. Drug Discov.* **2015,** *10*, 449-461.

[105] Suárez, D. and Díaz, N. Direct methods for computing single-molecule entropies from molecular simulations. *WIREs Comput. Mol. Sci.* **2015,** *5*, 1-26.

[106] Zhou, H.X. and Gilson, M.K., Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* **2009**, *109*, 4092–4107.

[107] van Speybroeck, V., Gani, R. and Meier, R.J., The calculation of thermodynamic properties of molecules. *Chem. Soc. Rev.* **2010**, *39*, 1764–1779.

[108] Polyansky, A.A., Zubac, R. and Zagrovic, B., Estimation of conformational entropy in protein-ligand interactions: A computational perspective. In computational drug discovery and design. Ed., Springer: Berlin, Germany, **2012**, *819*, 327–353.

[109] Baron, R. and McCammon, J.A., Molecular recognition and ligand association. *Ann. Rev. Phys. Chem.* **2013**, *64*, 151–175.

[110] Suárez, D. and Diaz, N. Direct methods for computing single-molecule entropies from molecular simulations., *Rev. Comput. Sci.* **2015**, *5*, 1–26.

[111] Kassem, S., Ahmed, M., El-Sheikh, S. and Barakat, K.H., Entropy in bimolecular simulations: A comprehensive review of atomic fluctuations-based methods. *J. Mol. Graph. Model.* **2015**, *62*, 105–117.

[112] Butler, K.T., Walsh, A., Cheetham, A.K. and Kieslich, G., Organised chaos: Entropy in hybrid inorganic-organic systems and other materials. *Chem. Sci.* **2016**, *7*, 6316–6324.

[113] Chong, S.H., Chatterjee, P. and Ham, S., Computer Simulations of Intrinsically Disordered Proteins. *Ann. Rev. Phys. Chem.* **2017**, *68*, 117–134.

[114] Wilson, E. B., Some mathematical methods for the study of molecular vibrations. *J. Chem. Phys.* **1941,** *9*, 76−84.

[115] Karplus, M. and Kushick, J.N., Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981,** *14*, 325−332.

[116] McQuarie, D. A. Statistical mechanics; Harper and Row: New York, 1978.

[117] Pascal, T.A., Lin, S.T. and Goddard Iii, W. A., Thermodynamics of liquids: standard molar entropies and heat capacities of common solvents from 2PT molecular dynamics. *Phys. Chem. Chem. Phys.* **2011,** *13*, 169−181.

[118] King, B.M., Silver, N.W. and Tidor, B., Efficient calculation of molecular configurational entropies using an information theoretic approximation. *J. Phys. Chem. B* **2012,** *116*, 2891-904.

[119] Cover, T.M. and Thomas, J.A., Elements of Information Theory. 2nd ed. Hoboken, NJ: *Wiley-Inter Science*; **2006**.

[120] Kassem, S., Ahmed, M., El-Sheikh, S. and Barakat, K.H., Entropy in bimolecular simulations: A comprehensive review of atomic fluctuations-based methods. *J. Mol. Graph. Model.* **2015,** *62*, 105−117.

[121] Evans, D.A. and Wales, D.J., Free energy landscapes of model peptides and proteins. *J. Chem. Phys.* 2003, *118*, 3891−3897.

[122] Suaŕez, E., Diaz, N. and Suaŕez, D., Entropy calculations of single molecules by combining the rigid-Rotor and harmonic-oscillator approximations with conformational entropy estimations from molecular dynamics simulations. *J. Chem. Theory Comput.* 2011, *7*, 2638−2653.

[123] Schlitter, J., Estimation of absolute and relative entropies of macromolecules using the covariance-matrix. *Chem. Phys. Lett.* 1993, *215*, 617−621.

[124] Schafer, H., Daura, X., Mark, A.E. and van Gunsteren, W.F., Entropy calculations on a reversibly folding peptide: Changes in solute free energy cannot explain folding behavior proteins. *Struct. Funct. Genet.* 2001, *43*, 45−56.

[125] Andricioaei, I. and Karplus, M., On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* 2001, *115*, 6289−6292.

[126] Genheden, S. and Ryde, U. Will molecular dynamics simulations of proteins ever reach equilibrium? *Phys. Chem. Chem. Phys.* 2012, *14*, 8662−8677.

[127] Hensen, U., Graẗer, F. and Henchman, R.H. Macromolecular entropy can be accurately computed from force. *J. Chem. Theory Comput.* 2014, *10*, 4777−4781.

[128] Hikiri, S., Yoshidome, T. and Ikeguchi, M. Computational methods for configurational entropy using internal and cartesian coordinates. *J. Chem. Theory Comput.* 2016, *12*, 5990−6000.

[129] Li, D.W. and Bruschweiler, R., In silico relationship between configurational entropy and soft degrees of freedom in proteins and peptides. *Phys. Rev. Lett.* 2009, *102*, 118108.

[130] Klefas-Stennett, M. and Henchman, R.H., Classical and quantum Gibbs free energies and phase behavior of water using simulation and cell theory. *J. Phys. Chem. B* 2008, *112*, 9769−9776.

Chapter 2. Entropy of Simulated Liquids Using Multiscale Cell Correlation

PAPER 1

Hafiz Saqib Ali[1,2], Jonathan Higham[1,2,3] and Richard H. Henchman[1,2*]

[1] Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

[2] School of Chemistry, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

[3] Human Genetics Unit, Institute of Genetics & Molecular Medicine, The University of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh EH4 2XU, UK

* Corresponding author:

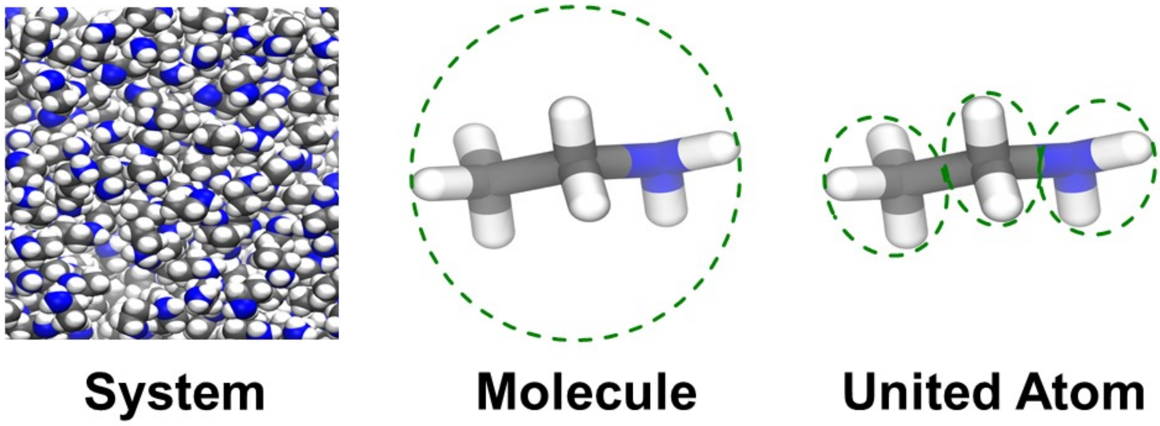
E-mail: henchman@manchester.ac.uk

Conceptualization, supervision and project administration, R.H.H.; investigation and funding acquisition, H.S.A.; methodology and software, J.H., H.A.S and R.H.H.; validation, visualization and writing—original draft preparation, H.S.A. and R.H.H.; formal analysis and writing—review and editing, H.S.A., J.H. and R.H.H.

Published in *Entropy*, Volume 21, Issue 8, Page 750, 31 July 2019.

Abstract

Accurately calculating the entropy of liquids is an important goal, given that many processes take place in the liquid phase. Of almost equal importance is understanding the values obtained. However, there are few methods that can calculate the entropy of such systems, and fewer still to make sense of the values obtained. We present our multiscale cell correlation (MCC) method to calculate the entropy of liquids from molecular dynamics simulations. The method uses forces and torques at the molecule and united-atom levels and probability distributions of molecular coordinations and conformations. The main differences with previous work are the consistent treatment of the mean-field cell approximation to the approriate degrees of freedom, the separation of the force and torque covariance matrices, and the inclusion of conformation correlation for molecules with multiple dihedrals. MCC is applied to a broader set of 56 important industrial liquids modeled using the Generalized AMBER Force Field (GAFF) and Optimized Potentials for Liquid Simulations (OPLS) force fields with 1.14*CM1A charges. Unsigned errors versus experimental entropies are 8.7 J K^{-1} mol^{-1} for GAFF and 9.8 J K^{-1} mol^{-} for OPLS. This is significantly better than the 2-Phase Thermodynamics method for the subset of molecules in common, which is the only other method that has been applied to such systems. MCC makes clear why the entropy has the value it does by providing a decomposition in terms of translational and rotational vibrational entropy and topographical entropy at the molecular and united-atom levels.



2.1 Introduction

Molecular liquids are present in numerous systems in chemistry and biology. However, methods to calculate their entropy are scarce or limited in scope. Entropy quantifies the probability distribution of quantum states of a system and, together with energy, determines a system's stability. The most common route used to determine entropy is indirect, being as a difference with respect to a reference state, typically the ideal gas or a non-interacting set of atoms. The entropy difference may be extracted from integrated heat capacity changes or from the Gibbs energy difference, either as its temperature derivative or as a difference with enthalpy [1]. While there is a range of methods to compute entropy [2–10], those that use single molecular dynamics or Monte Carlo simulations are advantageous because of the ease of using standard simulation methods and because such approaches directly yield and explain entropy and structure in terms of the full probability distribution of the system of interest. However, because the ensemble of molecular configurations generated by standard simulation methods is only a tiny fraction of the full ensemble corresponding to a system's entropy, special techniques are required to extrapolate to the full probability distribution and entropy.

The probability distributions to evaluate entropy are typically over the coordinates of the system, which may be Cartesian coordinates, bonds-angles-dihedrals, or interatomic distances. Histogram-based methods, because of their arbitrary bin-widths, can only give the entropy difference relative to a reference, which is typically the uniform distribution. Even then, the entropy difference may be unrealistic for strongly interacting systems such as those with covalent bonds because of the omission of quantum effects which necessarily keep the entropy non-negative. For this reason, histogram methods are often restricted to softer degrees of freedom such as dihedrals or atomic distances. The simplest approach ignores coordinate correlations by considering each coordinate separately, for example, in dihedral angles [11]. Higher-order correlations can be included such as the radial distribution function [12–14] or a mutual-information expansion [15,16] but at greater computational expense and complexity, even for second-order, although some correlations are small and can be excluded [17–19]. Extensions to higher orders are difficult and do not necessarily lead to more accuracy [15,20]. Mutual information in terms of discrete rotamers has been found to converge much faster, enabling up to eighth order [21]. An alternative strategy for high-dimensional data sets is the k-Nearest Neighbours method [16,22–24] which more adaptively estimates density from the distances between configurations but at the price of having many distances to compute and still requiring a lot of data to converge.

Significant simplification of the theory, greater speed of convergence and a route to the direct calculation of entropy is provided by assuming a multivariate Gaussian probability distribution [25]. Entropy is directly computed from the quantum states of the set of harmonic-oscillator eigenvectors [26,27]. The main limitation of the method is the suitability of the Gaussian distribution, given that typical potential energy surfaces for flexible molecules [28] or liquids [27,29] have multiple minima, compounded by the difficulty of how to specify the minima. A hybrid solution to this problem is to replace the diagonal elements of the coordinate covariance matrix with the entropy of the probability distributions [30,31]. Another solution is to incorporate multiple Gaussians [32]. An approach particularly relevant to the case of liquids is the 2-Phase Thermodynamics (2PT) method, which calculates entropy from the spectrum of vibrational frequencies derived from the velocity auto-correlation function and the gas-phase fluidicity [33]. Another viable method for liquid-phase entropy is the cell approximation which maps regions of the potential energy surface into single, representative energy wells, whose entropy is determined from the force [34] plus an entropy term for the probability distribution of the energy wells [35]. This is the method we have been working to generalise, progressing from liquid argon [34] to liquid water with its rotational vibration and orientational degrees of freedom [35–37], organic liquids with an internal one-dimensional dihedral entropy [38], single molecules with internal entropy based on force correlation [39], and molecular liquids in a multiscale framework from atom to united atom to molecule to system [40]. This development has been supported by extensive parallel studies on the entropy of aqueous solutions [41–47]. With the main ideas now in place to make the method general, to encapsulate the main features of the method we name it Multiscale Cell Correlation (MCC).

Here we extend MCC to calculate the entropy of 56 important industrial liquids. These represent a class of system which no other method has been capable of calculating entropy except for the 2PT method, which has been tested on a smaller subset of 14 liquids [48], argon [33], water [49], carbon dioxide [50], and methanol and hexane including torsional fluidicity [51]. The first improvement here in MCC is a more appropriate application of the mean-field cell approximation to the weakly correlated non-bonded and dihedral degrees of freedom and not to the correlated bonded and angular degrees of freedom as had been done in previous work [39,40]. Strong correlations for the bonded atoms invalidate the cell approximation and can be accounted for in the force covariance matrices. Related to this, force and torque covariances are evaluated separately because of their weak correlation [40]. The second key improvement is a new way to account for correlation between dihedrals by using a covariance matrix of conformation correlation, a method that scales with the square of the number of dihedrals. The 56 liquids are tested using two force fields: OPLS (Optimized Potentials for Liquid Simulations) with 1.14*CM1A charges [52] and GAFF [53] (Generalized AMBER Force Field), for both of which parameters can be generated in an automated fashion for a wide range of molecules. A decomposition of the entropy in six terms gives an insightful and intuitive explanation of why molecules have the entropy they do. Compared to our earlier study in which a comparison with 2PT was inconclusive because there were few liquids in common, MCC is found to be significantly closer to experiment than 2PT, which in most cases underestimates experiment. An analysis of entropy components suggests that the internal entropy of 2PT is responsible for this underestimation, even when torsional fluidicity is included [51]. The findings show that MCC is well placed to scale to complex multi-component systems with multiple length scales.

2.2 Theory

2.2.1 Entropy Decomposition

The entropy of molecular liquids is well captured at two different length scales [40]: the molecule (M) level and the united-atom (UA) level. A united atom is defined here as a non-hydrogen atom together with any bonded hydrogen atoms and is taken as a rigid body with both translational and rotational degrees of freedom rather than only translation as for a point-particle unless there are no hydrogens. Such an approach captures softer collective dihedral motion of hydrogens while ignoring their individual stretching and bending motions which have negligible entropy, owing to the low mass of hydrogen and its higher bond and angle vibrational frequencies. At the other extreme of the whole system, the entropy of its three translational and three rotational degrees of freedom is negligible on a per-molecule basis. Coordinate systems at the molecule and united-atom levels are defined as before [40]. For a molecule this is its three principal axes with the origin at the centre of mass. For a united atom the axes and centre of mass depend on the number of bonded united and hydrogen atoms. All non-linear molecules and united atoms have three translational degrees of freedom. Linear molecules in terms of their united atoms or linear united atoms in terms of their hydrogens have two rotational degrees of freedom. United atoms with no hydrogens have no rotational degrees of freedom.

In the cell approximation the potential energy surface is partitioned into energy wells, and in the multiscale approximation this partitioning is done at the molecule and united-atom levels. This brings about two kinds of entropy term: vibrational relating to the average size of the energy wells, termed a cell, and topographical relating to the probability of the energy wells. The vibrational term at each level is further partitioned according to the translational (transvib) and rotational (rovib) degrees of freedom. The translational component of the topographical entropy at the molecular level is zero for a pure liquid because exchanging identical molecules leads to no change. The rotational topographical entropy (topo) at the molecule level is termed the orientational entropy. At the united-atom level the translational topographical entropy is the conformational entropy, while the rotational component, corresponding to hydrogen-bond arrangements, is negligible for the liquids studied here. The total entropy per molecule for a liquid is therefore taken as the sum of six terms

$$S_{\text{total}} = S_{\text{M}}^{\text{transvib}} + S_{\text{M}}^{\text{rotvib}} + S_{\text{M}}^{\text{topo}} + S_{\text{UA}}^{\text{transvib}} + S_{\text{UA}}^{\text{rotvib}} + S_{\text{UA}}^{\text{topo}} \tag{2.1}$$

2.2.2 Molecular Vibrational Entropy

All four vibrational entropy terms are calculated in the harmonic approximation using the equation for a collection of N_{vib} quantum harmonic oscillators

$$S_{\text{vib}} = k_{\text{B}} \sum_{\text{i}=1}^{N_{\text{vib}}} \left(\frac{hv_i/k_{\text{B}}T}{\text{e}^{hv_i/k_{\text{B}}T} - 1} - \ln\left(1 - \text{e}^{-hv_i/k_{\text{B}}T}\right) \right) \tag{2.2}$$

where k_{B} is Boltzmann's constant, h is Planck's constant, T is temperature and v_i are the vibrational frequencies. Different to previous work [40], translational and rotational vibrational entropy are evaluated separately, justified by the absence of correlations between the forces and torques that are used to evaluate them. For $S_{\text{M}}^{\text{transvib}}$, $N_{\text{vib}} = 3$ and v_{i} are calculated using [39,40]

$$v_i = \frac{1}{2\pi} \sqrt{\frac{\lambda_i}{k_{\text{B}}T}} \tag{2.3}$$

where λ_i are the eigenvalues of the 3 × 3 mass-weighted force covariance matrix of the molecule with elements $\langle F_i^{'} F_j^{'} \rangle$, with i and j ranging over the three axes x, y, z and averaging over all molecules in all simulation frames. Mean-field, mass-weighted forces are defined as $F_i' = F_i/(2\sqrt{m})$ where m is the molecule's mass, and F_i is half the component of net force on all the atoms of the molecule rotated into the molecule's coordinate frame. In practice, this matrix is essentially diagonal because forces along different axes are negligibly correlated. The halving is done in the mean-field cell approximation [34,35]. whereby every pairwise energy term and therefore its negative coordinate derivative, the force, is partitioned equally between the atoms involved. The mean-field cell approximation is justified in liquids because average molecular energies and forces in many-body systems are weakly correlated with the position of any other neighbouring molecule. Only over the short duration of a repulsive collision is the correlation significant. To calculate $S_{\mathrm{M}}^{\mathrm{rovib}}$ with Equation (2.2), $N_{\mathrm{vib}} = 3$ unless the molecular is linear with respect to its united atoms, in which case $N_{\mathrm{vib}} = 2$. The vibrational frequencies v_i are calculated using Equation (2.3) with eigenvalues from the $N_{\mathrm{vib}} \times N_{\mathrm{vib}}$ moment-of-inertia-weighted torque covariance matrix of the molecule, whose elements are $\langle \tau_i' \tau_j' \rangle$ where $\tau_i' = \tau_i/\left(2\sqrt{I_i}\right)$ for each axis $i = x, y, z$ and I_i is the respective moment of inertia, with torque halving being done as for the forces.

2.2.3 United-Atom Vibrational Entropy

The procedure at the united-atom level to evaluate $S_{\mathrm{UA}}^{\mathrm{transvib}}$ and $S_{\mathrm{UA}}^{\mathrm{rovib}}$ in Equation (2.1) is similar to that at the molecule-level but with some differences. United atoms are used in place of molecules to evaluate the forces, torques, masses and moments of inertia. N_{vib} in Equation (2.2) for united-atom translation equals $3N - 6$, where N is the number of united atoms and the six vibrations removed correspond to the six largest eigenvalues which are already accounted for as molecular translation and rotation. N_{vib} in Equation (2.2) for united-atom rotation depends on the number of non-linear, linear and point united atoms, as well as the linearity of the whole molecule. Non-linear and linear united atoms contribute 3 and 2 degrees of freedom, respectively, and the largest six or five eigenvalues are removed if the molecule is non-linear or linear. A notable difference compared to the molecule level is that the mean-field cell approximation is not made for bonded atoms or bonded 1–3 interactions corresponding to angles. The forces of such atoms are strongly correlated, a correlation that is accounted for in the covariance matrix. However, the mean-field approximation is still made for united-atom rotation and dihedral vibration whose correlations with neighbours are weak relative to the overall torque or force or which largely average out to zero because of averaging in different reference frames. Consequently, forces in the united-atom matrix are not halved but united-atom torques are halved. To implement the cell approximation for dihedrals, the N_dih lowest eigenvalues of the united-atom force covariance matrix are halved twice (force-squared), where N_dih is the number of united-atom dihedrals, because these eigenvalues correspond to the soft conformational eigenvectors.

2.2.4 Molecular Topographical Entropy

The molecular topographical entropy $S_\mathrm{M}^\mathrm{topo}$ in Equation (2.4) only has a rotational contribution for a pure liquid, referred to as the orientational entropy. Based on the idea that neighbouring molecules discretize a molecule's rotational motion, $S_\mathrm{M}^\mathrm{topo}$ is estimated using an average of the number M of orientations weighted by the probability $p(N_\mathrm{C})$ of molecular coordination number N_C using [40]

$$S_\mathrm{M}^\mathrm{topo} = k_\mathrm{B} \sum_{N_\mathrm{c}} p(N_\mathrm{c}) \ln \left[\max{(N_\mathrm{c}^3 \pi)^{\frac{1}{2}}/\sigma}\right] \tag{2.4}$$

where σ is the symmetry number of the molecule according to its united atoms. The max function only takes effect for the very small values of N_c which are rare. Thus, there are $\sim N_\mathrm{c}^{1/2}$ orientations per rotational axis, and every orientation is taken to have the same probability, $\sigma/(N_\mathrm{c}^3\pi)^{1/2}$, justified by the weak correlation of these moderately polar molecules with their neighbours. For linear molecules with two axes of rotation [40], the equation is

$$S_\mathrm{M}^\mathrm{topo} = k_\mathrm{B} \sum_{N_\mathrm{c}} p(N_\mathrm{c}) \ln \left[\max{(1, \frac{N_\mathrm{c}}{\sigma})}\right] \tag{2.5}$$

Molecules with a single united atom may still have orientational entropy at the atom-level if their hydrogens sufficiently break symmetry, so as to form distinct energy wells. Ammonia is included in this category, as water had been earlier [40], but methane and hydrogen sulfide are not. N_C is evaluated using the parameter-free relative angular distance (RAD) method [54,55] according to the centre of mass of each molecule. RAD determines N_C from a single configuration in good agreement with those using a cut-off at the first minimum in the radial distribution function. It avoids the need for a mean-field, spherically-symmetric cut-off that must either be chosen arbitrarily or evaluated from the pre-computed radial distribution function.

2.2.5 United-Atom Topographical Entropy

The topographical entropy at the united-atom level, $S_{\mathrm{UA}}^{\mathrm{topo}}$, also called the conformational entropy, is derived from the distribution of discrete conformations for all flexible dihedrals involving united atoms. Unlike in the previous work on liquid entropy [40] in which the molecules only had a maximum of one flexible dihedral, a number of molecules here have multiple dihedrals. Given that they may be correlated, we present a new method to account for this using a conformation correlation matrix. Each molecule has N_{dih} dihedrals, taken as four consecutives, bonded united atoms. The topographial entropy of dihedrals at the atomic level and involving hydrogen are ignored, either because they have only one conformation by symmetry, such as a methyl group, or because they have negligibly more than one conformation, such as a hydroxyl, owing to limited variable hydrogen-bonding capability to neighbour molecules. The molecules considered here have three conformations per dihedral: trans (*t*), gauche− (*g*–) and gauche+ (*g*+), defined with boundaries in dihedral angle at 120°, 0° and −120°, respectively. Thus, each molecule has available $3N_{\mathrm{dih}}$ conformations. Every combination of conformations for each molecule is termed a conformer, and the total possible number of conformers is $3^{N_{\mathrm{dih}}}$. Overall, we ensure there is no double-counting of identical conformers by treating g^{-} and g^{+} as distinct and dividing by the rotational symmetry number in Equations (4) and (5). We construct the $3N_{\mathrm{dih}} \times 3N_{\mathrm{dih}}$ correlation matrix ρ which has elements

$$\rho_{ij} = p_{ij} r_{ij} / P_{m(i)} \tag{2.6}$$

where p_{ij} is the probability of simultaneously having the conformation pair *i* and *j*, normalised such that $\sum_{i=3m}^{3m+2} \sum_{j=3n}^{3n+2} p_{ij} = 1$ for the square sub-block over all conformation i and j of the respective dihedrals *m* and *n*, and r_{ij} is the Pearson correlation coefficient of conformations *i* and *j*, given by

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\left[(p_i - p_i^2)(p_j - p_j^2)\right]^{1/2}} \tag{2.7}$$

where p_i is the probability of conformation i. Note that $r_{ii} = 1$, $p_{ii} = p_i$, and $p_{ij} = 0$ if i and j belong to the same dihedral. $P_{m(i)}$ in Equation (2.6) are normalisation constants, one per dihedral m, that are defined to ensure $\sum_{i=3m}^{3m+2} \sum_{j=1}^{3N_{\mathrm{dih}}} p_{ij} = 1$ for each dihedral m. Thus, ρ_{ij} represents the fraction of correlation that conformation pair ij makes to the total correlation that i's dihedral m has with all conformations of all dihedrals. Similar to the von Neumann entropy of the density matrix [56], the total conformational entropy is given by

$$S_{\mathrm{UA}}^{\mathrm{topo}} = -k_{\mathrm{B}} \sum_{\mathrm{i}=1}^{3N_{\mathrm{dih}}} \lambda_i \ln (\lambda_{\mathrm{i}}) \tag{2.8}$$

where λ_i, the eigenvalues of ρ, are the probability of each conformer eigenvector, and each conformer eigenvector itself comprises the probabilities of each conformation. Unlike the density matrix, whose trace equals 1 [56], the trace of ρ ranges from 1, corresponding to full correlation between conformations, to a maximum of N_{dih}, corresponding to fully uncorrelated conformations. For a molecule with uncorrelated conformations or with only one dihedral [40], its eigenvalues would be the diagonal elements of ρ and the conformer eigenvectors would be the individual conformations. At the other extreme of full correlation, as would occur when there is only one single conformer, one eigenvalue would equal 1 with its eigenvector being that very conformer, while the remaining eigenvalues would be zero. For cases of intermediate correlation between conformations, the eigenvector conformers would have various contributions from the correlated conformations, with entropy ranging from zero to the fully disordered value for all N_{dih} dihedrals.

2.3 Methodology

2.3.1 Molecular Dynamics Simulations

The entropy was calculated for a series of 56 liquids using molecular dynamics simulations. All simulations were carried out with the sander module of the AMBER 14 simulation package [57]. Each system consists of 500 identical molecules in the liquid phase in a cubic box. The force fields used were GAFF [53] with AM1-BCC charges for all molecules and OPLS-AA with the 1.14*CM1A charges [52] for all molecules except acetonitrile, carbon dioxide, hydrogen sulfide and tetrafluoroethylene for which charges were not available on the LigParGen webserver [52]. In place of this for carbon dioxide, a simulation was run with the

TraPPE (Transferable Potentials for Phase Equilibria) force field [58]. All molecules were built in standard geometry using xleap of AMBER 14. GAFF force-field parameters were generated with antechamber [59] and all molecules were placed in a cubic box of side 6 nm using Packmol [60]. For OPLS, the GROMACS topology and coordinate files were obtained by uploading a pdb of each molecule to the LigParGen webserver [52] with the 1.14*CM1A charges, the coordinates of the box of molecules were generated in GROMACS 5.1 [61], and the topology and coordinate files were converted into AMBER format using the AMBER ParmEd tool. Note that these OPLS charges differ to those in previous work [40] with the OPLS force field which had charges fitted to liquid-phase properties [62]. TraPPE parameters for carbon dioxide were added directly in by hand.

For equilibration, each system was minimized with 500 steps of steepest descent minimization, thermalized in a 100 ps molecular dynamics simulation at constant volume and temperature using a Langevin thermostat with a collision frequency of 5 ps^{-1}, and brought to the correct density with 1 ns of molecular dynamics simulation at constant pressure using the Berendsen barostat with a time constant of 2 ps. For data collection, forces and coordinates were saved every 1 ps in a further 1 ns simulation under the same conditions, which earlier work had shown to be easily sufficient for converged values [40], in which as few as ten frames was often sufficient to achieve converged integer values in units of J K^{-1} mol^{-1}. The pressure was 1 bar and the temperature was 298 K unless the liquid was gaseous at that temperature, in which case the boiling temperature at 1 bar was used as listed in Table 2.1. The exception is carbon dioxide, which does not liquefy at ambient pressure and so the pressure was set to 5.99 bar and temperature 220 K which is in the liquid-phase region, close to the triple point and matches conditions used in a 2PT study [50]. Simulations used SHAKE on all bonds involving hydrogen atoms, a non-bonded cutoff of 8 Å, periodic boundary conditions, particle-mesh Ewald summation with default parameters in AMBER, and a 2 fs timestep. Table 2.4 contains all the liquids simulated, for five of which the following abbreviations are used: dimethylformamide (DMFA), dimethylsulfoxide (DMSO), N-methyl acetamide (NMA), tert-butyl alcohol (TBA) and tetrafluoroethylene (TFE). Entropies were calculated with in-house C++ and Perl code, reading in the force, coordinate and topology files and writing out eigenvalues and coordination numbers.

Table 2.1. Boiling Temperature of Liquids [63] that are Gaseous at Ambient Conditions.

Liquid	***T*/K**	**Liquid**	***T*/K**	**Liquid**	***T*/K**	**Liquid**	***T*/K**
ammonia	240	ethane	185	hydrogen sulfide	213	methylamine	267
butane	272	ethene	170	methane	112	Propane	231
carbon dioxide	220^a	ethylamine	291	methanethiol	279	TFE	197
diazene	275						

^{a} Pressure is 5.99 bar.

2.4 Results

2.4.1 Entropy Values

Figure 2.1 presents the entropy of 50 of the liquids calculated by MCC for the OPLS and GAFF force fields plotted against the respective experimental values [63–70]. Table 2.2 gives the mean unsigned and signed deviations, slopes, intercepts, Pearson correlation coefficients R^2, and zero-intercept slopes of entropies by the MCC and 2PT methods with respect to experiment. Table 2.4 contains the MCC entropy values for all 56 liquids, together with values from experiment, the MCC entropy of carbon dioxide with the TraPPE force field, and values using the 2PT method with the OPLS and GAFF force fields for fifteen liquids [48], carbon dioxide [50] and methanol and hexane including torsional fluidicity [51]. Statistical errors are negligible for the precision given.

Table 2.2. Statistical Data for MCC and 2-Phase Thermodynamics (2PT) versus Experiment.

Data Set (Number of Liquids)	$\langle\lvert S - S_{\text{expt}}\rvert\rangle$/ **J K^{-1} mol^{-1}**	$\langle\lvert S - S_{\text{expt}}\rvert\rangle$/ **J K^{-1} mol^{-1}**	**Slope**	**Y-Intercept/ J K^{-1} mol^{-1}**	**R^2**	**Zero-Intercept Slope**
MCC OPLS^a (46)	9.8	0.6	0.94	11.7	0.95	1.00
MCC GAFF (50)	8.7	–0.3	0.93	13.0	0.96	0.99
2PT OPLS^b (12)	15.5	–15.6	1.05	–25.3	0.84	0.92
2PT GAFF (14)	28.0	–24.4	0.97	–19.5	0.55	0.87
MCC OPLS^a (12)	4.9	2.3	0.87	26.7	0.89	1.01
MCC GAFF (14)	7.6	4.0	0.93	16.5	0.93	1.02

^{a} OPLS with 1.14*CM1A charges [52]; ^{b} OPLS with charges optimised to liquid-phase properties [62].

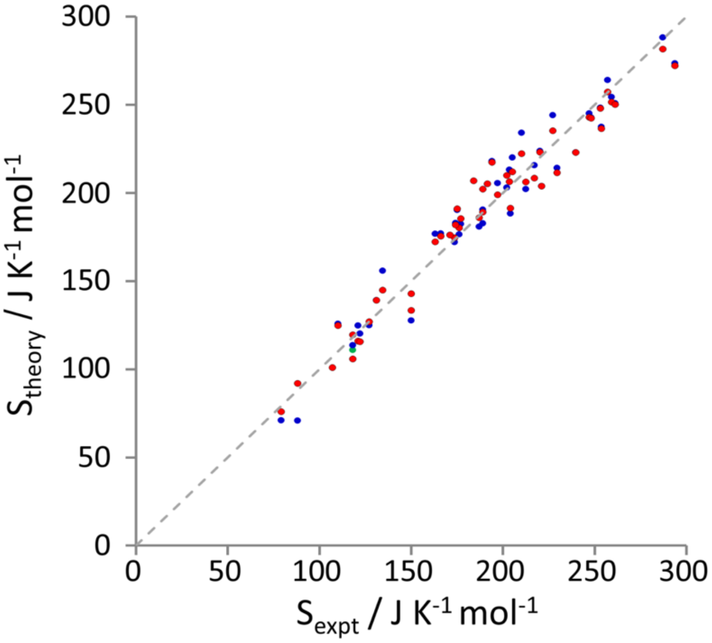


Figure 2.1. Multiscale cell correlation (MCC) entropy values versus experiment for OPLS (blue), GAFF (**red**), and TraPPE (**green**), together with the line of perfect agreement (**dotted**).

The entropy values calculated by MCC agree well with experiment, with Table 2.2 showing a mean unsigned error of less than 10 J K^{-1} mol^{-1}, GAFF being slightly better than OPLS. The small mean signed errors, the slopes being marginally less than one, and the positive y-intercept suggest that MCC is slightly missing the dependence on molecular size, although forcing the line through zero brings about the correct unity slope. The excessive entropies seen for larger molecules in the earlier version of the theory [40] no longer occur because we no longer halve forces for hard internal degrees of freedom in the mean-field approximation.

The experimental entropies for most liquids were taken from the NIST Chemistry Webbook [63]. If more than one value was reported by different authors, all values were included, although for acetic acid, ethanol, ethylene glycol, formic acid, propanol and pyridine the spread is substantial, exceeding 10 J K^{-1} mol^{-1}. Entropies were found elsewhere for ammonia [64], chloroform [65], methane [66], hydrogen peroxide [67], hydrogen sulfide [68] and carbon dioxide [69]. Values for ethylamine and triethylamine were calculated from the experimental gas-phase entropy, enthalpy of vaporization, and either heat capacity at constant pressure or partial pressure [63,70] in Table 2.3. For the remaining six liquids no values could be found in the literature. The experimental entropy is averaged if there is more than one value.

Table 2.3. Experimental Data to Calculate Liquid-Phase Entropy

Molecules	**$S_{\text{gas,bp}}$/ J K^{-1} mol^{-1}**	**ΔH_{vap}/ kJ mol^{-1}**	**$C_{\text{p,gas}}$/ J K^{-1} mol^{-1}**	**P_{vap}/ kPa**	**Equation**
ethylamine	284.8^{b}	28.0^{a}	72.6^{b}		$S_{\text{liq}}^{\text{o}} = S_{\text{gas}}^{\text{o}} + C_{\text{p}}\ln\frac{T_{\text{vap}}}{298} - \frac{\Delta H_{\text{vap}}}{T_{\text{vap}}}$
triethylamine	405.4^{b}	35.1^{a}		7.66^{a}	$S_{\text{liq}}^{\text{o}} = S_{\text{gas}}^{\text{o}} + R\ln\frac{100}{P_{\text{vap}}} - \frac{\Delta H_{\text{vap}}}{298}$

[a] Reference [63]. [b] Reference [70].

Comparisons with experiment are affected by the accuracy of the force field. To compare MCC with 2PT, Table 2.2 contains the statistical quantities for the liquids studied by the 2PT method [48] listed in Table 2.4, comprising 14 with the GAFF force field [53] and 12 with the OPLS force field [62] together with the corresponding MCC values with GAFF and OPLS with 1.14*CM1A charges [52]. For both force fields, the mean unsigned error for 2PT is three times that of MCC, largely because the 2PT values are too small, shown by the negative mean error, negative y-intercept, and poorer correlation. The slope is close to unity but decreases when forced through the origin. The difference between the two methods is unlikely solely due to the different OPLS force fields, given the trend is present for GAFF, that the 2PT values using the earlier OPLS force field better reproduce liquid-phase entropy, and that the same trend was observed earlier when comparing with the same force field [40,71]. The variability in experiment for acetic acid and ethanol may affect this comparison, in that MCC is closer to the higher value and 2PT closer to the lower value, but this would be insufficient to affect the overall trend. The poorer MCC performance of OPLS with 1.14*CM1A charges compared to GAFF likely reflects the over-polarization of the charges to optimise their free energy of hydration [52]. This also likely explains the better performance of OPLS than GAFF for 2PT. Including the localized bond charge corrections, an alternative provided by LigParGen, is unlikely to lead to any improvement in entropy, given their mixed performance in calculating enthalpies of vaporization and density of liquids [52]. The more positive signed error for OPLS indicates that its entropies overall are larger than the GAFF entropies, implying that the combined intermolecular and intramolecular OPLS interactions are marginally weaker than GAFF. Contrary to this trend, the largest deviations between the force fields are OPLS being ~20 J K^{-1} mol^{-1} lower than GAFF for ammonia and DMSO.

Table 2.4. Entropy by Experiment, MCC and 2PT (J K^{-1} mol^{-1}).

Liquid	Experiment [a]	MCC		2PT [48]	
		OPLS	GAFF	OPLS	GAFF
acetic acid	158, 194	177	180	147	128
acetone	200	202	206	198	187
acetonitrile	150		143		145
ammonia	87 [b]	71	92		
aniline	191, 192	205	205		
benzene	173, 175	183	182	172	161
benzyl alcohol	217	216	208		
benzaldehyde	221	204	204		
butane	227, 230, 231	214	212		
butanol	226, 228	244	235		
2-butaoxyethanol		293	301		
carbon dioxide	118 [c]	111[d]	106	112 [d]	
chloroform	202 [e]	203	210	193	226
cyclohexane	204, 206	220	212		
diazene	121	125	116		
dichloromethane	175	190	191		
diethanolamine		248	256		
diethyl ether	253, 254	237	236		
DMFA		214	222		
DMSO	189	183	202	164	159
1,4-dioxane	197	206	199	179	159
ethane	127	125	127		
ethanol	160, 161, 177	177	175	141	127
ethene	118	114	120		
ethyl acetate	259	254	252		
ethylamine	189[f]	181	185		
ethylene glycol	167, 180	172	175	141	121
formamide		151	153		
formic acid	128, 132, 143	156	145		
furan	177	181	186	168	157

hexane	290, 295, 296	273	272	251[g]	
hexanol	287	288	281		
hydrazine	122	120	116		
hydrogen peroxide	110[h]	126	125		
Hydrogen sulfide	106[i]		101		
methane	79[j]	73	78		
methanethiol	163	177	172		
methanol	127, 130, 136	139	139	117[g],122	109
methylamine	150	128	133		
NMA		205	206	181	168
octanol		335	331		
pentane	259, 263	251	250		
pentanol	255, 259	264	257		
piperidine	210	234	222		
propane	171	176	176		
propanol	193, 214	213	206		
pyridine	178, 179, 210	191	189		
styrene	238, 241	223	223		
TBA	190, 198	218	217		
tetrahydrofuran	204	188	192	197	159
TFE	184		207	195	185
toluene	219, 221	224	223	204	190
triethylamine	309[f]	292	295		
m-xylene	252, 254	248	248		
o-xylene	246, 248	245	245		
p-xylene	244, 247, 253	243	243		

[a] Reference [63] Experimental errors < 1 J K^{-1} mol^{-1} ; [b] Reference [64]; [c] References [50,69]; [d] TraPPE force field [58]; [e] Reference [65];[f] Derived in Table S2 using References [63,70]; [g] Reference [51]; [h] Reference [67]; [i] Reference [68];[j] Reference [66].

2.4.2 Entropy Components

To give deeper understanding into the values of the entropies, their six components in Equation (2.1) are illustrated in Figure 2.2 for the case of GAFF while the numerical values for both force fields are shown in Table 2.6. Plotted in Figure 2.3 are the entropy components as a function of molecular mass, and Table 2.5 lists data for the lines of best fit. The first observation is the dominance of the molecular translational and rotational entropy, being more than half the total entropy for all but the largest molecules. $S_{\mathrm{M}}^{\mathrm{transvib}}$ has a weak dependence on mass, deviating lower for systems at colder temperatures. $S_{\mathrm{M}}^{\mathrm{rosvib}}$has a stronger mass-dependence and is lower for colder and linear molecules and those forming hydrogen bonds. One point to emphasise about our decomposition is that linear molecules in terms of united atoms, such as ethane or acetonitrile, have negligible rotational entropy about their long axis at the molecule level. The entropy about this axis including hydrogens is instead assigned to the united-atom level.

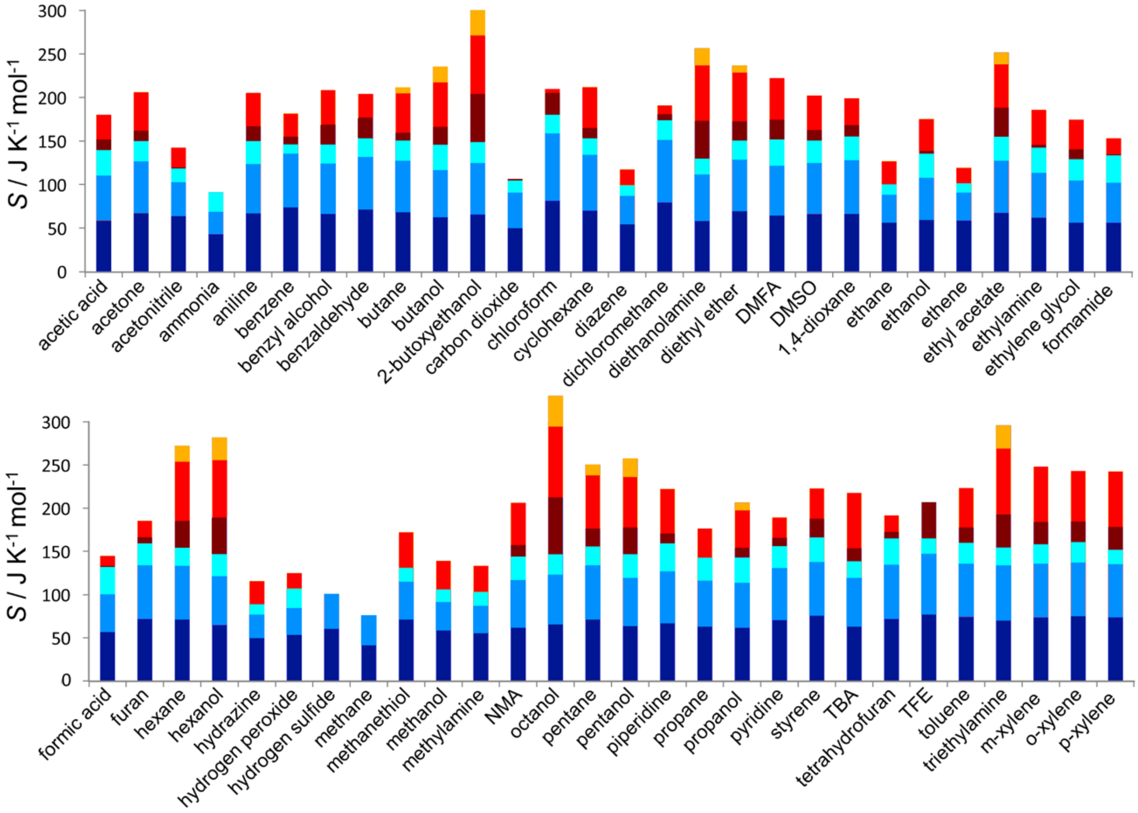


Figure 2.2. MCC entropy components for GAFF (bottom to top): molecular-translational (dark blue), molecular rotational (blue), molecular topographical (cyan), united-atom translational (dark red) united-atom rotational (red), and united-atom topographical (orange).

$S_{\mathrm{UA}}^{\mathrm{transvib}}$ is slightly smaller, making up about a quarter of the total. It primarily comprises the twisting of united atoms such as methyls (∼17 J K^{-1} mol^{-1}) and hydroxyls (∼13 J K^{-1} mol^{-1}) as well as hydrogen bending, such as in benzene, and thus relates more specifically to the number of hydrogens. As mentioned earlier, for linear molecules with two united atoms, it also includes the entropy of rotation about the long axis because this term would otherwise be zero without hydrogen. For example, for ethene $S_{\mathrm{M}}^{\mathrm{rosvib}}$ is smaller than for other molecules, and most of its $S_{\mathrm{UA}}^{\mathrm{rosvib}}$ is rotational entropy about the long axis, leaving about 3 J K^{-1} mol^{-1} for internal motion. The remaining three terms are more variable and together make up about a quarter to a third of the total. The orientational term $S_{\mathrm{M}}^{\mathrm{topo}}$ weakly increases with mass and is smaller for molecules with higher symmetry or those that form hydrogen bonds, which tend to reduce N_c. $S_{\mathrm{UA}}^{\mathrm{transvib}}$ mainly comprises dihedral vibration of united atoms and has a strong dependence on mass, as does the conformational term $S_{\mathrm{UA}}^{\mathrm{topo}}$, which is one of the smallest terms and only present for 13 liquids. The lines of best fit for each component indicate moderate predictability based on mass, but a thorough treatment is beyond the scope of this work. Comparing the force fields, GAFF has marginally higher molecular vibrational entropy (1.5 J K^{-1} mol^{-1}) and higher $S_{\mathrm{UA}}^{\mathrm{topo}}$ (5.2 J K^{-1} mol^{-1}) whereas OPLS has more $S_{\mathrm{UA}}^{\mathrm{rovib}}$ (2.2 J K^{-1} mol^{-1}). Of the most extreme deviations, $S_{\mathrm{UA}}^{\mathrm{topo}}$ of GAFF is 14 J J K^{-1} mol^{-1} higher than OPLS for 2-butoxyethanol and 12 J K^{-1} mol^{-1} higher for diethanolamine. Why this is so is revealed by an inspection of the probability distributions in Table 2.7 which indicate that the reduced $S_{\mathrm{UA}}^{\mathrm{topo}}$ for OPLS is because of stronger internal hydrogen-bonding. In more detail than looking at overall entropy, these trends imply that GAFF compared to OPLS has weaker intermolecular interactions, consistent with the charge over-polarisation of OPLS mentioned earlier [52], more evenly occupied conformations, and stronger intramolecular interactions, particularly relating to united-atom rotation.

Table 2.5. Lines of Best Fit for the Entropy Components versus Molecular Mass.

Components	Slope/ J K^{-1} mol^{-1}	Y-Intercept/ J K^{-1} mol^{-1}	R^2	Components	Slope/ J K^{-1} mol^{-1}	Y-Intercept/ J K^{-1} mol^{-1}	R^2
$S_{\mathrm{M}}^{\mathrm{transvib}}$	0.21	50	0.54	$S_{\mathrm{UA}}^{\mathrm{transvib}}$	0.42	14	0.63
$S_{\mathrm{M}}^{\mathrm{rovib}}$	0.35	28	0.70	$S_{\mathrm{UA}}^{\mathrm{rovib}}$	0.43	6	0.34
$S_{\mathrm{M}}^{\mathrm{top}}$	0.09	16	0.13	$S_{\mathrm{UA}}^{\mathrm{top}}$	0.39	16	0.87

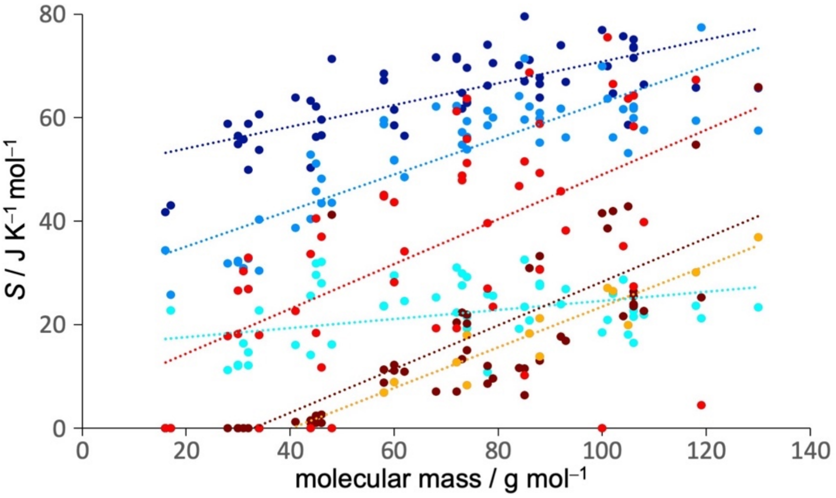


Figure 2.3. MCC entropy components for GAFF versus molecular mass for all liquids. Molecular-translational (dark blue), molecular rotational (blue), molecular topographical (cyan), united-atom translational (dark red) united-atom rotational (red), and united-atom topographical (orange).

A direct comparison of entropy components with 2PT for the 15 liquids in common [48] cannot be done because different OPLS force fields are used, but in general the 2PT molecular translational and rotational entropies are larger than the equivalent MCC terms, and the MCC terms become slightly larger upon inclusion of the orientational term. However, the three MCC united-atom terms are larger than the internal vibrational 2PT term, which in that work did not include a fluidicity term, as noted earlier [40]. However, later formulation of such a term [51] applied to ethane, methanol and hexane shows that the torsional fluidicity is only a few percent of the vibrational term, thus not being responsible for the difference with MCC.

Table 2.6: MCC Entropy Components (J K^{-1} mol^{-1}) for OPLS and GAFF

Liquid	$S_{\mathrm{M}}^{\mathrm{transvib}}$		$S_{\mathrm{M}}^{\mathrm{rovib}}$		$S_{\mathrm{M}}^{\mathrm{topo}}$		$S_{\mathrm{UA}}^{\mathrm{transvib}}$		$S_{\mathrm{UA}}^{\mathrm{rotvib}}$		$S_{\mathrm{UA}}^{\mathrm{topo}}$	
	OPLS	**GAFF**	**OPLS**	**GAFF**	**OPLS**	**GAFF**	**OPLS**	**GAFF**	**OPLS**	**GAFF**	**OPLS**	**GAFF**
acetic acid	58.3	58.5	52.0	51.8	30.0	29.5	6.7	12.2	29.6	28.2		
acetone	65.4	67.3	58.3	59.5	25.2	23.5	7.4	11.3	45.9	44.7		
acetonitrile		63.9		38.7		16.0		1.2		22.7		
ammonia	34.8	41.7	14.5	20.4	22.0	21.1						
aniline	66.7	67.0	56.1	56.2	26.6	26.9	16.1	16.8	39.5	38.3		
benzene	73.1	74.1	61.5	61.4	11.7	10.8	7.9	8.6	28.7	27.0		
benzyl alcohol	68.2	66.4	59.4	57.6	22.1	21.9	22.4	22.7	43.6	39.8		
benzaldehyde	70.2	71.6	60.0	59.9	22.9	21.5	21.5	23.6	29.2	27.4		

butane	67.7	68.5	57.1	58.7	23.7	23.6	10.8	8.7	51.3	45.1	3.7	6.9
butanol	64.1	62.9	54.8	53.9	28.9	29.2	22.6	20.2	55.7	51.2	17.9	17.9
2-butaoxyethanol	64.9	65.8	55.3	59.4	24.1	23.6	60.7	54.8	72.6	67.3	15.6	29.7
carbon dioxide	53.6	50.3	42.3	40.6	13.6	14.2	1.5	1.5				
chloroform	80.6	81.5	77.3	77.4	21.3	21.3	20.7	25.3	3.3	4.4		
cyclohexane	70.2	70.1	64.2	64.2	19.0	19.2	13.3	11.6	53.4	46.8		
diazene	57.6	54.9	32.4	30.8	11.6	12.2	0.0	0.0	23.4	18.2		
dichloromethane	80.3	79.5	71.6	71.4	23.4	23.4	6.8	6.4	8.2	10.2		
diethanolamine	58.4	58.7	48.1	53.2	21.5	18.1	46.7	42.9	65.6	63.7	7.9	19.9
diethyl ether	69.3	69.7	58.8	59.4	21.9	21.5	23.1	21.9	56.5	55.9	7.7	8.0
DMFA	64.6	64.8	56.7	57.2	30.6	29.9	13.3	22.3	48.5	47.9		
DMSO	62.1	66.2	54.8	58.5	25.6	25.9	7.9	12.1	32.5	39.6		
1,4-dioxane	66.6	66.5	61.3	61.0	27.6	27.9	13.7	13.0	36.4	30.6		
ethane	55.0	56.6	30.8	32.0	12.2	12.0	0.0	0.0	27.1	26.5		
ethanol	60.7	59.7	48.7	48.2	27.8	28.0	2.9	2.6	36.9	37.0		
ethene	54.5	58.8	29.7	31.8	12.2	11.2	0.0	0.0	17.3	17.8		
ethyl acetate	66.1	67.7	58.7	59.9	28.8	27.5	35.4	33.3	51.0	49.3	14.4	13.9
ethylamine	59.8	62.0	47.4	50.9	29.8	29.6	2.2	2.3	41.0	40.0		
ethylene glycol	57.0	56.5	45.2	48.5	26.1	24.6	8.7	10.9	35.2	34.1		
formamide	56.7	56.3	44.1	45.8	31.6	31.8	1.1	1.0	17.2	18.4		
formic acid	61.1	56.6	48.3	43.5	30.5	32.1	1.3	1.0	14.8	11.7		
furan	69.5	71.7	60.1	62.2	27.0	25.3	7.5	7.1	18.5	19.3		
hexane	70.7	71.2	60.7	62.2	20.1	20.8	34.9	30.9	75.1	68.7	11.5	18.0
hexanol	65.7	64.7	56.7	56.2	25.5	26.0	45.8	41.9	73.3	66.5	20.5	26.1
hydrazine	51.4	50.0	26.4	26.9	12.0	12.1	0.0	0.0	30.4	26.9		
hydrogen peroxide	54.0	53.8	29.0	30.4	24.1	22.7	0.0	0.0	18.7	18.0		
hydrogen sulfide		60.6		40.4								
methane	40.2	42,8	33.0	35.3								
methanethiol	71.0	71.4	43.7	43.6	16.9	16.1	0.0	0.0	45.3	41.3		
methanol	58.9	58.9	32.7	32.7	14.6	14.6	0.0	0.0	33.0	33.0		
methylamine	53.0	55.8	27.6	31.0	15.8	16.4	0.0	0.0	31.3	30.3		
NMA	59.7	61.7	52.3	54.8	28.4	27.6	13.7	13.3	50.5	48.8		
octanol	66.4	65.7	57.3	57.5	22.7	23.4	71.9	66.0	90.6	81.9	26.0	36.3
pentane	70.8	71.4	60.8	62.2	21.7	22.3	22.9	20.4	66.4	61.2	7.9	12.6
pentanol	65.1	63.8	56.0	55.2	27.2	27.6	33.8	30.7	64.5	58.9	16.9	21.0
piperidine	67.8	67.0	61.0	59.6	32.9	32.6	11.8	11.5	60.8	51.5		
propane	62.2	63.3	51.5	52.8	25.6	25.5	1.2	1.0	35.7	33.7		
propanol	62.9	61.6	52.9	51.8	29.5	29.5	12.1	11.1	47.0	43.7	8.9	8.9
pyridine	69.9	70.5	59.4	60.0	27.2	25.5	8.8	9.6	25.2	23.5		
styrene	73.4	75.7	61.6	61.7	29.6	28.7	19.8	21.7	38.6	35.2		

TBA	64.1	63.1	57.2	56.3	20.1	19.3	16.8	15.1	60.1	63.7		
TFE		77.0		70.0		18.5		41.6				
tetrahydrofuran	69.5	71.8	60.1	62.3	32.7	31.1	7.5	7.1	18.5	19.3		
toluene	72.7	74.0	61.4	61.8	25.2	23.9	16.9	17.7	47.7	45.8		
triethylamine	69.4	69.9	63.2	63.7	21.9	20.9	38.7	38.6	79.3	75.5	19.2	26.1
m-xylene	71.8	73.5	61.2	62.1	24.3	22.6	24.6	25.5	66.3	64.3		
o-xylene	73.6	75.1	61.6	62.2	24.9	23.4	23.2	23.9	62.0	58.4		
p-xylene	71.8	73.7	60.8	61.7	18.4	16.5	25.6	26.3	66.2	64.3		

2.4.3 Covariance Matrices and Coordination and Dihedral Distributions

Representative plots in Figure 2.4 and Figure 2.5 show the force and torque covariance matrices respectively for the liquids using the GAFF force field. Similar to the combined force-torque matrices in earlier work [40], force covariance matrices show maximum auto-correlation along the diagonal and strong anti-correlation for bonded atoms. Correlations between more distant atoms are only evident for more rigid molecules, consistent with their lower vibrational entropy. Torque covariance matrices have weak correlations, most ranging from negligible up to a tenth of the diagonal self-correlation, consistent with the mean-field approximation made for united-atom rotation. Only very rigid molecules such as ethene display large correlations but their associated entropy is very small. Molecule-level matrices are not shown, being near-purely diagonal.

Representative $p(N_c)$ distributions of all liquids with the GAFF force field are shown in Figure 2.6. As expected for liquids, these distributions are broad and roughly Gaussian, most peaking between N_c = 5 and 10. As Equation (2.4) makes clear, larger coordination brings about larger orientational entropy. The outliers with higher coordination are the six-membered rings such as cyclohexane, piperidine and 1,4-dioxane, and carbon dioxide, versus the hydrogen-bonded molecules whose hydrogen-bonds bring about more directed interactions and lower N_C, such as methanol, diethanolamine and octanol, the last of which is slightly liquid-crystalline.

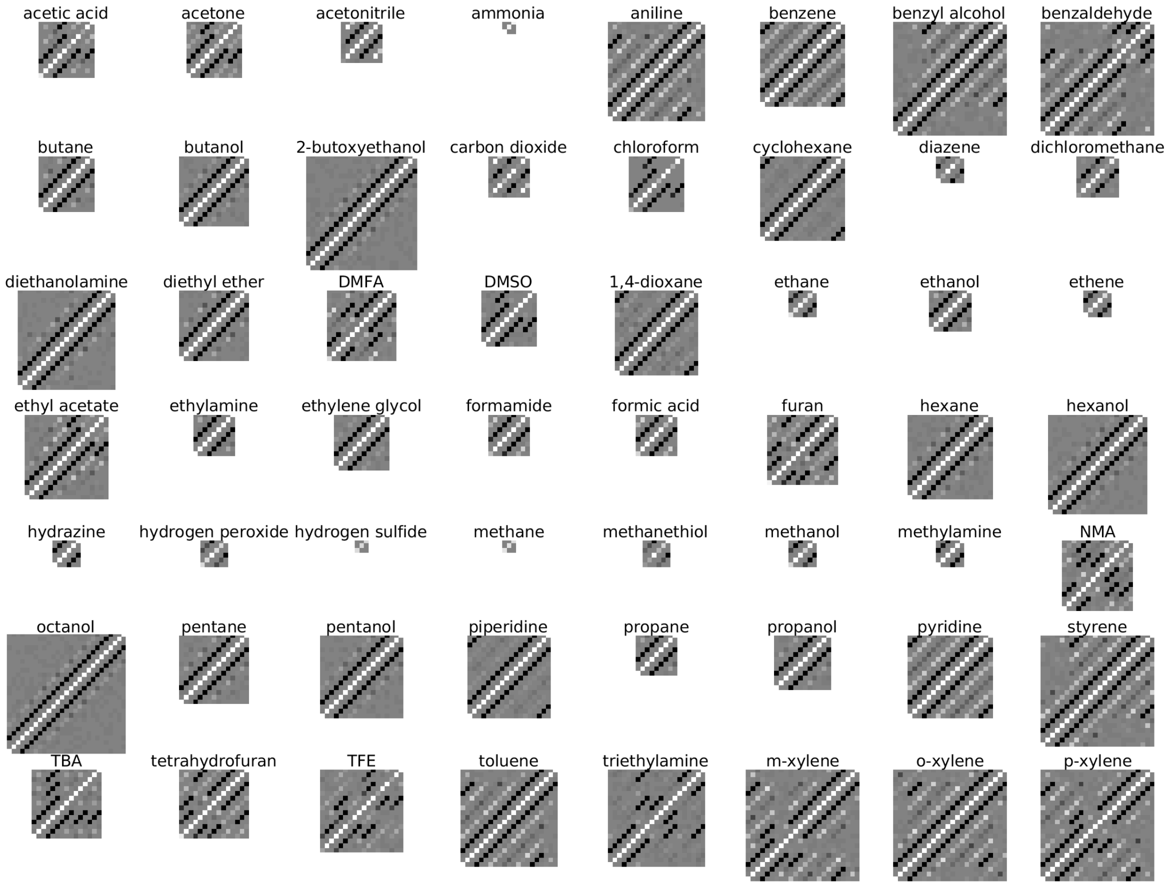


Figure 2.4. United-atom (UA) force covariance matrices for each liquid (GAFF), with the origin at the lower left. White and black represent correlations of 1 and −1, respectively, with grey in between.

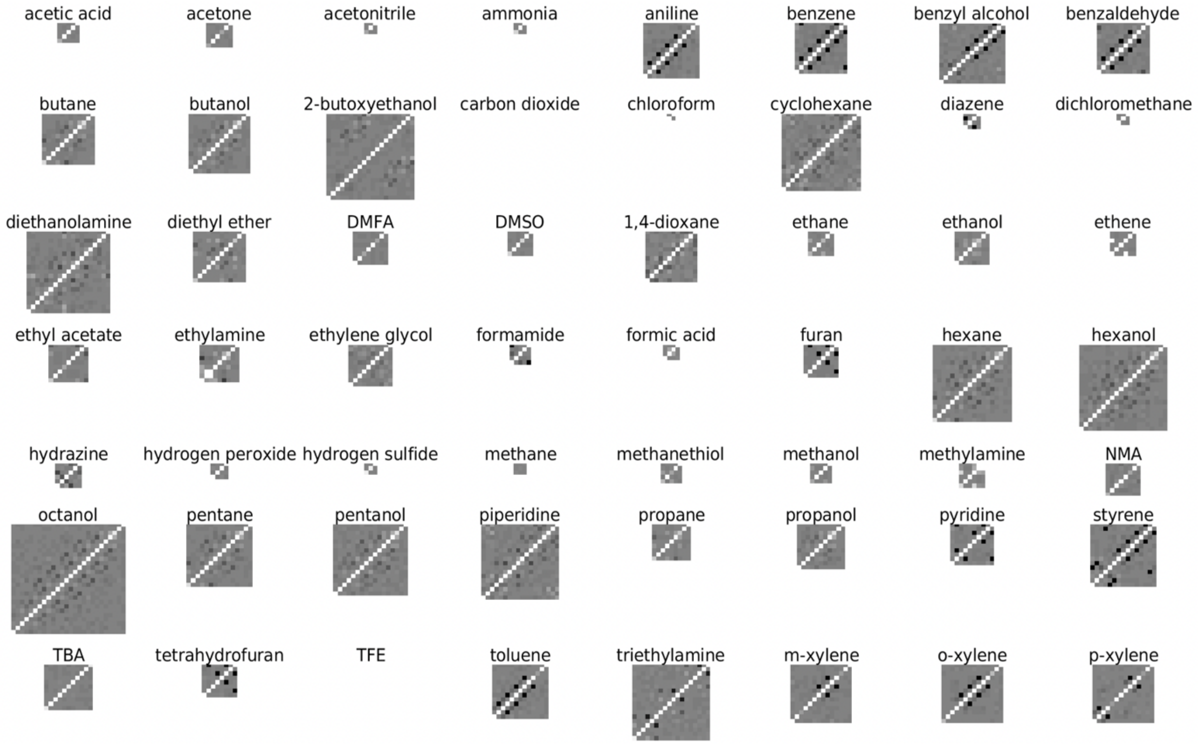


Figure 2.5. UA torque covariance matrices for each liquid and otherwise as for Figure 2.4.

Table 2.7. Conformation Probabilities [a]

Liquid	***p*(t)**		***p*(g−)**		***p*(*g*+)**	
	OPLS	**GAFF**	**OPLS**	**GAFF**	**OPLS**	**GAFF**
butane	0.88	0.69	0.88	0.15	0.06	0.15
butanol	0.24	0.23	0.43	0.43	0.43	0.34
(C terminus)	0.27	0.28	0.35	0.35	0.38	0.37
2-butaoxyethanol	0.79	0.71	0.10	0.15	0.10	0.15
(C terminus)	0.94	0.51	0.04	0.24	0.03	0.25
	0.86	0.88	0.07	0.06	0.07	0.06
	0.84	0.86	0.08	0.07	0.08	0.070
	1.00	0.04	0.00	0.48	0.00	0.48
diethanolamine	1.00	0.10	0.00	0.50	0.00	0.41
	0.88	0.94	0.06	0.03	0.06	0.03
	0.84	0.95	0.09	0.03	0.07	0.03
	1.00	0.09	0.00	0.44	0.00	0.47
diethyl ether	0.87	0.86	0.07	0.07	0.06	0.07
	0.87	0.86	0.07	0.07	0.07	0.07
ethyl acetate	1.00	1.00	0.00	0.00	0.00	0.00
(acetate C-terminus)	0.00	0.00	0.50	0.50	0.50	0.50
	0.51	0.58	0.25	0.21	0.25	0.22
hexane	0.87	0.73	0.07	0.13	0.06	0.13
	0.87	0.73	0.07	0.12	0.07	0.11
	0.86	0.75	0.07	0.13	0.07	0.12
hexanol	0.86	0.75	0.07	0.13	0.07	0.12
(C terminus)	0.87	0.76	0.07	0.12	0.07	0.11
	0.87	0.77	0.07	0.12	0.07	0.11
	0.43	0.51	0.29	0.24	0.28	0.25
octanol	0.88	0.75	0.06	0.13	0.06	0.13
(C terminus)	0.89	0.79	0.06	0.10	0.05	0.11
	0.90	0.80	0.05	0.10	0.05	0.10
	0.89	0.78	0.06	0.11	0.06	0.11
	0.88	0.77	0.06	0.11	0.06	0.12
	0.44	0.52	0.30	0.24	0.26	0.24

pentane	0.86	0.74	0.07	0.13	0.07	0.13
	0.86	0.73	0.07	0.14	0.07	0.14
pentanol	0.85	0.70	0.08	0.15	0.08	0.15
(C terminus)	0.87	0.77	0.07	0.11	0.06	0.11
	0.44	0.50	0.28	0.24	0.28	0.25
propanol	0.24	0.23	0.43	0.44	0.33	0.33
triethylamine	0.43	0.394	0.11	0.27	0.46	0.34
(C terminus)	0.43	0.389	0.09	0.25	0.49	0.36
	0.43	0.390	0.08	0.27	0.48	0.34

[a] Dihedrals are ordered sequentially from the terminus given.

The dihedral probability distributions p_i are given in Table 2.7 for the 13 molecules with united-atom dihedrals. Of the 11 molecules with more than one dihedral, the correlation matrix brings about only a small reduction in entropy relative to the ideal value for independent dihedrals, indicating that conformations in these non-ring systems are weakly correlated. The largest reductions are −4.2 and −1.0 J K^{-1} mol^{-1} for OPLS and GAFF triethylamine, followed by −1.0 and −0.4 J K^{-1} mol^{-1} for OPLS and GAFF 2-butoxyethanol and −0.6 J K^{-1} mol^{-1} for both OPLS and GAFF octanol. However, for the ring molecules, such as cyclohexane, piperidine and 1,4-dioxane, which have six fully correlated dihedrals the method correctly picks out their two possible conformers as eigenvectors with eigenvalues according to their probability, with all other eigenvalues being zero. In the short timescale here, only a few molecules in each system convert to the other conformer. Achieving equilibrium is unnecessary for cyclohexane and 1,4-dioxane because both conformers are identical and contribute no entropy. However, the equatorial and axial conformers of piperidine are distinct, with the equatorial hydrogen on the nitrogen being lower in energy by 1.7 K J mol^{-1} [72], which would increase entropy by ~5 J K^{-1} mol^{-1}.

2.5 Discussion

We have extended our MCC method to calculate entropy for a much broader range of 56 liquids than the 14 liquids studied previously [40]. To emphasise the advantages of MCC, it is simple in its theoretical formulation, informative by giving an entropy decomposition over all degrees of freedom, rapidly convergent in the number of simulation frames required, scalable to large systems with its multiscale formulation, near-general and applicable to a huge range of molecular systems, and accurate to the level of the thermal energy k_BT for the liquids studied here.

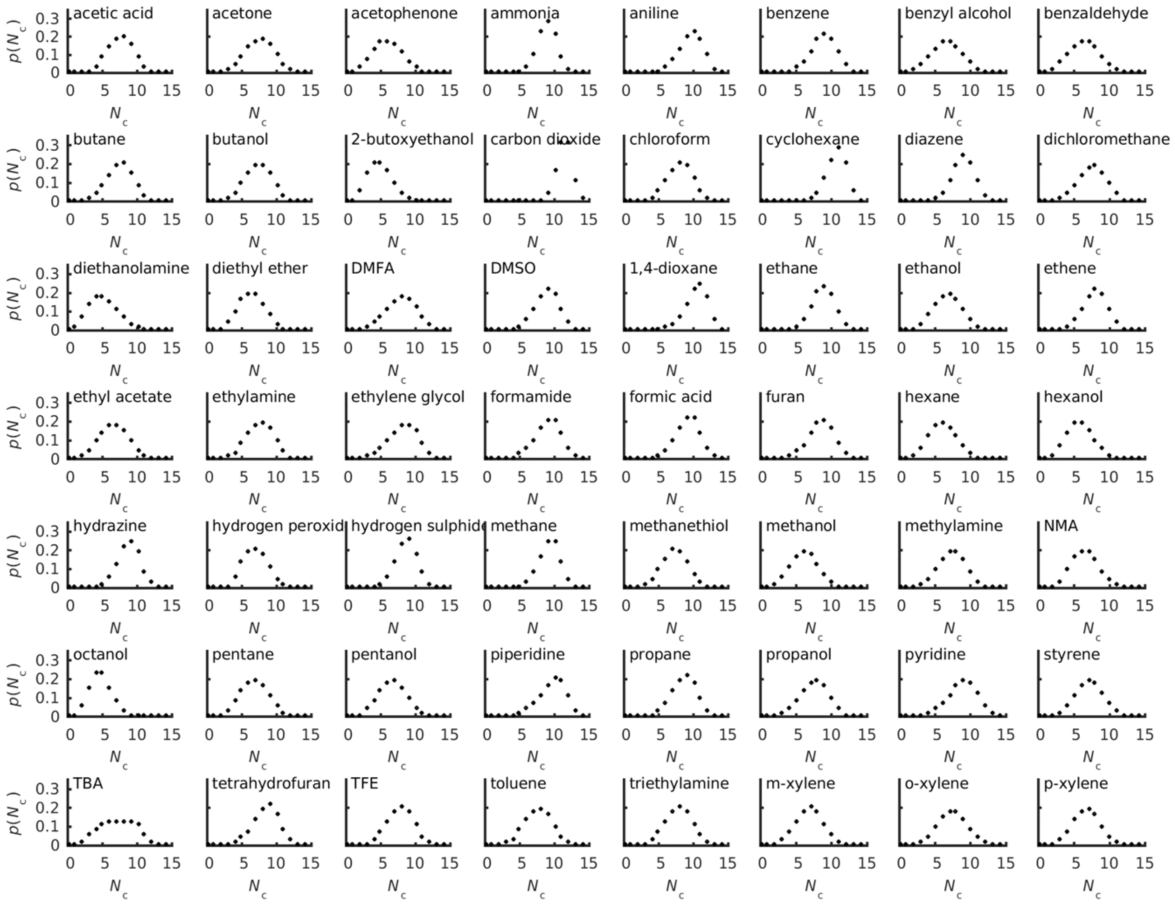


Figure 2.6. Probability distribution functions $p(N_C)$ of coordination number N_C for each liquid (GAFF).

Of the improvements incorporated in this work, the first is the recognition that the force-halving arising from the mean-field cell approximation should not be applied to bonded united atoms because of their strong correlation. This leads to lower entropies than previously [40], which is especially important for the larger molecules such as toluene or cyclohexane. The good agreement obtained earlier for single flexible molecules [39] was likely obtained due to a cancellation of errors, with the missing rotational entropy of united atoms offsetting the larger entropy due to force halving in the force covariance matrix. Nonetheless, averaged out correlations in the force and torque covariance matrices owing to conformational fluctuations may account for MCC entropies being lower than experiment for larger molecules. A more minor modification from previous work [40] relates to the use of separate force and torque covariance matrices, rather than a combined force-torque covariance matrix, owing to weak correlation between forces and torques, a change which improves the computational efficiency of the method. This work shows that subunit torques are weakly correlated in most cases, meaning that even the torque covariance may be unnecessary.

The second principle improvement is the correlation matrix to account for the correlation of dihedral conformations by expressing the conformational distribution in terms of a basis of conformers. A key feature of the correlation-matrix method is that it efficiently scales to large systems, with matrix size increasing as N_{dih}^2. Considering each conformer separately goes exponentially as $3^{N_{\mathrm{dih}}}$ and would become unfeasible beyond $N_{\mathrm{dih}} > 10$. The traditional approach using correlations in continuously valued dihedral angles has an even worse exponential dependence and goes as $N_{\mathrm{bin}}^{N_{\mathrm{dih}}}$, where N_{bin} is the number of bins. This is already problematic for $N_{\mathrm{dih}} > 2$, but it can be somewhat relieved by nearest-neighbour methods [16,22–24]. It is reasonable to assume that dihedral correlations need only be considered for local energy wells rather than for the numerical value of the dihedral, given that this correlation is unlikely to change on the timescale of molecular vibration.

A third issue to consider in future work is the multiscale approximation in how different levels of hierarchy are defined, how to avoid the double-counting of entropy between different levels of hierarchy, and how to streamline the theory further so that it is essentially equivalent at every level of hierarchy to maximise generality. Ideally, the determination of each level would be automated and dynamic, adjusting to the level of order in the system. Care is needed to ensure that the translational or rotational entropy duplicates that at the higher level for every level of hierarchy so that it is cleanly removed. The theory for vibrational entropy is already quite general for any level of hierarchy, while the topographical terms require more work to fuse Equations (2.4) and (2.8) into the same formulation. This would involve generalising the orientational entropy to be non-ideal so that orientations have different weightings according to the orientations of the neighbouring molecules, as has been already studied for water with its strongly directional hydrogen bonds [35,37,42,44]. Including the united-atom orientational entropy could be extended to other molecules such as alcohols and amines. Nonetheless, the framework is in place to scale the method to simulated systems of greater complexity.

2.6 Conclusions

We have presented the multiscale cell correlation method to calculate the entropy of 56 molecular liquids from molecular dynamics simulations. The entropies are in excellent agreement with experiment for the OPLS and GAFF force field, with GAFF performing slightly better. Agreement is better than that of the 2PT method, which can also calculate the entropy of molecular liquids. The components of entropy give an insightful and intuitive understanding of the values obtained. With suitably chosen levels of hierarchy, the method is readily scalable to larger and more complex systems.

2.7 References

[1] Peter, C., Oostenbrink, C., van Dorp, A. and van Gunsteren, W.F., Estimating entropies from molecular dynamics simulations. *J. Chem. Phys.* **2004**, *120*, 2652–2661.

[2] Zhou, H.X. and Gilson, M.K., Theory of free energy and entropy in noncovalent binding. *Chem. Rev.* **2009**, *109*, 4092–4107.

[3] van Speybroeck, V., Gani, R. and Meier, R.J., The calculation of thermodynamic properties of molecules. *Chem. Soc. Rev.* **2010**, *39*, 1764–1779.

[4] Polyansky, A.A., Zubac, R. and Zagrovic, B., Estimation of conformational entropy in protein-ligand interactions: A computational perspective. In *Computational Drug Discovery and Design*; Baron, R., Ed.; Springer: Berlin, Germany, **2012**, *819*, 327–353.

[5] Baron, R. and McCammon, J.A., Molecular recognition and ligand association. *Ann. Rev. Phys. Chem.* **2013**, *64*, 151–175.

[6] Suárez, D. and Diaz, N., Direct methods for computing single-molecule entropies from molecular simulations. *Rev. Comput. Sci.* **2015**, *5*, 1–26.

[7] Kassem, S., Ahmed, M., El-Sheikh, S. and Barakat, K.H., Entropy in bimolecular simulations: A comprehensive review of atomic fluctuations-based methods. *J. Mol. Graph. Model.* **2015**, *62*, 105–117.

[8] Butler, K.T., Walsh, A., Cheetham, A.K. and Kieslich, G., Organised chaos: Entropy in hybrid inorganic-organic systems and other materials. *Chem. Sci.* **2016**, *7*, 6316–6324.

[9] Chong, S.H., Chatterjee, P. and Ham, S. Computer simulations of intrinsically disordered proteins. *Ann. Rev. Phys. Chem.* **2017**, *68*, 117–134.

[10] Huggins, D.J., Biggin, P.C., Dämgen, M.A., Essex, J.W., Harris, S.A., Henchman, R. H., Khalid, S., Kuzmanic, A., Laughton, C.A., Michel, J., Mulholland, A.J., Rosta, E., Sansom, M.S.P. and van der Kamp, M.W., Biomolecular simulations: From dynamics and mechanisms to computational assays of biological activity. *WIREs Comput. Mol. Sci.* **2019**, *9*, e1393.

[11] Edholm, O. and Berendsen, H.J.C., Entropy estimation from simulations of non-diffusive systems. *Mol. Phys.* **1984**, *51*, 1011–1028.

[12] Wallace, D.C., On the role of density-fluctuations in the entropy of a fluid. *J. Chem. Phys.* **1987**, *87*, 2282–2284.

[13] Baranyai, A. and Evans, D.J., Direct entropy calculation from computer-simulation of liquids. *Phys. Rev. A* **1989**, *40*, 3817–3822.

[14] Lazaridis, T.and Karplus, M., Orientational correlations and entropy in liquid water. *J. Chem. Phys.* **1996**, *105*, 4294–4316.

[15] Killian, B.J., Kravitz, J.Y. and Gilson, M.K., Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.* **2007**, *127*, 024107.

[16] Hnizdo, V., Tan, J., Killian, B.J. and Gilson, M.K., Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *J. Comput. Chem.* **2008**, *29*, 1605–1614.

[17] King, B.M., Silver, N.W. and Tidor, B., Efficient calculation of molecular configurational entropies using an information theoretic approximation. *J. Phys. Chem. B* **2012**, *116*, 2891–2904.

[18] Suárez, E. and Suárez, D., Multibody local approximation: Application to conformational entropy calculations on biomolecules. *J. Chem. Phys.* **2012**, *137*, 084115.

[19] Goethe, M., Gleixner, J., Fita, I. and Rubi, J.M. Prediction of protein configurational entropy (popcoen). *J. Chem. Theory Comput.* **2018**, *14*, 1811–1819.

[20] Goethe, M., Fita, I. and Rubi, J.M. Testing the mutual information expansion of entropy with multivariate Gaussian distributions. *J. Chem. Phys.* **2017**, *147*, 224102.

[21] Suárez, E., Diaz, N. and Suárez, D., Entropy calculations of single molecules by combining the rigid-rotor and Harmonic-oscillator approximations with conformational entropy estimations from molecular dynamics simulations. *J. Chem. Theory Comput.* **2011**, *7*, 2638–2653.

[22] Hnizdo, V., Darian, E., Fedorowicz, A., Demchuk, E., Li, S. and Singh, H., Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *J. Comput. Chem.* **2007**, *28*, 655–668.

[24] Hensen, U., Lange, O.F. and Grubmüller, H., Estimating absolute configurational entropies of macromolecules: The minimally coupled subspace approach. *PLoS ONE* **2010**, *5*, e9179.

[25] Huggins, D.J., Estimating translational and orientational entropies using the k-Nearest neighbors algorithm. *J. Chem. Theory Comput.* **2014**, *10*, 3617–3625.

[26] Karplus, M. and Kushick, J.N., Methods for estimating the configuration entropy of macromolecules. *J. Am. Chem. Soc.* **1981**, *14*, 325–332.

[27] Schlitter, J., Estimation of absolute and relative entropies of macromolecules using the covariance-matrix. *Chem. Phys. Lett.* **1993**, *215*, 617–621.

[28] Andricioaei, I. and Karplus, M., On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.* **2001**, *115*, 6289–6292.

[29] Chang, C.E., Chen, W. and Gilson, M.K., Evaluating the accuracy of the quasiharmonic approximation. *J. Chem. Theory. Comput.* **2005**, *1*, 1017–1028.

[30] Reinhard, F. and Grubmüller, H., Estimation of absolute solvent and solvation shell entropies via permutation reduction. *J. Chem. Phys.* **2007**, *126*, 014102.

[31] Dinola, A., Berendsen, H.J.C. and Edholm, O., Free-energy determination of polypeptide conformations generated by molecular-dynamics. *Macromolecules* **1984**, *17*, 2044–2050.

[32] Hikiri, S., Yoshidome, T. and Ikeguchi, M., Computational methods for configurational entropy using internal and cartesian coordinates. *J. Chem. Theory Comput.* **2016**, *12*, 5990–6000.

[33] Gyimesi, G., Zavodszky, P. and Szilagyi, A., Calculation of configurational entropy differences from conformational ensembles using Gaussian mixtures. *J. Chem. Theory Comput.* **2017**, *13*, 29–41.

[34] Lin, S.T., Blanco, M. and Goddard, W.A., The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids. *J. Chem. Phys.* **2003**, *119*, 11792–11805.

[35] Henchman, R.H., Partition function for a simple liquid using cell theory parametrized by computer simulation. *J. Chem. Phys.* **2003**, *119*, 400–406.

[36] Henchman, R.H., Free energy of liquid water from a computer simulation via cell theory. *J. Chem. Phys.* **2007**, *126*, 064504.

[37] Klefas-Stennett, M. and Henchman, R.H., Classical and quantum Gibbs free energies and phase behavior of water using simulation and cell theory. *J. Phys. Chem. B* **2008**, *112*, 3769–3776.

[38] Henchman, R.H. and Irudayam, S.J., Topological hydrogen-bond definition to characterize the structure and dynamics of liquid water. *J. Phys. Chem. B* **2010**, *114*, 16792–16810.

[39] Green, J.A., Irudayam, S.J. and Henchman, R.H., Molecular interpretation of Trouton's and Hildebrand's rules for the entropy of vaporization of a liquid. *J. Chem. Thermodyn.* **2011**, *43*, 868–872.

[40] Hensen, U., Gräter, F. and Henchman, R.H., Macromolecular entropy can be accurately computed from force. *J. Chem. Theory Comput.* **2014**, *10*, 4777–4781.

[41] Higham, J., Chou, S.Y., Gräter, F. and Henchman, R.H., Entropy of flexible liquids from hierarchical force-torque covariance and coordination. *Mol. Phys.* **2018**, *116*, 1965–1976.

[42] Irudayam, S.J. and Henchman, R.H., Entropic cost of protein-ligand binding and iIts dependence on the entropy in solution. *J. Phys. Chem. B* **2009**, *113*, 5871–5884.

[43] Irudayam, S.J. and Henchman, R.H., Solvation theory to provide a molecular interpretation of the hydrophobic entropy loss of noble gas hydration. *J. Phys. Condens. Matter* **2010**, *22*, 284108.

[44] Irudayam, S.J., Plumb, R.D. and Henchman, R.H., Entropic trends in aqueous solutions of common functional groups. *Faraday Discuss.* **2010**, *145*, 467–485.

[45] Irudayam, S.J. and Henchman, R.H., Prediction and interpretation of the hydration entropies of monovalent cations and anions. *Mol. Phys.* **2011**, *109*, 37–48.

[46] Gerogiokas, G., Calabro, G., Henchman, R.H., Southey, M.W.Y., Law, R.J. and Michel, J., Prediction of small molecule hydration thermodynamics with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10*, 35–48.

[47] Michel, J., Henchman, R.H., Gerogiokas, G., Southey, M.W.Y., Mazanetz, M.P. and Law, R.J., Evaluation of host-guest binding thermodynamics of model cavities with Grid Cell Theory. *J. Chem. Theory Comput.* **2014**, *10*, 4055–4068.

[48] Gerogiokas, G., Southey, M.W.Y., Mazanetz, M.P., Heifetz, A., Bodkin, M., Law, R.J., Henchman, R.H. and Michel, J., Assessment of hydration thermodynamics at protein interfaces with Grid Cell Theory. *J. Phys. Chem. B* **2016**, *120*, 10442–10452.

[49] Pascal, T.A., Lin, S.T. and Goddard, W.A., Thermodynamics of liquids: Standard molar entropies and heat capacities of common solvents from 2PT molecular dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13*, 169–181.

[50] Lin, S.T., Maiti, P.K. and Goddard, W.A., Two-phase thermodynamic model for efficient and accurate absolute entropy of water from molecular dynamics simulations. *J. Phys. Chem. B* **2010**, *114*, 8191–8198.

[51] Huang, S.N., Pascal, T.A., Goddard, W.A., Maiti, P.K. and Lin, S.T., Absolute entropy and energy of carbon dioxide using the two-phase thermodynamic model. *J. Chem. Theory Comput.* **2011**, *7*, 1893–1901.

[52] Lai, P.K. and Lin, S.T., Rapid determination of entropy for flexible molecules in condensed phase from the two-phase thermodynamic model. *RSC Adv.* **2014**, *4*, 9522–9533.

[53] Dodda, L.S., Cabeza de Vaca, I., Tirado-Rives, J. and Jorgensen, W.L., LigParGen web server: An automatic OPLS-AA parameter generator for organic ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336.

[54] Wang, J.M., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[55] Higham, J. and Henchman, R.H., Locally adaptive method to define coordination shell. *J. Chem. Phys.* **2016**, *145*, 084108.

[56] Higham, J. and Henchman, R.H., Overcoming the limitations of cutoffs for defining atomic coordination in multicomponent systems. *J. Comput. Chem.* **2018**, *39*, 699–703.

[57] von Neumann, J. *Mathematical Foundations of Quantum Mechanics*; Princeton University Press: Princeton, NJ, USA, 1955.

[58] Case, D.A., Berryman, J.T., Betz, R.M., Cerutti, D.S., Cheatham, T.E., III, Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., et al. *AMBER 2015*; University of California: San Francisco, CA, USA, 2015.

[59] Potoff, J.J. and Siepmann, J.I., Vapor-liquid equilibria of mixtures containing alkanes, carbon dioxide, and nitrogen. *AICHE J.* **2001**, *47*, 1676–1682.

[60] Wang, J.M., Wang, W., Kollman, P.A. and Case, D.A., Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

[61] Martinez, L., Andrade, R., Birgin, E.G. and Martinez, J.M., PACKMOL: A Package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

[62] Abraham, M.J., Murtola, T., Schultz, R., Páll, S., Smith, J.C., Hess, B. and Lindahl, E., GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.

[63] Jorgensen, W.L. and Tirado-Rives, J., Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6665–6670.

[64] National Institute of Standards and Technology. Standard Reference Database Number 69. In *NIST Chemistry Webbook*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2018.

[65] Overstreet, R. and Giaque, W.F., Ammonia. The heat capacity and vapor pressure of solid and liquid. Heat of vaporization. The entropy values from thermal and spectroscopic data. *J. Am. Chem. Soc.* **1937**, *59*, 254–259.

[66] Lide, D.R., (Ed.) *CRC Handbook of Chemistry and Physics*, 99th ed.; CRC Press: Boca Raton, FL, USA, 2018.

[67] Younglove, B. and Ely, J., Thermo-physical properties of fluids. II. Methane, ethane, propane, isotutane, and normal butane. *J. Phys. Chem. Ref. Data* **1987**, *16*, 577–798.

[68] Giguère, P.A., Liu, I.D., Dugdale, J.S. and Morrison, J.A., Hydrogen peroxide—The low temperature heat capacity of the solid and the 3rd law entropy. *Can. J. Chem.* **1954**, *32*, 117–128.

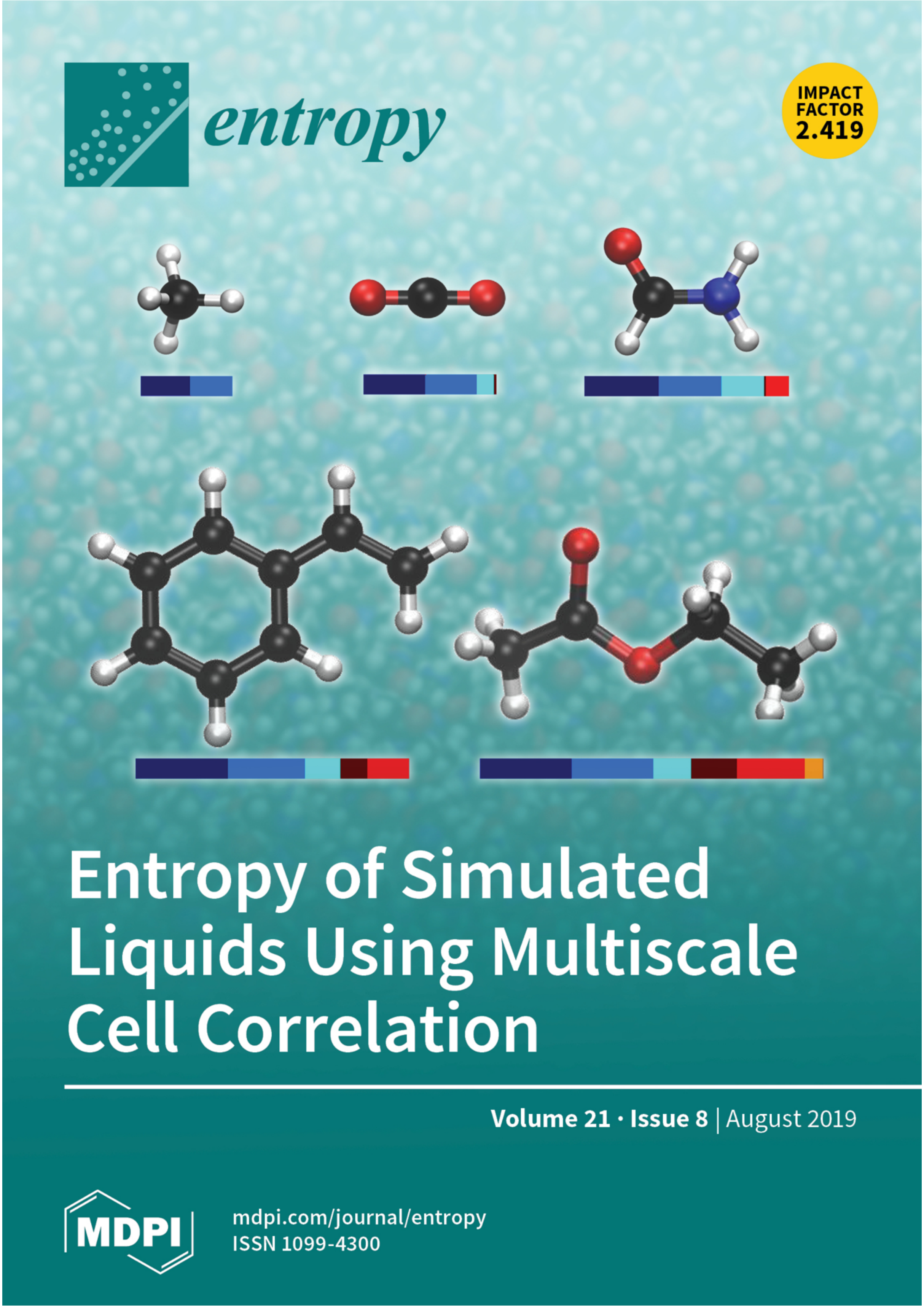
[69] Giaque, W.F. and Blue, R.W., Hydrogen sulfide. The heat capacity and vapor pressure of solid and liquid. The heat of vaporization. A comparison of thermodynamic and spectroscopic values of the entropy. *J. Am. Chem. Soc.* **1936**, *58*, 831–837.

[70] Perry, R.H., Green, D.W. and Maloney, J.O., *Perry's Chemical Engineers' Handbook*, 8th ed.; McGraw-Hill: New York, NY, USA, 2007.

[71] Stull, D.R., Westrum, E.F., Jr. and Sinke, G.C., *The Chemical Thermodynamics of Organic Compounds*; Wiley: New York, NY, USA, 1969.

[72] Caleman, C., van Maanen, P.J., Hong, M.Y., Hub, J.S., Costa, L.T. and van der Spoel, D., Force field benchmark of organic liquids: Density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. *J. Chem. Theory Comput.* **2012**, *8*, 61–74.

[73] Blackburne, I.D., Katritzky, A.R. and Takeuchi, Y., Conformation of piperidine and of derivatives with additional ring heteroatoms. *Acc. Chem. Res.* **1975**, *8*, 300–306.



entropy

IMPACT FACTOR 2.419

Entropy of Simulated Liquids Using Multiscale Cell Correlation

Volume 21 · Issue 8 | August 2019

MDPI

mdpi.com/journal/entropy
ISSN 1099-4300

Chapter 3. Comparison of Free-Energy Methods to Calculate the Barriers for the Nucleophilic Substitution of Alkyl Halides by Hydroxide

PAPER 2

Hafiz Saqib Ali[1,2], Jonathan Higham[1,2,3], Sam P. de Visser[1,4], and Richard H. Henchman[1,2]

[1] Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[2] School of Chemistry, The University of Manchester, Oxford Road, Manchester M13 9PL, UK
[3] Human Genetics Unit, Institute of Genetics & Molecular Medicine, The University of Edinburgh, Western General Hospital, Crewe Road South, Edinburgh EH4 2XU, UK
[4] Department of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

* Corresponding authors:

E-mail: henchman@manchester.ac.uk

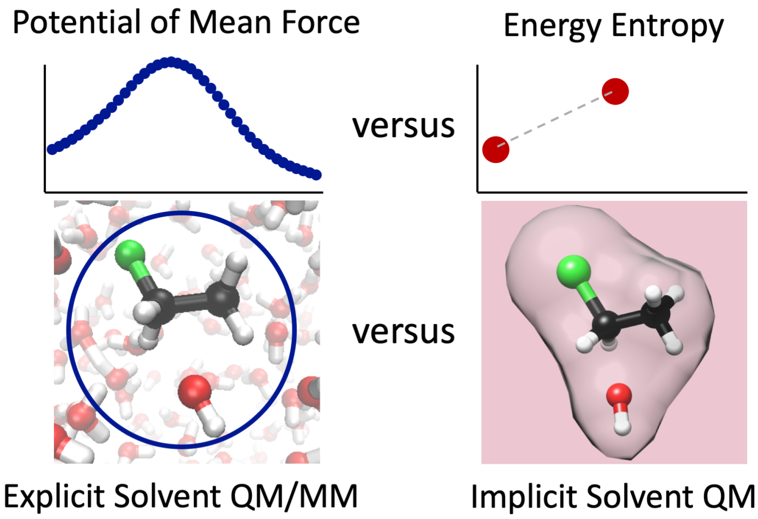
E-mail: sam.devisser@manchester.ac.uk

Conceptualization, supervision and project administration, R.H.H. and S.P.V. investigation and funding acquisition, H.S.A.; methodology and software, J.H., H.S.A. and R.H.H.; validation, visualization and writing—original draft preparation, H.S.A., S.P.V. and R.H.H.; formal analysis and writing—review and editing, H.S.A., J.H., S.P.V and R.H.H.

Published in *The Journal of Physical Chemistry B*, Volume 124, Issue 31, Page 6835-6842, 10 July 2020.

Abstract

Calculating the free-energy barriers of liquid-phase chemical reactions with explicit solvent is a considerable challenge. Most studies use the energy and entropy of minimized single-point geometries of the reactants and transition state in implicit solvent using normal mode analysis (NMA). Explicit-solvent methods instead make use of the potential of mean force (PMF). Here, we propose a new energy-entropy (EE) method to calculate the Gibbs free energy of reactants and transition states in explicit solvent by combining quantum mechanics/molecular mechanics (QM/MM) molecular dynamics simulations with multiscale cell correlation (MCC). We apply it to six nucleophilic substitution reactions of the hydroxide transfer to methyl and ethyl halides in water, where the halides are F, Cl, and Br. We compare EE-MCC Gibbs free energy barriers using two Hamiltonians, self-consistent charge density functional based tight-binding (SCC-DFTB) and B3LYP/6-31+G* density functional theory (DFT) with respective PMF values, EE-NMA values using B3LYP/6-31+G* and M06/6-31+G* DFT in implicit solvent and experimental values derived via transition state theory. The barriers using SCC-DFTB are found to agree well with the PMF and experiment and previous computational studies, being slightly higher but improving on the lower values obtained for the implicit solvent. Achieving convergence over many degrees of freedom remains a challenge for EE- MCC in explicit solvent QM/MM systems, particularly for the more expensive B3LYP/6-31+G* and M06/6 31+G* DFT methods, but the insightful decomposition of entropy over all degrees of freedom should make EE-MCC a valuable tool for deepening the understanding of chemical reactions.



3.1 Introduction

Calculating the kinetics of chemical reactions has been a major ongoing target of theoretical and computational chemistry for many decades [1-3]. This is a particular challenge in the liquid phase because it involves doing quantum mechanics calculations in a system comprising thousands of molecules. A number of different techniques have been developed for determining the kinetics of chemical reactions. The most common approach treats the reacting atoms by quantum mechanics, often at a high level [4,5], and approximates the solvent implicitly as a continuum with a dielectric constant and interfacial properties [6,7]. Implicit-solvent models include the polarizable continuum model (PCM) [8], the self-consistant isodensity PCM (SCIPCM) [9], the Onsager model [10] or the universal solvent model (SMD) [11]. Similar to the treatment of gas-phase reactions, energies and Hessians are evaluated at the minima of the reactants and transition state, corresponding to absolute zero temperature, the internal vibrational energy and entropy at the temperature of interest are accounted for using normal mode analysis (NMA), the translational and rotational energy and entropy are calculated using the ideal-gas values, and rate constants are evaluated using Transition State Theory (TST) [12,13] Energy may be understood via the many energy decomposition methods proposed [14] while entropy is commonly interpreted in terms of normal modes for ideal gas molecules.

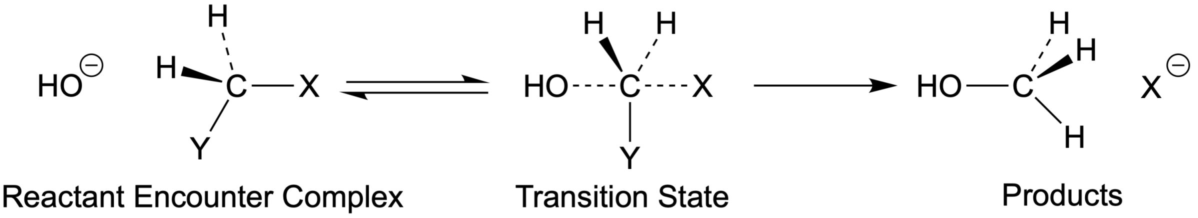
Two main advances have occurred beyond this standard implicit-solvent approach. First, explicit treatment of the surrounding non-reacting atoms was made possible by the Quantum Mechanics / Molecular Mechanics (QM/MM) method [15], which treats the small reacting region with QM and the rest of the system with MM [16]. Second, moving away from single-point minimum calculations, extensive sampling of the thermalized ensemble was made possible with simulation methods such as molecular dynamics (MD) or Monte Carlo. While energy and enthalpy can be directly calculated from the average system Hamiltonian, existing entropy methods such as NMA are not easily applicable to multi-molecular systems, given the huge number of minima [17], meaning that the direct calculation of free energy from energy and entropy has not been possible for explicit-solvent systems. Other free-energy methods are used instead, such as Umbrella Sampling (US) [18], which yields the free energy as a function of the reaction coordinate, also known as the Potential of Mean Force (PMF). Quantitatively accurate free-energy barriers can be extracted from the PMF. Entropy barriers can be obtained using the restraint-release method [19], from the temperature dependence of the PMF [20-23], or from the free-energy difference using a number of different formulations [24]. However, these methods do not produce absolute entropy over all degrees of freedom and provide little direct understanding at the molecular level in the way that the energy-entropy NMA method does in implicit-solvent. While energy typically dominates kinetics in gas-phase reactions, entropy may make a more comparable contribution in solvent or with catalysts, requiring it to be better understood.

In this work we seek to address this gap in capability of calculating and understanding the free energy barrier of a chemical reaction directly from the energy and entropy in explicit solvent. We compare implicit-solvent QM and explicit-solvent QM/MM methods to determine the reaction kinetics for the model reaction of second-order nucleophilic substitution of alkyl halides reacting with hydroxide OH^{-1} in water. Nucleophilic substitution, having equation $X^- + RY \rightarrow RX + Y^-$ where X = halide, R = alkyl group and Y = leaving group, is an extensively studied reaction [25-30], via a single transition state with a concerted ligand switch. Most theoretical studies have been in the gas phase [25,26], others in implicit solvent [27,28], and more recent studies in explicit solvent [31,32]. We employ four different Hamiltonians and three free-energy methods. The four Hamiltonians are QM density functional theory (DFT) with the B3LYP functional (HIB) and M06 (HIM) in implicit water solvent, QM/MM with self-consistent charge density functional based tight-binding (SCC-DFTB) [33,34] in explicit solvent (HES) and QM/MM with B3LYP in explicit solvent (HEB). The first free-energy method, which we designate the Energy Entropy Normal Mode Analysis (EE-NMA), uses the standard energy and entropy evaluated using NMA added to the energy minimum of the reactant encounter complex (REC) and transition state (TS). The second method uses the barrier of the PMF along a pre-defined reaction coordinate. The third method, newly proposed here and termed Energy Entropy Multiscale Cell Correlation (EE-MCC) calculates the barrier from the energy and entropy of the REC and TS, where the TS is taken as the maximum of the PMF. The energies are calculated from the average of the simulation Hamiltonian and the entropies are calculated using MCC [35,36]. Experimental rate constants for the methyl halide reactions are converted into free energy barriers using TST as a point of comparison. Both PMF methods are found to give the best agreement with experiment, the SCC-DFTB being slightly better than B3LYP. EE-MCC SCC-DFTB values are in satisfactory agreement, performing better than EE-NMA B3LYP but EE-MCC B3LYP values could not be converged. EE-MCC is affected by statistical noise arising from having to calculate free energy of all molecules combined with the limited sampling possible in a multimolecular QM/MM system, but it provides substantial detail about the entropic contributions of every degree of freedom of the system.

3.2 Methods

3.2.1 Systems of Interest.

Six nucleophilic substitution reactions in water are considered. Hydroxide displaces the halide atom of CH_2XY, where X = F, Cl or Br and Y = H or CH_3. Assuming that the binding of reactants is not rate-determining, we consider the reactant encounter complex (REC) which reacts in a unimolecular process via the transition state (TS) to the product alcohol and halide ion as shown in Scheme 3.1:



Scheme 3.1. Reaction mechanism of the nucleophilic substitution of the alkyl halide CH_2XY where X = F, Cl, Br and Y = H, CH_3 with hydroxide.

3.2.2. System Hamiltonians.

HIB: QM DFT B3LYP in Implicit Solvent. The REC in implicit solvent is illustrated in Figure 3.1.

The B3LYP [37] DFT method with the 6-31+G* basis set was used. The Integral Equation Formalism variant of the Polarizable Continuum Model (IEFPCM) [7] was used to model the solvent water. The solute molecules were built with GaussView for the REC. Minima and transition states were located by full geometry optimizations assisted by initial constraint geometry scan calculations in the Gaussian 09 software package [38].

HIM: QM DFT M06 in Implicit Solvent. This is the same as HIB except that the M06 [39,40] DFT method with the 6-31+G* basis set was used.

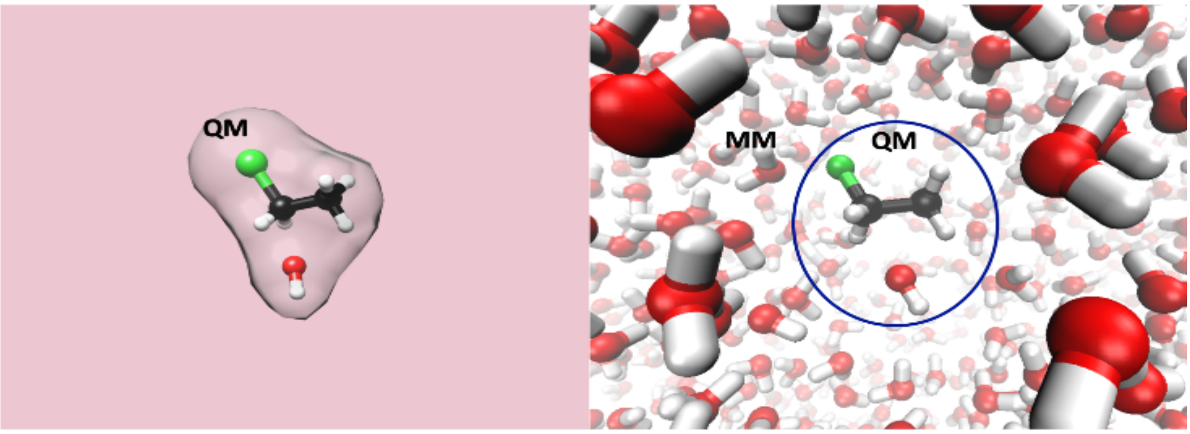


Figure 3.1. QM and MM regions for the implicit-water QM (left) and explicit-water QM/MM simulations (right).

HES: QM/MM SCC-DFTB in Explicit Solvent. The alkyl halide and hydroxide comprise the QM region where the chemical reaction takes place and the water solvent comprises the MM region, as illustrated in Figure 3.1. The QM region was modeled with the Self-Consistent Charge Density Functional based Tight-Binding (SCC-DFTB) [33] method implemented in AMBER 16 [41]. SCC-DFTB is less expensive than other methods such as B3LYP and has good accuracy for structure and relative energy [42]. The AMBER formulation contained the necessary parameters for organic molecules [43] and the halogen-related parameters [44] were taken from the DFTB website [45]. The MM region comprises explicit solvent TIP3P water molecules [9]. The reactants interact with the MM region using the fixed van der Waals parameters from the Generalized AMBER Force Field (GAFF) [46] generated by Antechamber [47] and variable configuration-specific RESP (Restrained Electrostatic Potential) charges for the electrostatic interaction using the AM1-BCC method. The reactant solute molecules were solvated with 1500 water molecules in a cubic 40 Å box using the xleap module of AMBER [41].

Minimization of the REC of each system was done with 500 steps of steepest-descent minimization, followed by a 100 ps NVT (constant number, volume temperature) MD simulation and a 100 ps NPT simulation (constant pressure) with a 1 ps time constant. The resulting structure was used as the starting structure in each window in the US simulations, to be discussed in the Free Energy Methods section, enabling the simulations of each window to be run in parallel. For each window, there were 2000 steps of minimization, 100 ps of NPT MD equilibration, and 1 ns of MD data collection for the PMF calculations. Taking the TS at the maximum in this PMF, the REC and TS windows were each run for 10 ns to provide better sampling for the EE method. Forces and coordinates were saved every 1 ps. The simulations were performed at 1 bar pressure using the Berendsen barostat [48] and 298.15 K temperature using the Langevin thermostat. Periodic boundary conditions were used together with a non-bonded cut-off of 8 Å, particle-mesh Ewald with default AMBER parameters, and a time step of 1 fs.

HEB: QM/MM B3LYP DFT in Explicit Solvent. The same QM method was used as in HIB, namely DFT with the B3LYP hybrid density functional and 6-31+G* basis set for all atoms. This is implemented in AMBER with Gaussian as an external tool for the QM part [49]. The same GAFF-TIP3P MM force field and simulation protocol are used as in HEB, except that 10 ps of sampling was used for data collection in the US simulations because this is a much slower and computationally demanding method.

3.2.3 Free Energy Methods

Energy Entropy-Normal Mode Analysis (EE-NMA). The Gibbs free energy barrier was calculated using

$$\Delta G = H_{\mathrm{TS}} - H_{\mathrm{REC}} - T(S_{\mathrm{TS}} - S_{\mathrm{REC}}) \tag{3.1}$$

where T is the temperature and H_{TS} and H_{REC} are the enthalpies of the TS and REC, and S_{TS} and S_{REC} are the corresponding entropies. For an ideal-gas molecular complex at the level of molecular translation and rotation, all but the rotational entropy cancels in the difference between REC and TS, which is given by

$$S_{\mathrm{rot}} = \frac{8\pi^2}{\sigma} \prod_{i=x,y,z} \frac{h}{\sqrt{2\pi I_i k_{\mathrm{B}} T}} \tag{3.2}$$

where σ is the symmetry number, taken as 1 for all reactions, h is Planck's constant, I_i are the moments of inertia for the three principal axes x, y, z, and k_{B} is Boltzmann's constant. The equations for the intramolecular enthalpy and entropy of the REC are

$$H = E_0 + PV + \sum_{i=1}^{3N-6} h\nu_i \left(\frac{1}{2} + \frac{1}{\mathrm{e}^{\frac{h\nu_i}{k_{\mathrm{B}}T}} - 1} \right) \tag{3.3}$$

and

$$S = \sum_{i=1}^{3N-6}\left(\frac{h\nu_i}{T}\frac{1}{\mathrm{e}^{\frac{h\nu_i}{k_\mathrm{B}T}}-1} - k_\mathrm{B}\ln\left(1-\mathrm{e}^{-\frac{h\nu_i}{k_\mathrm{B}T}}\right)\right) \tag{3.4}$$

where E_0 is the energy of the minimum, P is pressure, V is volume, N is the number of atoms, and ν_i are the $3N-6$ vibrational frequencies, which are calculated using NMA in Gaussian 09. The six lowest frequencies, being whole-molecule translations and rotations, are excluded, having already been accounted for. The same procedure is used for the TS except that there are $3N-7$ frequencies because the imaginary frequency along the reaction coordinate is excluded. The ideal gas value $PV = k_\mathrm{B}T$ used by Gaussian cancels between the unimolecular REC and TS.

Potential of Mean Force (PMF). The PMF along the reaction coordinate ξ is given by:

$$\mathrm{PMF}(\xi) = -k_\mathrm{B}T\ln p(\xi) \tag{3.5}$$

where $p(\xi)$ is the probability distribution of the system along ξ, which for all reactions is defined to be $\xi = R_{\mathrm{C-X}} - R_{\mathrm{C-O}}$, where $R_{\mathrm{C-X}}$ and $R_{\mathrm{C-O}}$ are the C–X and C–O bond lengths respectively. $p(\xi)$ was evaluated using Umbrella Sampling (US). ξ was divided into 31 windows separated by a 0.1 Å spacing over the range of −1.5 Å to 1.5 Å. The system was restrained at each value of ξ using a harmonic potential

$$U_i(\xi) = k_{\mathrm{US}}(\xi - \xi_i^0)^2 \tag{3.6}$$

with a force constant k_{US} of 300 kcal mol^{-1} Å^{-2} for all reactions [50,51]. The associated Gaussian distribution has a standard deviation of 0.03 Å, which is sufficient to span each window. MD simulations were used to generate probability distributions of ξ for each window, which were converted to the full probability distribution $p(\xi)$ using the weighted histogram analysis method (WHAM) [52,53] ΔG was calculated as

$$\Delta G = \mathrm{PMF}(\xi_{\mathrm{TS}}) - \mathrm{PMF}(\xi_{\mathrm{REC}}) \tag{3.7}$$

Energy Entropy-Multiscale Cell Correlation (EE-MCC). The Gibbs free energy barrier is evaluated from the enthalpy and entropy using Equation 3. 1 as in EE-NMA plus the Gibbs free energy ΔG_R for adding the restraint on the REC, which is given by

$$\Delta G_{\mathrm{R}} = k_{\mathrm{B}}T \ln\langle \exp[U_{\mathrm{REC}}(\xi)/k_{\mathrm{B}}T]\rangle_{\xi} \tag{3.8}$$

The TS is at ξ of the maximum in the HES and HEB PMFs. The enthalpy is evaluated using

$$H = E_{\mathrm{QM}} + K_{\mathrm{MM}} + U_{\mathrm{MM}} + U_{\mathrm{QM/MM}} + PV + \sum_{i=1}^{N_{\mathrm{vib}}} h\nu_i \left(\frac{1}{2} + \frac{1}{e^{\frac{h\nu_i}{k_{\mathrm{B}}T}} - 1}\right) - N_{\mathrm{vib}}k_{\mathrm{B}}T \tag{3.9}$$

where E_{QM} is the average energy of the QM region, K_{MM} and U_{MM} are the average kinetic and potential energy of the MM region, and $U_{\mathrm{QM/MM}}$ is the average interaction energy between the QM and MM regions. In the harmonic approximation, the classical energy $k_{\mathrm{B}}T$ is subtracted off for all N_{vib} vibrational degrees of freedom and replaced with the quantum energy of a harmonic oscillator with frequency ν_i, with the exception of the restrained reaction coordinate of the TS, for which the quantum energy is not included. Frequencies ν_i are calculated using MCC [35,36] at two length scales: the molecule level and the united-atom level, where a united atom is each heavy atom together with any bonded hydrogens, treated as a rigid body. Note that this approach ignores the negligible entropy in high-frequency covalent bonds involving hydrogen atoms. Three vibrations are whole-molecule translations and three are whole-molecule rotations, termed "transvibrational" and "rovibrational". Their frequencies are derived from the eigenvalues λ_i of the mass-weighted force covariance matrix and moment-of-inertia-weighted torque covariance matrix for the three principal axes of the solute complex using

$$\nu_i = \frac{1}{2\pi}\sqrt{\frac{\lambda_i}{k_{\mathrm{B}}T}} \tag{3.10}$$

There are also $3N - 6$ internal vibrations that relate to translation of the N united atoms. Their frequencies are evaluated using Equation 3.10 with the eigenvalues of the mass-weighted force covariance matrix. The remaining vibrations are rotations of the united atoms, being three if non-linear, two if linear, and none if a point. All matrices are constructed and diagonalized from the force and coordinate trajectories and connectivity information in the topology file using in-house C++ code.

S is calculated using Equation 3.4 with the same vibrational frequencies as for the vibrational energy at molecule and united-atom levels. For the TS the entropy of the vibration along the restrained reaction coordinate is excluded. The translational and rotational entropy of the solvent water is evaluated in the same way as the solute, except that the force and torque covariance matrices are averaged over all water molecules, and the total entropy is multiplied by the number of water molecules.

The rotational topographical entropy, relating to the number of solute orientations and termed "rotopographical", is calculated for the REC and TS with the equation [35,36]

$$S = k_{\mathrm{B}} \sum_{N_{\mathrm{c}}} p(N_{\mathrm{c}}) \ln \left[\max \left(1, \frac{(N_{\mathrm{c}}^3 \pi)^{1/2}}{\sigma} \right) \right] \tag{3.11}$$

where, $p(N_{\mathrm{c}})$ is the probability distribution of water coordination number N_{c} of the solute complex, and σ is its symmetry number, taken as 1 for all reactions. The term inside the logarithm is the number of solute orientations, all assumed to have equal probability. N_{c} is evaluated with the Relative Angular Distance (RAD) [54,55] algorithm using the centres of mass of the solute complex and each water molecule. A similar procedure is used for the rotational topographical entropy of the solvent, with $p(N_{\mathrm{c}})$ being averaged over all water molecules, σ = 2, and the number of orientations is divided by 4 to account for orientational correlations of hydrogen-bonded neighbours [36]. Translational topographical entropy is omitted, being constant and canceling between the REC and TS.

Transition State Theory (TST). Experimental rate constants k_{expt} from their Arrhenius parameters at 298 K for the nucleophilic substitution of methyl halides by hydroxide were converted into ΔG using the unimolecular TST equation [26,56]

$$k_{\mathrm{expt}} = \frac{k_{\mathrm{B}} T}{h} \exp \left(\frac{-\Delta G}{k_{\mathrm{B}} T} \right) \tag{3.12}$$

This assumes that the chemical step of the reaction (Scheme 3.1) is rate-determining and that k_{expt} is not influenced by the rate of reactant binding [56].

3.2.4 Error Analysis in the Calculation of Gibbs Free Energy Barriers

The dominant sources of error in the calculation of ΔG in Equation 3.1 are the energy of the system E (which enters as E_0 in Equation 3.3) and the solvent vibrational Gibbs free energy

(Equations 3.3 and 3.4). The solute entropy is negligible, being just one of 1501 molecules and the solvent topographical entropy is small.

Error in Energy: The standard error $\boldsymbol{\delta}E$ in the energy of system was calculated using

$$\delta E = \frac{\sigma_E}{\sqrt{N}} \tag{3.13}$$

where $\boldsymbol{\sigma}_E$ is the standard deviation of the energy and N is the number of independent data points, which was taken as 10000 to match the 1 ps spacing of the forces over 10 ns. The standard error for the energy barrier (Equation 3.1) is then given by

$$\delta\Delta E = \sqrt{\delta E_{\mathrm{TS}}^2 + \delta E_{\mathrm{REC}}^2} \tag{3.14}$$

Values of $\boldsymbol{\sigma}_E$ are found to lie between 68 and 69 kcal mol^{-1} for the TS and REC for all six reactions. This gives an overall error $\delta\Delta E$ of 0.98 kcal mol^{-1}.

Error in Water Vibrational Gibbs Free Energy: The standard error δG in the vibrational Gibbs free energy of water was calculated using the relationship

$$\delta G_i \approx \frac{1}{2} k_{\mathrm{B}} T \frac{\delta\langle F_i^2\rangle}{\langle F_i^2\rangle} \tag{3.15}$$

which can be derived by differentiating the relationship

$$G_i = k_{\mathrm{B}}T\ln\left(1 - \mathrm{e}^{-h\nu_i/k_{\mathrm{B}}T}\right) \approx k_B T\ln\frac{h\nu_i}{k_{\mathrm{B}}T} = k_{\mathrm{B}}T\ln\frac{h}{2\pi k_{\mathrm{B}}T}\sqrt{\frac{\lambda_i}{k_{\mathrm{B}}T}} \approx k_B T\ln\frac{h}{2\pi k_{\mathrm{B}}T}\sqrt{\frac{\langle F_i^2\rangle}{m k_{\mathrm{B}}T}} \tag{3.16}$$

where the first approximate equality is made in the limit of low frequency, which is valid for molecular vibration at room temperature, the second equality uses Equation 3.10 and the third equality uses the property of a molecular force covariance matrix that it is near-diagonal, such that the eigenvalues are $\lambda_i \approx F_i^2/m$.

It is found from a simulation of water that $\sigma_{\langle F_i^2\rangle}$ = 68 kcal^2 mol^{-2} Å^{-2} and $\langle F_i^2\rangle$ = 36 kcal^2 mol^{-2} Å^{-2}, where F_i is the isotropic force on the molecule over all orientations. Given that N = 10000, Equation 3.13 for $\langle F_i^2\rangle$ gives $\delta\langle F_i^2\rangle$ = 0.68 kcal mol^{-1} Å^{-1}, which in Equation 3.15 gives δG_i = 0.011 kcal mol^{-1}.

Taking the torques to depend linearly on forces and assuming isotropic force and torque, this variance applies to all six degrees of freedom, thereby introducing a factor $\sqrt{6}$ for the molecule. Given that F_i is averaged over 1500 molecules but then scaled multiplicatively to 1500 molecules, this introduces a further scaling of $1500/\sqrt{1500}$. Thus, the error in vibrational Gibbs free energy for all water is given by $\delta G = \delta G_i \times \sqrt{6} \times \sqrt{1500}$ = 0.53 kcal mol^{-1}. Using Equation 3.14 but for Gibbs free energy, $\delta \Delta G$ = 0.75 kcal mol^{-1}.

3.3 Results

3.3.1 Gibbs Free Energy Barriers.

ΔG values are given in Table 3.1 for the HIB and HIM Hamiltonians using the EE-NMA method, both HES and HEB Hamiltonians using the PMF method, the HES Hamiltonian using the EE-MCC method, and experimental rate constants converted into ΔG using TST. The value of ξ at the TS is not known from experiment but its ΔG value is placed at the HES PMF maximum for convenience. The EE-MCC ΔG values for the HEB simulations are not included because the associated energies and entropies were not converged over 10 ps. Figure 3.2 shows the HES and HEB PMFs for all reactions and both EE Gibbs free energies of the TS relative to the REC. The HES EE-MCC ΔG values take the TS at the maximum of the corresponding HES PMF, which for each reaction are at ξ = –0.1, 0.3, 0.5, –0.1, 0.2 and 0.5 Å respectively.

The TSs occur at larger values of ξ for the larger halides because of the longer C-X bond. They are slightly closer to the REC for HEB than for HES and for HIM than for HIB. The most accurate method to calculate ΔG with respect to experiment is HES PMF, whose values are only a few kcal mol^{-1} lower. However, it predicts a slightly higher barrier for CH_3Cl the CH_3F. The barriers by the HEB PMF method are smaller by a few kcal mol^{-1} and decrease in the order F, Cl and Br, which is the same trend as in experiment and as reported elsewhere, [56-59] and aligns with the strengths of the C–X bonds: the bond dissociation energies of C–F are 110 and 108 kcal mol^{-1} for the methyl and ethyl halide, respectively, of C–Cl they are 85 kcal mol^{-1} and 80 kcal mol^{-1} and of C–Br they are 71 and 68 kcal mol^{-1} [60].

Table 3.1. Gibbs Free Energy Barriers for Each Reaction and Method versus Experiment (kcal mol^{-1}).

System	Implicit Water		Explicit Water			Experiment/TST [57,58]
	$\Delta G^{\mathrm{HIB}}_{\mathrm{EE-NMA}}$	$\Delta G^{\mathrm{HIM}}_{\mathrm{EE-NMA}}$	$\Delta G^{\mathrm{HES}}_{\mathrm{PMF}}$	$\Delta G^{\mathrm{HEB}}_{\mathrm{PMF}}$	$\Delta G^{\mathrm{HES}}_{\mathrm{EE-MCC}}$	ΔG
CH_3F	16.4	15.1	20.0	18.4	21.6 ± 1.2	25.9 ± 0.1
CH_3Cl	10.1	9.9	22.3	17.0	17.2 ± 1.2	24.5 ± 0.1
CH_3Br	7.5	8.7	19.1	15.3	31.0 ± 1.2	22.7 ± 0.2
C_2H_5F	20.5	20.2	21.0	17.6	25.9 ± 1.2	-
C_2H_5Cl	10.5	11.9	18.4	14.8	22.1 ± 1.2	-
C_2H_5Br	10.1	10.9	18.5	11.3	31.6± 1.2	-

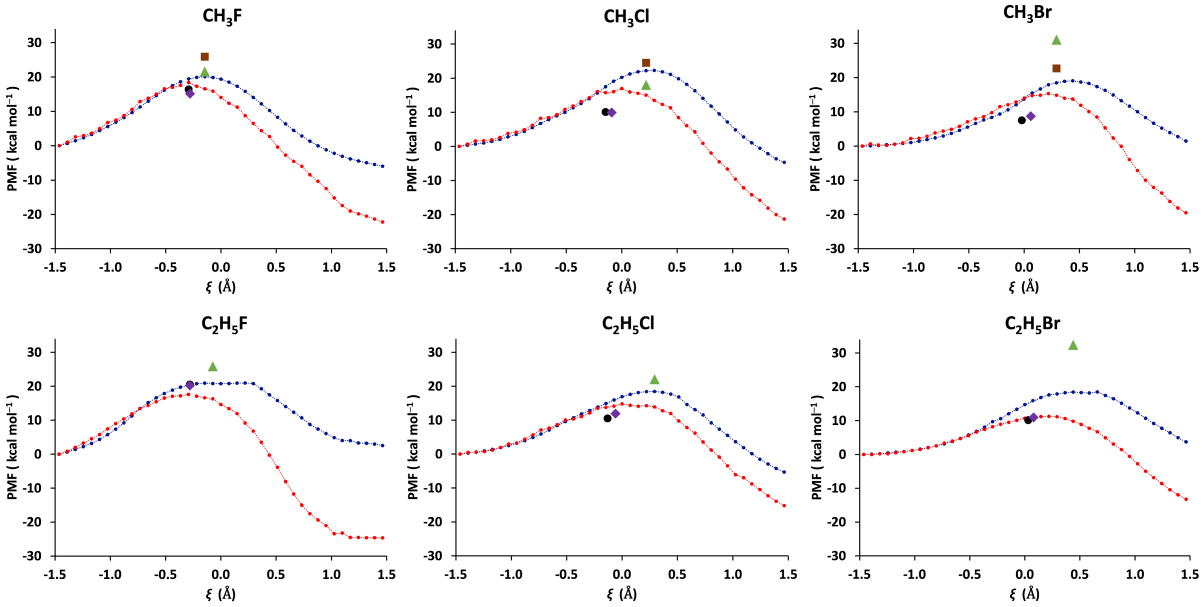


Figure 3.2. Potentials of Mean Force (PMFs) along the reaction coordinate ξ for OH^- reacting with CH_2XY (X = F, Cl, Br; Y = H, CH_3) using HES (blue) or HEB (red). Gibbs free energies using EE-NMA for HIB (black circles) or HIM (purple diamonds), EE-MCC for HES (green triangles) and experiment via TST (brown squares) of the transition state relative to the reactant encounter complex.

The lower barrier for CH_3F may be due to F^- being more strongly solvated in water than other halides [57]. The barriers for the ethyl halides are slightly lower for Cl and Br which is in line with the bond dissociation energies above [60] but higher for F. The EE-NMA ΔG barriers by both the HIB and HIM methods are much lower for methyl halides but higher for ethyl fluoride and more comparable to the PMF values. This likely reflects the inaccuracy of the implicit solvent model, because water is known to raise the Δ*G* for these reactions relative to the gas phase [56]. It also highlights the need to explicitly include solvent, despite the greater computational cost. The HES EE-MCC barriers are in the appropriate range but are more variable and display a different trend, being lower for CH_3Cl and higher for both alkyl bromides. These trends can be better understood by examining the many component quantities on which they depend, examined next.

Table 3.2. Enthalpy and Entropy Barriers (kcal mol^{-1}) of All Reactions Using the EE Methods.

	EE-NMA				EE-MCC		
	HIB		HIM		HES		
	Δ*H*	*T*ΔS	Δ*H*	*T*ΔS	Δ*H*	*T*Δ*S*	ΔG_R
CH_3F	12.5	–4.9	13.4	–3.0	25.6	4.6	0.6
CH_3Cl	6.1	–4.9	7.5	–2.1	23.7	6.4	0.4
CH_3Br	5.7	–1.8	6.7	–2.8	37.2	6.3	0.6
C_2H_5F	18.9	–0.8	17.9	–4.0	26.5	1.2	0.7
C_2H_5Cl	9.4	–2.1	10.0	–1.1	25.9	4.1	0.7
C_2H_5Br	8.1	–1.5	10.4	–1.2	35.9	3.8	0.7

3.3.2 Enthalpy and Entropy Components.

The enthalpy and entropy components of the TS minus the REC are given in Table 3.2 for each reaction and EE method. As expected for most chemical reactions, the enthalpy change dominates the entropy change. Δ*H* clearly explains the greater Δ*G* in explicit solvent, especially for both the bromohalides. Table 3.3 shows that the QM energy is large and positive, which is consistent with the destabilization of the intramolecular dipole by water [56,59] but it is partially compensated by the QM/MM and MM energy which together are negative and stabilising. The vibrational energy and Δ*PV* contributions are small, as are the Gibbs free energies for removing the US restraint on the REC (Table 3.2).

Δ*S* displays opposing trends for the two solvent models, being negative in implicit solvent but positive in explicit solvent. An inspection of the entropy terms in Tables 3.4 and 3.5

indicates that the dominant contribution to this difference is the rotational entropy of the solvent, both vibrational and topographical. This implies that there is a weakening of solvent interactions in the TS, possibly because of the more delocalised charge, even though Table 3.3 indicates that the QM/MM and MM energy is stabilizing [56,59]. For the bromohalides, it is not clear what entropy term should compensate for their large ΔH but it is likely to be an even larger gain in solvent entropy, so necessitating a more refined approachs [61-64] than that used here.

Table 3.3. Energy Components of Transition States Relative to Reactant Encounter Complexes for Each Reaction by NMA and MCC Methods (kcal mol^{-1}).

	HIB		HIM		HES			
	ΔE_0	$\Delta E_{\rm vib}$	ΔE_0	$\Delta E_{\rm vib}$	$\Delta E_{\rm QM}$	$\Delta U_{\rm QM/MM}+\Delta U_{\rm MM}$	$\Delta E_{\rm vib}$	ΔPV
CH_3F	13.6	−1.0	14.7	−1.3	54.0	−29.3	1.1	−0.2
CH_3Cl	7.0	−0.9	7.1	0.3	62.9	−40.0	0.6	−0.1
CH_3Br	5.8	−0.1	7.4	−0.7	78.8	−42.4	0.4	−0.2
C_2H_5F	18.1	0.8	17.4	−0.6	47.8	−22.0	0.8	−0.1
C_2H_5Cl	10.3	−0.9	9.3	0.8	52.2	−26.0	−0.4	−0.2
C_2H_5Br	7.5	0.5	11.1	−0.7	68.4	−32.8	−0.7	−0.1

ΔE_0 is the change of energy zero for the HIB and HIM system and $\Delta E_{\rm vib}$ is the corresponding change in quantum vibrational energy. $\Delta E_{\rm QM}$ is the energy change of the QM region in the H2a system, $\Delta U_{\rm QM/MM}$ is the energy change of the QM/MM interaction, $\Delta U_{\rm MM}$ is the energy change of the MM region, and ΔPV is the change in pressure times volume.

Table 3.4. Entropy Components of Transition States Relative to Reactant Encounter Complexes for Each Reaction by NMA and MCC Methods (cal mol^{-1} K^{-1}).

	HIB			HIM			MCC$_{\rm solute}$			MCC$_{\rm solvent}$		
	$\Delta S_{\rm int}$	$\Delta S_{\rm trans}$	$\Delta S_{\rm rot}$	$\Delta S_{\rm int}$	$\Delta S_{\rm tran}$	$\Delta S_{\rm rot}$	$\Delta S_{\rm int}$	$\Delta S_{\rm tran}$	$\Delta S_{\rm rot}$	$\Delta S_{\rm int}$	$\Delta S_{\rm tran}$	$\Delta S_{\rm rot}$
CH_3F	−15.0	0	−1.4	−8.5	0	−1.5	1.8	0.3	−0.9	0	0.7	13.5
CH_3Cl	−15.0	0	−1.4	−5.6	0	−1.5	1.5	0.8	−0.9	0	−2.0	22.2
CH_3Br	−5.3	0	−0.9	−8.4	0	−0.8	2.6	1.1	−2.0	0	−1.1	20.8
C_2H_5F	−0.3	0	−2.4	−10.9	0	−2.4	2.5	−1.1	0.7	0	−3.3	5.4
C_2H_5Cl	−6.3	0	−0.6	−3.5	0	−0.2	3.4	−0.1	1.5	0	0.7	8.5
C_2H_5Br	−4.0	0	−1.2	−3.6	0	−0.4	2.8	0.5	0.5	0	−3.8	12.7

HIB and HIM are evaluated by NMA, HES is evaluated by MCC, and $\Delta S_{\rm int}$, $\Delta S_{\rm trans}$ and $\Delta S_{\rm rot}$ are the internal and molecular translational and rotational entropy changes.

Another differing trend in entropy between the implicit and explicit solvent models evident in Table 3.4 is that the HIB and HIM internal entropy terms decrease in the TS but slightly increase in the HES system. This increase occurs despite the entropy reduction for all systems partly owing to the missing degree of freedom along ξ and the formation of covalent bonding between the reactants. It would appear that the presence of explicit solvent may dampen this reduction, possibly along with the difference in functionals. The other weak but curious trend at the molecular level is that the translational entropy increases and rotational entropy decreases for the methyl halides, but the other way around for the ethyl halides.

Table 3.5. Vibrational and Topographical MCC Entropy Components of Transition States Relative to Reactant Encounter Complexes for Each Reaction (cal mol^{-1} K^{-1}).

	MCC$_{\text{solute}}$				MCC$_{\text{solvent}}$	
	ΔS_{int}		ΔS_{rot}		ΔS_{rot}	
	$\Delta S_{\text{int-transvib}}$	$\Delta S_{\text{int-rovib}}$	ΔS_{rovib}	ΔS_{rotopo}	ΔS_{rovib}	ΔS_{rotopo}
CH_3F	0.5	1.3	−1.9	1.1	7.1	6.5
CH_3Cl	0.7	0.9	−2.0	1.1	9.3	12.9
CH_3Br	1.1	1.5	−1.4	−0.3	9.6	11.1
C_2H_5F	1.1	1.4	−0.04	0.7	2.9	2.5
C_2H_5Cl	1.1	2.1	0.6	0.9	4.6	3.9
C_2H_5Br	0.9	1.9	0.7	−0.2	7.7	5.0

$\Delta S_{\text{int-transvib}}$ and $\Delta S_{\text{int-rovib}}$ are the internal vibrational entropy changes for translation and rotation, and ΔS_{rovib} and ΔS_{rotopo} are the vibrational and topographical entropy changes for molecular rotation, all evaluated for HES.

3.4 Discussion

The novel insights provided into reaction thermodynamics by EE-MCC come at the price of some accuracy and the need for sufficient sampling. There are sizeable errors for the HES simulations and converged values could not be obtained for the shorter and more expensive HEB simulations, even though reasonable PMFs were still produced. Evidently, it is much more difficult to obtain converged probability distributions over all molecular coordinates than just one. Figure 3.3 illustrates how G, TS and H using HES vary as a function of ξ. The values moderately well reproduce the reaction profile but are still with errors of ~5 kcal mol^{-1}, which, being based on 1 ns of sampling, are larger than the errors in Table 1 based on 10 ns. The inhomogeneous nature of a solution means that there are many more molecules to average over per mole of solute than in a pure liquid as was done in earlier work [35,36] compounded by the slower speed of QM/MM simulations. A minimal QM region of the only reacting molecules was adopted here to minimise the slow-down but more accurate studies should include additional water molecules, particularly for the solvation of OH^- [65,66], and possibly in an adaptive scheme to account for solvent diffusion [67,68].

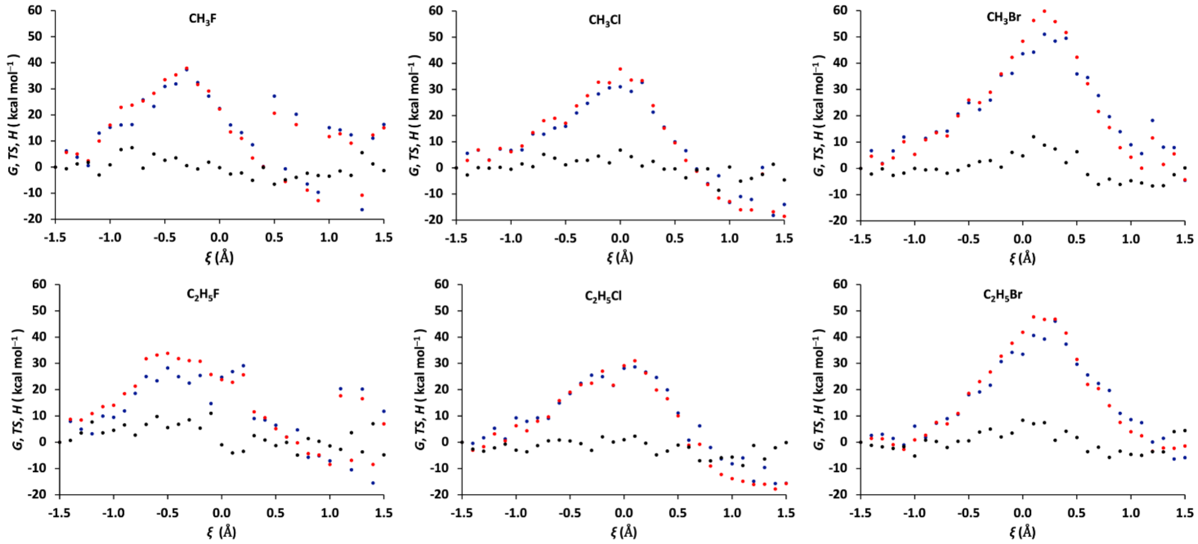


Figure 3.3: *G* (blue), *TS* (black), and *H* (red) calculated by MCC and HES versus reaction coordinate ξ for each reaction. ΔE_{vib} and ΔPV, being small (Table 3.4), are excluded.

Another problematic issue is that a PMF calculation is still needed to locate the TS and that an umbrella potential must be added to keep the system localized to the TS. However, this requirement could be alleviated by running a short series of simulations in the expected region to locate the TS followed by a longer simulation at the TS. This may be especially valuable when the PMF is difficult to converge due to a long path or there is a difficulty in identifying a suitable path. Furthermore, MCC contains a number of approximations, particularly relating to the solvent topographical entropy, in order to make tractable the calculation of the full probability distribution. However, it represents a more accurate treatment of the solvent as explicit molecules compared to a continuum model that ignores the molecular detail of the solvent and treats the solute as an ideal-gas molecule. It currently represents the state-of-the-art in liquid-phase entropy, given the limitations of more accurate methods such as inhomogeneous solvation theory [69] that require many more configurations to converge higher-dimensional integrals, limiting them to small rigid molecules.

3.5 Conclusions

A new free energy method, EE-MCC, has been proposed to calculate the Gibbs free energy barriers of chemical reactions in explicit solvent using QM/MM simulations. Energy and entropy are evaluated from the system Hamiltonian and entropy using Multiscale Cell Correlation together with the system Hamiltonian. EE-MCC has been applied to six nucleophilic substitution reactions between alkyl halides and hydroxide modelled with the two QM/MM methods SCC-DFTB and B3LYP DFT. EE-MCC SCC-DFTB Gibbs free energy barriers using are in reasonable agreement with the corresponding PMF and experiment. However, accuracy is affected by the difficulty in obtaining converged entropy and energy over many molecules in an expensive QM/MM simulation. EE-MCC values are better than implicit-solvent values using NMA but this is primarily due to the more accurate explicit-solvent energy. EE-MCC still requires the use of a PMF to identify the TS, but its primary advantages for chemical reactions are the direct route to Gibbs free energy, as done in implicit solvent, and the insightful entropy decomposition that has not previously been available in explicit solvent chemical reactions. This capability should be valuable in liquid-phase and catalyzed reactions where entropy is expected to play a larger role more comparable to that of enthalpy in determining the kinetics of chemical reactions.

3.6 References

[1] Schaleger, L.L. and Long, F.A., Entropies of activation and mechanisms of reactions in solution. *Adv. Phys. Org. Chem.* **1963**, 1, 1-33.

[2] Mardirossian, N. and Head-Gordon, M., Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017,** *115*, 2315-2372.

[3] Hu, H. and Yang, W., Free energies of chemical reactions in solution and in enzymes with ab initio quantum mechanics/molecular mechanics methods. *Annu. Rev. Phys. Chem.* **2008,** *59*, 573-601.

[4] Zhang, J., Zhang, H., Wu, T., Wang, Q. and van der Spoel, D., Comparison of implicit and explicit solvent models for the calculation of solvation free energy in organic solvents. *J. Chem. Theory Comput.* **2017,** *13*, 1034-1043.

[5] Kundi, V. and Ho, J., Predicting octanol–water partition coefficients: Are quantum mechanical implicit solvent models better than empirical fragment-based methods? *J. Phys. Chem. B* **2019,** *123*, 6810-6822.

[6] Ren, P., Chun, J., Thomas, D.G., Schnieders, M.J., Marucho, M., Zhang, J. and Baker, N.A., Biomolecular alectrostatics and solvation: A computational perspective. *Q. Rev. Biophys.* **2012,** *45*, 427-491.

[7] Tomasi, J., Mennucci, B. and Cammi, R., Quantum mechanical continuum solvation models. *Chem. Rev.* **2005,** *105*, 2999-3094.

[8] Cossi, M. Scalmani, G., Rega, N. and Barone, V., New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution. *The J. Chem. Phys.* **2002,** *117*, 43-54.

[9] Foresman, J.B., Keith, T.A., Wiberg, K.B., Snoonian, J. and Frisch, M. J., Solvent effects. Influence of cavity shape, truncation of electrostatics, and electron correlation on ab initio reaction field calculations. *J. Phys. Chem.* **1996,** *100*, 16098-16104.

[10] Onsager, L., Electric moments of molecules in liquids. *J. Am. Chem. Soc.* **1936,** *58*, 1486-1493.

[11] Marenich, A.V., Cramer, C.J. and Truhlar, D.G., Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009,** *113*, 6378-6396.

[12] Truhlar, D.G., Garrett, B.C. and Klippenstein, S.J., Current status of transition-state theory. *J. Phys. Chem.* **1996,** *100*, 12771-12800.

[13] Laidler, K.J. and King, M.C., Development of transition-state theory. *J. Phys. Chem.* **1983,** *87*, 2657-2664.

[14] Andrés, J., Ayers, P.W., Boto, R.A., Carbó-Dorca, R., Chermette, H., Cioslowski, J., Contreras-García, J., Cooper, D.L., Frenking, G., Gatti C., et al. Nine questions on energy decomposition analysis. *J. Comput. Chem.* **2019,** *40*, 2248-2283.

[15] van der Kamp, M.W. and Mulholland, A.J., Combined quantum mechanics/molecular mechanics (QM/MM) methods in computational enzymology. *Biochemistry* **2013,** *52*, 2708-2728.

[16] Quesne, M.G., Borowski, T. and de Visser, S.P., Quantum mechanics/molecular mechanics modeling of enzymatic processes: Caveats and breakthroughs. *Chem.: Eur. J.* **2016,** *22*, 2562-2581.

[17] Grimme, S. and Schreiner, P.R., Computational chemistry: The fate of current methods and future challenges. *Angew. Chem.* **2018,** *57*, 4170-4176.

[18] Bernardi, R.C., Melo, M.C.R. and Schulten, K., Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochm. Biophy. Acta.* **2015,** *1850*, 872-877.

[19] Villà, J., Štrajbl, M., Glennon, T.M., Sham, Y.Y., Chu, Z.T. and Warshel, A., How important are entropic contributions to enzyme catalysis? *Proc. Natl. Acad. Sci.* **2000,** *97*, 11899.

[20] Kazemi, M. and Åqvist, J., Chemical reaction mechanisms in solution from brute force computational Arrhenius plots. *Na. Commun.* **2015,** *6,* 7293.

[21] Åqvist, J. and Kamerlin, S.C.L., The conformation of a catalytic loop is central to GTPase activity on the ribosome. *Biochemistry,* **2015,** *54*, 546-556.

[22] Kazemi, M., Himo, F. and Åqvist, J., Enzyme catalysis by entropy without circe effect. *Proc. Natl. Acad. Sci.* **2016,** *11*, 2406.

[23] Åqvist, J., Kazemi, M., Isaksen, G.V. and Brandsdal, B.O., Entropy and enzyme catalysis. *Acc. Chem. Res.* **2017,** *50*, 199-207.

[24] Peter, C., Oostenbrink, C., van Dorp, A. and van Gunsteren, W.F., Estimating entropies from molecular dynamics simulations. *J. Chem. Phys.* **2004,** *120*, 2652-2661.

[25] Shojaie, F. and Dehestani, M., Vibrational mode analysis for the multichannel reaction of CH3Cl + OH. *Int. J. Quantum Chem.* **2012,** *112*, 2450-2455.

[26] Espinosa-García, J., Coitiño, E.L., González-Lafont, A. and Lluch, J.M. Reaction-path and dual-level dynamics calculations of the CH3F + OH reaction. *J. Phys. Chem. A.* **1998,** *102*, 10715-10722.

[27] Kim, Y., Cramer, C.J. and Truhlar, D.G., Steric effects and solvent effects on S_N2 reactions. *J. Phys. Chem. A* **2009,** *113*, 9109-9114.

[28] Cai, C., Tang, W., Qiao, C., Jiang, P., Lu, C., Zhao, S. and Liu, H., A reaction density functional theory study of the solvent effect in prototype S_N2 reactions in aqueous solution. *Phys. Chem. Chem. Phys.* **2019,** *21*, 24876-24883.

[29] Kubelka, J. and Bickelhaupt, F.M., Activation strain analysis of S_N2 reactions at C, N, O, and F centers. *J. Phys. Chem. A* **2017,** *121*, 885-891.

[30] Giri, S., Echegaray, E., Ayers, P.W., Nuñez, A.S., Lund, F. and Toro-Labbé, A., Insights into the mechanism of an S_N2 reaction from the reaction force and the reaction electronic flux. *J. Phys. Chem. A* **2012,** *116*, 10015-10026.

[31] Hamlin, T.A., Swart, M. and Bickelhaupt, F.M., Nucleophilic substitution (S_N2): Dependence on nucleophile, leaving group, central atom, substituents, and solvent. *Chem. Phys. Chem.* **2018,** *19*, 1315-1330.

[32] Ensing, B. and Klein, M.L., Perspective on the reactions between F– and CH3CH2F: The free energy landscape of the E_2 and S_N2 reaction channels. *Proc. Natl. Acad. Sci. U.S.A.* **2005,** *102*, 6755.

[33] Elstner, M., Porezag, D., Jungnickel, G., Elsner, J., Haugk, M., Frauenheim, T., Suhai, S. and Seifert, G. Self consistent charge density functional tight binding method for simulations of complex materials properties. *Phys. Rev. B,* **1998,** *58*, 7260-7268.

[34] Miriyala, V.M. and Řezáč, J. Description of non-covalent interactions in SCC-DFTB methods. *J. Compu. Chem.* **2017,** *38*, 688-697.

[35] Ali, S.H., Higham, J. and Henchman, H.R., Entropy of simulated liquids using multiscale cell correlation. *Entropy* **2019,** *21*, 750.

[36] Higham, J., Chou, S.Y., Gräter, F. and Henchman, R.H., Entropy of flexible liquids from hierarchical force–torque covariance and coordination. *Mol. Phys.* **2018,** *116,* 1965-1976.

[37] Tirado-Rives, J. and Jorgensen, W.L., Performance of B3LYP density functional methods for a large set of organic molecules. *J. Chem. Theory Comput.* **2008,** *4*, 297-306.

[38] 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., et al., Fox, Gaussian, Inc., Wallingford CT, 2010.

[39] Zhao, Y. and Truhlar, D.G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: Two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008,** *120*, 215-241.

[40] Zhao, Y. and Truhlar, D.G., Density functionals with broad applicability in chemistry. *Acc. Chem. Res.* **2008,** *41*, 157-167.

[41] Case, D.A., Ben-Shalom, I.Y., Cerutti, D.S., Cheatham, T.E., III, Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Homeyer, N., et al., AMBER 2018. *University of California, San Francisco.* **2018**.

[42] Senn, H.M. and Thiel, W., QM/MM studies of enzymes. *Curr. Opin. Chem. Biol.* **2007,** *11*, 182-187.

[43] Gaus, M., Goez, A. and Elstner, M., Parametrization and benchmark of DFTB3 for organic molecules. *J. Chem. Theory Comput.* **2013,** *9*, 338-354.

[44] Kubillus, M., Kubař T., Gaus, M., Řezáč, J. and Elstner, M., Parameterization of the DFTB3 method for Br, Ca, Cl, F, I, K, and Na in organic and biological systems. *J. Chem. Theory Comput.* **2015**, 11, 332-342.

[45] The DFTB Website. http://www.dftb.org (accessed October 10, 2018)

[46] Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25*, 1157-1174.

[47] Wang, J., Wang, W., Kollman, P.A. and Case, D.A., Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006,** *25*, 247-260.

[48] Lin, Y., Pan, D., Li, J., Zhang, L. and Shao, X., Application of Berendsen Barostat in Dissipative Particle Dynamics for Nonequilibrium Dynamic Simulation. *J. Chem. Phys.* **2017,** *146*, 124108.

[49] Götz, A.W., Clark, M.A. and Walker, R.C., An extensible interface for QM/MM molecular dynamics simulations with AMBER. *J. Comput. Chem.* **2014,** *35*, 95-108.

[50] Torrie, G.M. and Valleau, J.P., Monte carlo study of a phase-separating liquid mixture by umbrella sampling. *J. Chem. Phys.* **1977,** *66*, 1402-1408.

[51] Kästner, J., Umbrella Sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011,** *1*, 932-942.

[52] Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H. and Kollman, P.A., The weighted histogram analysis method for free energy calculations on biomolecules. The method. *J. Comput. Chem.* **1992,** *13*, 1011-1021.

[53] Souaille, M. and Roux, B., Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations. *Comput. Phys. Commun.* **2001,** *135*, 40-57.

[54] Higham, J. and Henchman, R.H., Locally adaptive method to define coordination shell. *J. Chem. Phys.* **2016,** *145*, 084108.

[55] Higham, J. and Henchman, R.H., Overcoming the limitations of cutoffs for defining atomic coordination in multicomponent systems. *J. Comput. Chem.* **2018,** *39*, 705-710.

[56] Yin, H., Wang, D. and Valiev, M., Hybrid quantum mechanical/molecular mechanics study of the S_N2 reaction of CH3Cl+OH– in water. *J. Phys. Chem. A* **2011,** *115*, 12047-12052.

[57] Glew, D.N., Moelwyn-Hughes, E.A. and Norrish, R.G.W., The kinetics of the acid and alkaline hydrolysis of methyl fluoride in water. *P. Roy. Soc. Lond. A. Math. Phys. Sci.* **1952,** *211*, 254-265.

[58] Moelwyn-Hughes, E.A. and Norrish, R.G.W., The kinetics of certain reactions between methyl halides and anions in water. *P. Roy. Soc. Lond. A. Math. Phys. Sci.* **1949,** *196*, 540-553.

[59] Chen, J., Xu, Y. and Wang, D. A., Multilayered representation, quantum mechanical and molecular mechanics study of the CH3F + OH− reaction in water. *J. Comput. Chem.* **2014,** *35*, 445-450.

[60] McMillen, D.F. and Golden, D.M. Hydrocarbon bond dissociation energies. *Annu. Rev. Phys. Chem.* **1982,** *33*, 493-532.

[61] Nguyen, C.N., Kurtzman, Y.T. and Gilson, M.K., Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys.* **2012,** *137*, 044101.

[62] Gerogiokas, G., Southey, M.W.Y., Mazanetz, M.P., Heifetz, A., Bodkin, M., Law, R. J., Henchman, R.H. and Michel, J., Assessment of hydration thermodynamics at protein interfaces with grid cell theory. *J. Phys. Chem. B* **2016,** *120*, 10442-10452.

[63] Gerogiokas, G., Calabro, G., Henchman, R.H., Southey, M.W.Y., Law, R.J. and Michel, J., Prediction of small molecule hydration thermodynamics with grid cell theory. *J. Chem. Theory Compu.* **2014,** *10*, 35-48.

[64] Pattni, V., Vasilevskaya, T., Thiel, W. and Heyden, M., Distinct protein hydration water species defined by spatially resolved spectra of intermolecular vibrations. *J. Phys. Chem. B* **2017,** *121*, 7431-7442.

[65] Choi, T.H., Liang, R., Maupin, C.M. and Voth, G.A. Application of the SCC-DFTB method to hydroxide water clusters and aqueous hydroxide solutions. *J. Phys. Chem. B* **2013,** *117*, 5165-5179.

[66] Goyal, P., Qian, H.J., Irle, S., Lu, X., Roston, D., Mori, T., Elstner, M., Cui, Q., Molecular simulation of water and hydration effects in different environments: Challenges and developments for DFTB based models. *J. Phys. Chem. B* **2014,** *118*, 11007-11027.

[67] Park, K., Götz, A.W., Walker, R.C. and Paesani, F., Application of adaptive QM/MM methods to molecular dynamics simulations of aqueous systems. *J. Chem. Theory Comput.* **2012,** *8*, 2868-2877.

[68] Duster, A.W., Wang, C.H., Garza, C.M., Miller, D.E. and Lin, H., Adaptive quantum/molecular mechanics: What have we learned, where are we, and where do we go from here? *WIRES Comput. Mol. Sci.* **2017,** *7*, e1310.

[69] Lazaridis, T., Inhomogeneous fluid approach to solvation thermodynamics. *J. Phys. Chem. B* **1998**, *102*, 3531−3541.

Chapter 4. What determines the selectivity of arginine dihydroxylation by the nonheme iron enzyme OrfP?

PAPER 3

Hafiz Saqib Ali[1,2], Richard H. Henchman[1,2], and Sam P. de Visser[1,3]

[1] Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

[2] School of Chemistry, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

[3] Department of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

* Corresponding authors:

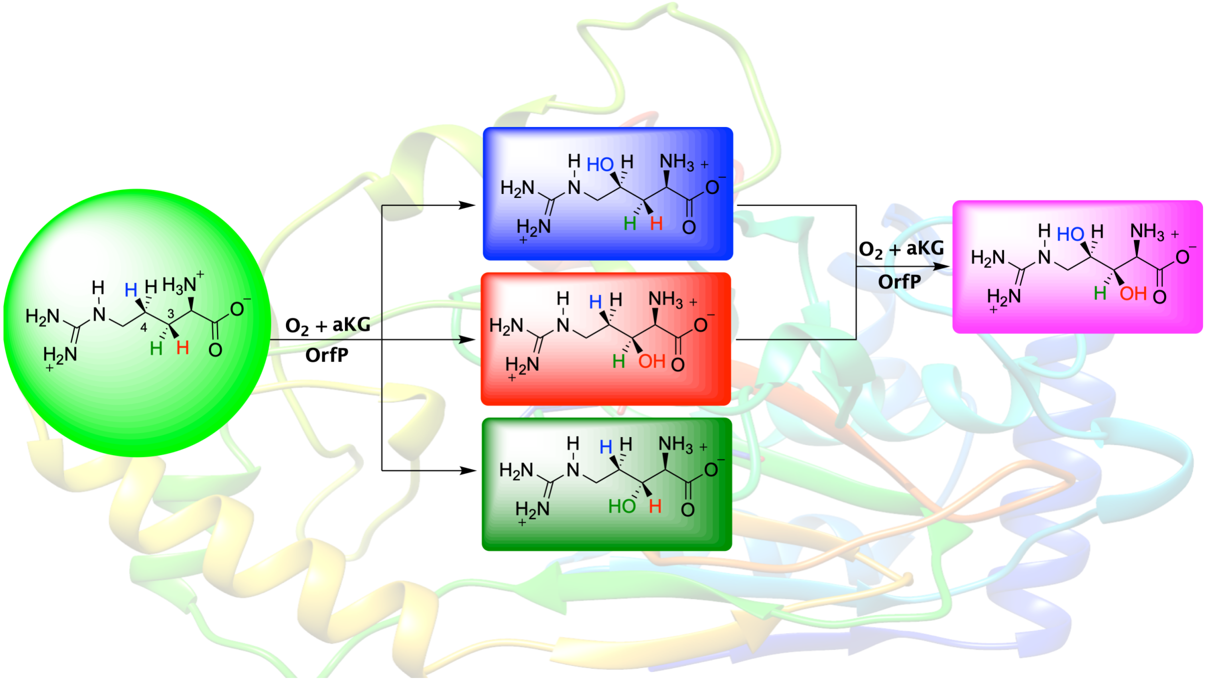
E-mail: sam.devisser@manchester.ac.uk

Conceptualization, supervision and project administration, S.P.V.; investigation and funding acquisition, H.S.A.; validation, visualization and writing—original draft preparation, H.S.A. and S.P.V.; formal analysis and writing—review and editing, H.S.A., R.H.H and S.P.V.

Published in *Chemistry — A European Journal*, Volume 27, Issue 5, Page 1795-1809, 23 September 2020.

Abstract

The nonheme iron enzyme OrfP reacts with L-Arg selectively to form the 3*R*,4*R*-dihydroxyarginine product, which in mammals can inhibit the nitric oxide synthase enzymes involved in blood pressure control. To understand the mechanisms of dioxygen activation of L-Arg by OrfP and how it enables two sequential oxidation cycles on the same substrate, we performed a density functional theory study on a large active site cluster model. We show that substrate binding and positioning in the active site guides a highly selective reaction through C^3–H hydrogen atom abstraction. This happens despite the fact that the C^3–H and C^4–H bond strengths of L-Arg are very similar. Electronic differences in the two hydrogen atom abstraction pathways drive the reaction with an initial C^3–H activation to a low-energy $^5\sigma$-pathway, while substrate positioning destabilizes the C^4–H abstraction and sends it over the higher-lying $^5\pi$-pathway. We show that substrate and monohydroxylated products are strongly bound in the substrate binding pocket and hence product release is difficult and consequently its lifetime will be long enough to trigger a second oxygenation cycle.



4.1 Introduction

Natural products including antibiotics often are build up from sugar and amino acid components [1-6]. Nature utilizes a range of enzymes to modify amino acids to give the natural products its structure and function. These biosynthesis processes often require a high regio- and chemoselectivity for the reaction, which is difficult to achieve in chemical catalysis. Understanding of how the natural product synthesis by enzymes is achieved and how this high selectivity is obtained is important for biotechnology and could enable the biosynthesis of useful products with high regio-, enantio- and stereoselectivity and minimize waste products. A class of metalloenzymes involved in the biosynthesis of many natural products are the nonheme iron dioxygenases, which are found in most organisms [7-14]. For instance, in humans the biosynthesis of the amino acid analogue 4*R*-hydroxyproline is performed by proline-4-hydroxylase enzymes in an enantio- and stereoselective reaction mechanism, which is an essential component of collagen and gives it its functional and structural properties [15-18]. Many nonheme iron dioxygenases react via a highly selective reaction mechanism, where the substrate is tightly bound in the substrate binding pocket near the nonheme iron cofactor.

Interestingly, several nonheme iron dioxygenases activate an arginine residue as part of a natural product synthesis such as in antibiotics in bacteria [1,2,19]. Thus, VioC hydroxylates a free arginine molecule selectively at the C^3-position as precursor to the biosynthesis of viomycin [20-24], while NapI desaturates an arginine substrate at the C^3–C^4 bond through two sequential hydrogen atom abstraction reactions [25]. These enzymes were studied spectroscopically and kinetically and their substrate specificity tested. Recent studies on the nonheme iron enzyme GetI showed it to hydroxylate L-Arg at the C^4-position, although it was not clear whether arginine is its natural substrate [26]. A similar situation was found in the ethylene forming enzyme that apart from catabolizing succinate to three CO_2 molecules, also appears to bind and activate an arginine residue but hydroxylate it at the C^5-position [27-29]. Therefore, several nonheme iron dioxygenases appear to activate an L-Arg substrate differently to give either C^3-hydroxylation, C^4-hydroxylation, C^5-hydroxylation or C^3–C^4 desaturation through the use of O_2 and α-ketoglutarate (αKG) on an iron center.

A recently discovered and characterized nonheme iron dioxygenase from *Streptomyces*, named OrfP, was shown to activate a free arginine amino acid and produce yet another product in a reaction with dioxygen, namely the 3*R*,4*R*-dihydroxyarginine [30]. Subsequently, the

3*R*,4*R*-dihydroxoarginine is converted into streptolidine and attached to a glucosamine sugar scaffold to form the streptothricin product. OrfP; therefore, performs the dihydroxylation of L-Arg as a precursor reaction to the biosynthesis of the streptothricin antibiotics. Isolation of OrfP and studies of its product distributions showed that the major component was 3*R*,4*R*-dihydroxyarginine; however, small amounts of 3*R*-hydroxyarginine, 3*S*-hydroxyarginine and 4*S*-hydroxyarginine, are also observed (Figure 4.1), although their origins remain unclear. Note that 3*R*,4*R*-dihydroxyarginine was found to have inhibitory effects on inducible nitric oxide synthase enzymes in mammals; an enzyme involved in blood pressure control and inflammation [31]. Hence, its selective biosynthesis pathways are important in medicinal and biotechnology work and may have important applications.

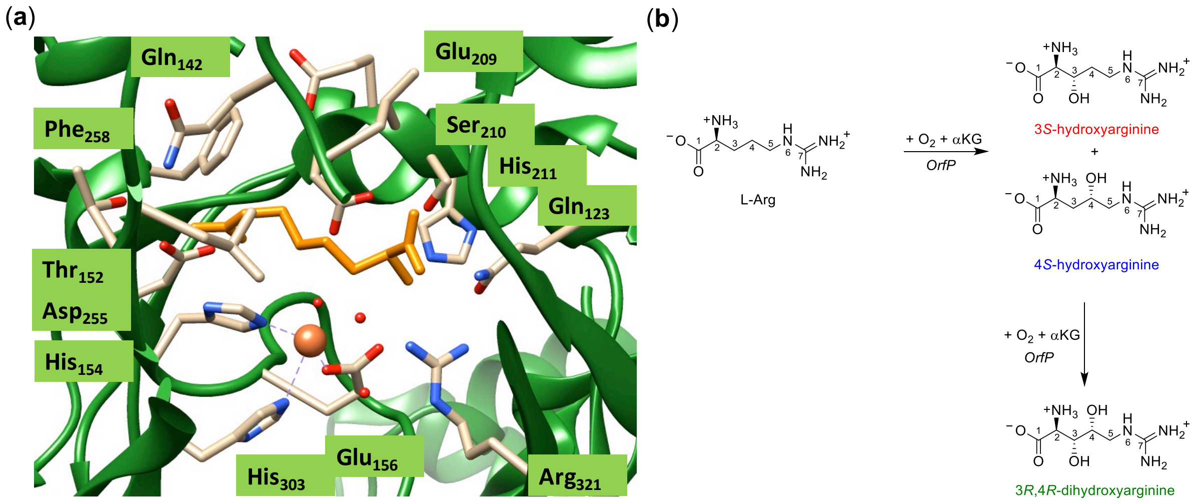


Figure 4.1. a) Extract of the crystal structure coordinates (4M2E pdb file) of L-homo-arginine (amber) bound OrfP with key residues and iron (light brown) highlighted; b) Products and by-products formed in OrfP activation of L-Arg.

An extract of the active site of OrfP is given in Figure 4.1 as taken from the 4M2E protein databank (pdb) file [32,33]. The iron is linked to the protein through interactions with the side chains of His_{154}, Glu_{156} and His_{303} group in a typical facial 2-His/1-Glu triad, which is common for many nonheme iron dioxygenases [34-36]. In the 4M2E pdb [32], the other ligand sites of the metal are occupied by water molecules. It is known that OrfP utilizes α-ketoglutarate (αKG or also called 2-oxoglutarate) and dioxygen and binds these on its iron center and presumably forms an iron(IV)-oxo active species. For several other nonheme iron dioxygenases the iron(IV)-oxo species has been characterized with UV-Vis absorption, electroparamagnetic resonance and Mössbauer spectroscopic studies and shown to be the active species that reacts with substrate [37-42].

How OrfP can perform a double dioxygenation reaction by preventing release of hydroxyarginine in favor of a second catalytic cycle is unknown. Although crystal structure coordinates of OrfP have been obtained, they do not give a clear understanding on the product release and substrate oxidation selectivity, which warrants a detailed computational study. In the OrfP pdb structure, the substrate analogue L-homo-arginine (Figure 4.1) binds tightly into the substrate-binding pocket and is surrounded by polar groups. In particular, its guanidinium group forms a salt bridge with Asp_{255} and experiences hydrogen bonding interactions from the alcohol group of Thr_{152} and the oxo group of Gln_{142}. The carboxylate group of L-homo-arginine forms a salt bridge with Arg_{321}, which is also in hydrogen bonding distance to the iron ligand Glu_{156}. The amine group of the substrate is positioned with hydrogen bonding and salt-bridge interactions to the side chains of the Gln_{123}, Asp_{208}, Ser_{210} and His_{211} residues, which will influence the reactivity of the substrate with the cofactor. Hence we reasoned that a large cluster model will be needed to describe the OrfP structure and reactivity well. In this work, a large cluster model incorporating the substrate and cofactor and their direct environment is studied to explore the reaction pathways of L-Arg activation leading to 3*S*-hydroxyarginine, 3*R*-hydroxyarginine, 4*S*-hydroxyarginine and 3*R*,4*R*-dihydroxyarginine products. Moreover, we focus on how the hydroxyarginine release is prevented in favor of a second oxygenation cycle.

4.2 Experimental Section

4.2.1 Model Set-Up

A density functional theory (DFT) cluster model was created using procedures described previously [43-46]. The 4M2E pdb file was used [32,33], whereby we selected chain A and replaced L-homo-Arg substrate manually by L-Arg and docked αKG in the structure using the SwissDock web server [47]. Subsequently, hydrogen atoms were added in Chimera at pH = 7 [48], and we made sure to check all polar residues: All Asp and Glu side-chains were deprotonated and all Arg and Lys side-chains protonated. The His residues in the protein were visually inspected for neighbouring hydrogen bonding donor and acceptor groups and based on the analysis all were chosen to be singly protonated on either the N_δ or N_ε atom. In the active site model we replaced the iron(II) ion by an iron(IV)-oxo species with a starting distance of 1.62 Å and the oxo trans to His_{303}, while αKG was replaced by succinate to create the active oxidant species in the enzyme. These changes kept the overall charge of the model the same, which was thereafter solvated with water in Chimera. The protein is solvated in 10 Å cubic box of water molecules using the TIP3P force field [49] and the system was neutralized with Na^+ of Cl^- ions.

4.2.2 Molecular Dynamics Simulation

To check that translational stability of the protein complex, a molecular dynamics (MD) simulation was performed. The MD simulation was carried out using the PMEMD (Particle Mesh Ewald Molecular Dynamics) module of the AMBER18 simulation package [50]. The GAFF [51] force field with AM1-BCC charges were used for the substrate and α-ketoglutarate while the ff14SB [52] force field was used for the protein system. The GAFF force field parameters were generated using the Antechamber [53] while the parameters for penta coordinated Fe(II) active site and its coordinated ligand were generated with the Metal Centre Parameter Builder (MCPB) using the MCPB.py [54] tool. The parameter and coordinate files of enzyme complex were generated using the xleap module in AMBER18. The equilibration of the enzyme structure was started with 500 steps of steepest decent minimization, 100 ps controlled heating from 0 to 300 K temperature at a constant volume and temperature molecular dynamic simulation with the Langevin thermostat [55] and 5 ps^{-1} collision frequency. After that the equilibration was completed with a 100 ps MD simulation at constant pressure using the Berendsen barostat [56] with 2 ps time constant. At the end, for data collection 20 ns MD simulation was performed at constant pressure and temperature with the Berendsen barostat and 2 ps time constant. The MD simulation gave very little changes to the protein structure and kept all features of the active site in tact during the full run. As such, OrfP is a rigid protein, where the active site and substrate binding pocket show little changes over time (Figure 4.2).

Next, we analysed the first and second coordination sphere of the iron and substrate environment in detail and included in the model all groups that determine the shape and constraints of the substrate and cofactor binding environments and particularly included key hydrogen bonding interactions of charged residues and salt-bridges. Thus, the cluster model includes two long protein chains that circumvent the cofactor and substrate-binding pocket, namely the peptide chains Leu_{151}-Thr_{152}-Trp_{153}-His_{154}-Thr_{155}-Glu_{156} and Asp_{208}-Glu_{209}-Ser_{210}-His_{211}. Amino acid residues pointing away from the active site where shortened to Gly residues in the model, i.e. Trp_{153}, Thr_{155} and Glu_{209}, but their protein backbone was kept in the model. The side chains of the residues Gln_{142}-Leu_{143}, Asp_{255}, Phe_{258}, and Arg_{321} were included in the model. A hydrogen atom was added to the positions where a bond was cut to restore the valencies of the atoms. The axial histidine ligand (His_{303}) of iron was shortened to methylimidazole. Finally, based on the solvated structure we decided to include two water molecules in the model: one positioned near the carboxylate group of the substrate and the other near the guanidinium group. Overall, our DFT cluster model consists of 275 atoms, has overall charge of –1 and was calculated with multiplicity singlet, triplet and quintet spin. As the model contains many internal hydrogen bonding interactions, no constraints on the system were used. An overlay of the optimized reactant structure and the original pdb file (Figure 4.3) indeed shows little differences on the overall shape and structure of the chemical system.

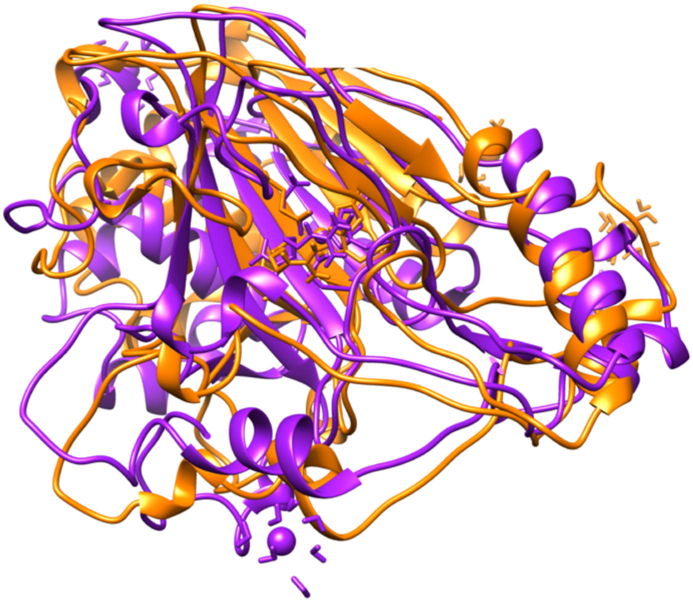


Figure 4.2. The overlay of X-ray crystal structure (in amber) and the equilibrated structure (in purple) obtained after a 20 ns MD-simulation.

4.2.3 Quantum Mechanical Calculations Procedures

The Gaussian-09 software package was used for all quantum chemical calculations [57]. Following previous experience with cluster models of nonheme iron dioxygenases [58,59], we utilized the unrestricted B3LYP density functional method for geometry optimizations, constraint geometry scans and frequency calculations. The basis set used was a LANL2DZ (with electron core potential, ECP) on the iron and 6-31G on the rest of the atoms (H, C, N, O) designated basis set BS1 [60-62]. Test calculations with alternative density functional methods were performed on analogous systems and showed little change in spin-state-ordering, optimized geometries and overall reaction mechanism [63]; hence B3LYP/BS1 was used for geometry optimizations. All local minima were verified by the presence or absence of negative eigenvalues in the vibrational frequency analysis while all the transition state structures were found using the Berny algorithm [64], and confirmed by vibrational analysis to have one imaginary mode, which was animated and shown to correspond to the reaction coordinate. In order to correct the energetics and account for the effect of solvent, single-point energies of the optimized geometries were evaluated with the UB3LYP-D3/BS2 level of theory [60-62,65], whereby BS2 is 6-311+G* on H, C, O, N and LACV3P+ (with electron core potential) on iron. The latter set of calculations included a conductor-like polarizable continuum model (CPCM) with a dielectric constant of ε = 5.7 mimicking chlorobenzene [66], which has been shown to be a good representative of an enzyme active site. For several structures the geometries were reoptimized at the UB3LYP/BS3 level of theory (basis set BS3 has LACV3P+ with core potential on iron and 6-31G* on the other atoms), but similar geometries and reaction barriers were obtained. Therefore, most of the work was done using the UB3LYP/BS2//UB3LYP/BS1 approach. Since the zero-point energies (ZPE) and vibrational entropy contributions are sometimes affected by small real frequencies for internal motions, we excluded vibrations with values below 50 cm^{-1} and recalculated the ZPE, thermal and entropy corrections. Free energies are calculated at 298.15 K and 1 atm, and include the thermal corrections evaluated from the unscaled vibrational frequencies at the UB3LYP/BS1 level of theory, with solvent, dispersion and entropy contributions.

Kinetic isotope effects (KIEs) were estimated using the classical Eyring equation 4.1 based on differences in free energy of activation ($\Delta G^{\ddagger}$) of the hydrogen and deuterium-substituted systems as described previously [67,68]. Tunneling corrections (Q_{t}) to the KIE were estimated using the Wigner correction as described in equation 4.2 and is based on the change in imaginary frequency of the transition state (ν), see equations 4.2 and 4.3. In equations 4.1 – 4.3, R is the gas constant, T is the temperature (298 K), h is Planck's constant and k_{B} is Boltzmann's constant.

$$\mathrm{KIE}_{\mathrm{Eyring}} = \exp\{(\Delta G_{\mathrm{D}}^{\ddagger} - \Delta G_{\mathrm{H}}^{\ddagger})/RT\} \tag{4.1}$$

$$\mathrm{KIE}_{\mathrm{Wigner}} = \mathrm{KIE}_{\mathrm{Eyring}} \times {}^{\mathrm{Q_{tH}}}/_{\mathrm{Q_{tD}}} \tag{4.2}$$

$$Q_t = 1 + \left({}^{hv}/_{k_B T} \right)^2 /24 \quad (4.3)$$

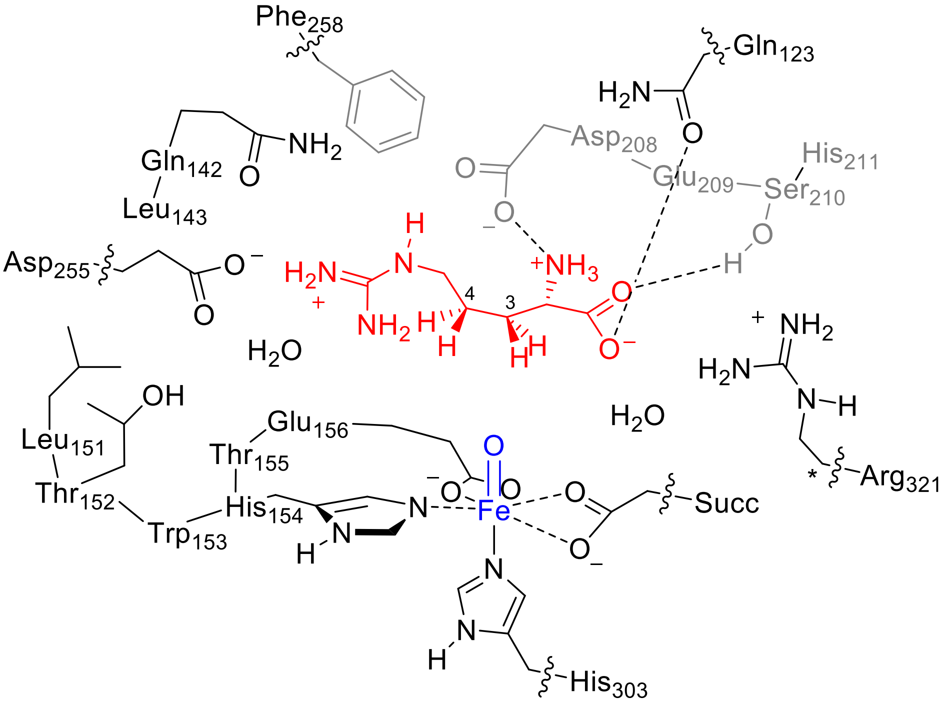
The substrate binding free energy (BFE) was calculated by taking the optimized geometries of the three hydroxyarginine complexes (**IM2**$_{C4S}$, **IM2**$_{C3S}$ and **IM2**$_{C3R}$). A single point frequency calculation at UB3LYP/BS1 with CPCM included was done on all structures. Thereafter, the structures were split into two components, namely the hydroxyarginine and the protein, i.e. **IM2** minus the hydroxyarginine. We then did a frequency at the same level of theory on both the hydroxyarginine and the protein structures. The binding free energy is evaluated from the difference in free energy of the three individual components.

4.3 Results

4.3.1 First Oxygenation Cycle of L-Arg by OrfP

We created an OrfP cluster model, see Scheme 4.1, from the crystal structure coordinates of the OrfP enzyme deposited in the protein databank under the 4M2E pdb file [32,33]. This structure represents an OrfP tetramer with L-homo-arginine as substrate mimic bound. We took chain A of the protein and replaced L-homo-arginine by L-arginine manually. The structure has iron coordinated to the side chains of two histidine groups (His$_{154}$ and His$_{303}$) and the carboxylate group of Glu$_{126}$. As α-ketoglutarate (αKG) is missing from the pdb file, it was docked in position with the SwissDock web server bound to iron in the same plane as the side chains of His$_{154}$ and Glu$_{156}$ [69]. A 20 ns molecular dynamics simulation shows the structure to be highly rigid and little changes to the active site and second-coordination sphere structure of the iron center was found. Next, a model containing the first and second coordination sphere of the metal center was created that included the oxidant with its direct ligands and the main components of the substrate (L-Arg) and succinate binding pocket: see Scheme 4.1.

Subsequently, we explored the first hydroxylation pathway of L-Arg substrate using the model described in Scheme 4.2. OrfP enzymes perform two consecutive hydroxylation reactions on L-Arg and each of these cycles uses one molecule of O_2 and a molecule of αKG to form an iron(IV)-oxo species and succinate. Spectroscopic studies showed this iron(IV)-oxo to react with L-Arg [11]. We will start with the substrate activation steps in the first catalytic cycle, where L-Arg is bound to an iron(IV)-oxo(succinate) complex and follow the reaction mechanism of substrate hydroxylation leading to 3*S*-hydroxyarginine, 3*R*-hydroxyarginine and 4*S*-hydroxyarginine products using the cluster model described in Scheme 4.1 (Model **A**). Thereafter, the second hydroxylation cycle was investigated from these mono-hydroxylated arginine product complexes by replacing the iron(II) site by an iron(IV)-oxo species (Model **B**). Model **A** and **B**; therefore, have the same overall charge. The model type is given in subscript after the label of the structure.



Scheme 4.1. DFT cluster model investigated in this work with substrate highlighted in red and iron(IV)-oxo in blue. The wiggly lines identify where the protein chain was cut and a hydrogen atom inserted.

Our cluster model of the iron(IV)-oxo species with L-Arg bound was geometry optimized in the low-lying triplet and quintet spin states (^{3,5}**Re**$_A$) using UB3LYP/BS1, while the quintet spin structure was also minimized with a larger basis set at UB3LYP/BS3 (^{5}**Re**$_{A,BS3}$). Figure 4.4 gives optimized geometries of the reactant complexes ^{3,5}**Re**$_A$ as obtained with DFT using cluster model **A**. The Fe–O distance in the quintet spin state is short and typical of a double bond, while it is somewhat larger in the triplet spin state. The axial histidine ligand is at a distance of 2.078 Å in the quintet spin state and at 2.035 Å in the triplet spin state, while a geometry optimization with a larger basis set, i.e. BS3 basis set, elongates it slightly to 2.157 Å. These structures of the nonheme iron(IV)-oxo species match previous calculations on similar species well [70-79]. Moreover, experimental studies on analogous nonheme iron enzymes, such as TauD and P4H established an Fe–O distance of about 1.62 Å and characterized them as quintet spin ground states [37-42]. Therefore, the calculations match previously reported experimental structures of enzymatic nonheme iron(IV)-oxo complexes well. Furthermore, an overlay of the $^{5}\mathbf{Re}_{\mathrm{A}}$ structure with the crystal structure coordinates shows that little changes to the structure have occurred during the geometry optimization and the model still has all features of the protein (Figure 4.3).

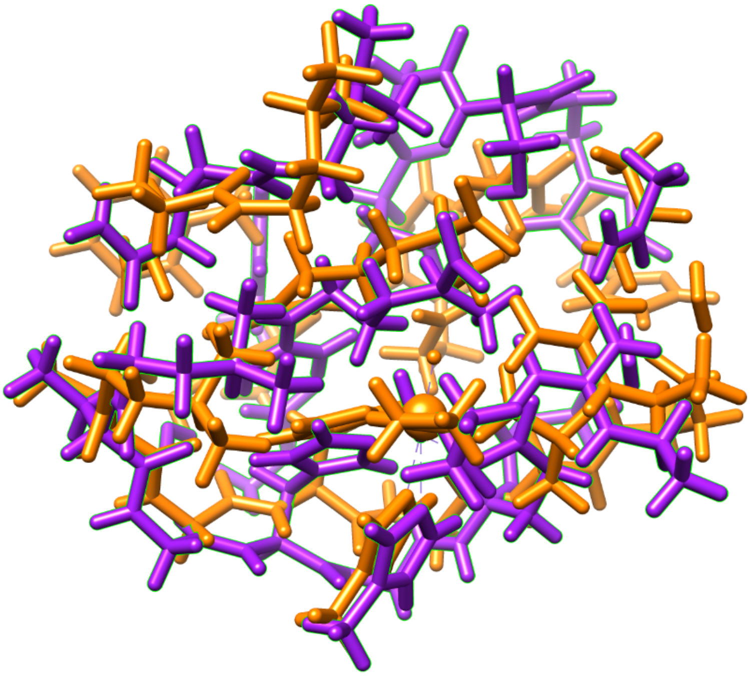


Figure 4.3. Overlay of the UB3LYP/BS1 optimized geometry of $^{5}\mathbf{Re}_{\mathrm{A}}$ (in orange) with the crystal structure coordinates (in purple).

The nearest hydrogen atom from the oxo group is the pro-*S* hydrogen atom at the C^3-position and its distance is about 2.667 Å in the quintet spin state, while the pro-*R* hydrogen atom is located at 4.061 Å and the pro-*S* hydrogen atom on the C^4-position at 5.008 Å. Based on the substrate positioning in OrfP; therefore, it appears that the C^3-position is located closest to the metal center and hence should be the preferred site of activation. The quintet spin state is the ground state and is well separated from the triplet spin state by $\Delta E+ZPE = 7$ kcal mol^{-1}. As such, the triplet spin state plays little role in the reaction mechanism and the reaction takes place through single-state-reactivity of the quintet spin state surface only. This is as expected for an iron(IV)-oxo species in trigonal bipyramidal configuration that usually gives a high-spin ground state [80,81]. The quintet spin state is characterized with a molecular orbital occupation of $\pi^*_{xy}{}^1$ $\pi^*_{xz}{}^1$ $\pi^*_{yz}{}^1$ $\sigma^*_{x2\text{-}y2}{}^1$, while in the triplet spin state it is is $\pi^*_{xy}{}^2$ $\pi^*_{xz}{}^1$ $\pi^*_{yz}{}^1$. The spin-state-ordering, spin-state energies and orbital configurations match previous work on nonheme iron enzymes and biomimetic models well and show that OrfP has the usual features of the first-coordination sphere [70-79,82-89].

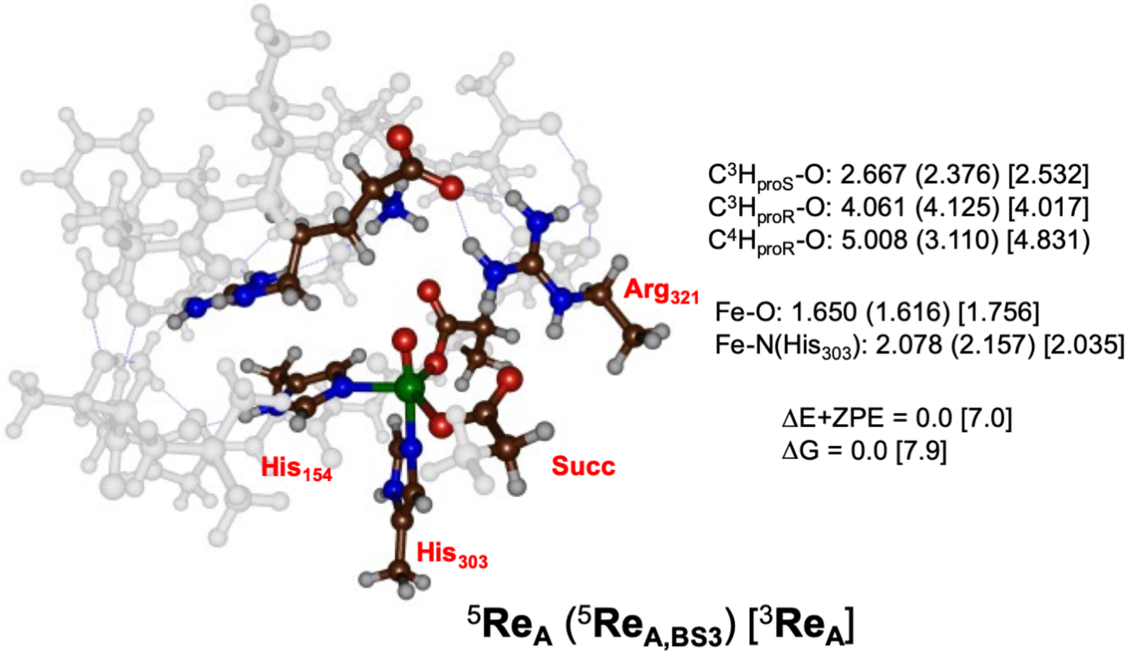
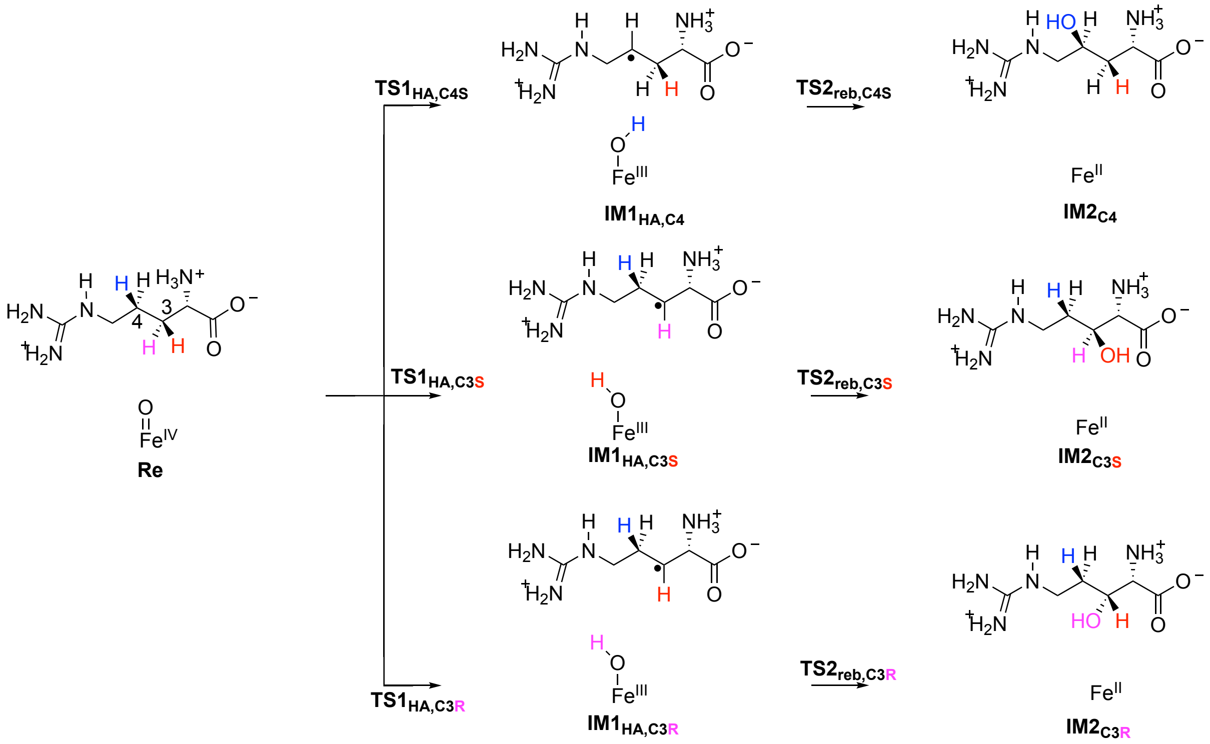

C^3H_{proS}-O: 2.667 (2.376) [2.532]
C^3H_{proR}-O: 4.061 (4.125) [4.017]
C^4H_{proR}-O: 5.008 (3.110) [4.831]
Arg_{321}
Fe-O: 1.650 (1.616) [1.756]
Fe-N(His_{303}): 2.078 (2.157) [2.035]
ΔE+ZPE = 0.0 [7.0]
ΔG = 0.0 [7.9]
His_{154}
Succ
His_{303}
$^5\mathbf{Re}_A$ ($^5\mathbf{Re}_{A,BS3}$) [$^3\mathbf{Re}_A$]

Figure 4.4. UB3LYP/BS1 optimized geometries of $^{5,3}\mathbf{Re}_A$ with bond lengths in angstroms and the relative (free) energies in kcal mol^{-1}. UB3LYP/BS3 optimized geometry $^5\mathbf{Re}_{A,BS3}$ data are given in parenthesis.

Next we considered substrate hydroxylation by the reactant complexes $^{3,5}\mathbf{Re}_A$ at the C^3 and C^4 positions to produce 3*S*-hydroxyarginine, 3*R*-hydroxyarginine and 4*S*-hydroxyarginine. The overall reaction scheme that was investigated with the definition of the various intermediate and transition state structures is given in Scheme 4.2. To this end we calculated the hydrogen atom abstraction from the two different hydrogen atoms on the C^3-position, designated the C3R (or pro-*R*) and C3S (or pro-*S*) hydrogen atoms, and the nearest hydrogen from the C^4 group, i.e. the C4S hydrogen atom. Transition states ($\mathbf{TS1}_{HA}$) for all positions were located and lead to a radical intermediate (**IM1**) representing an iron(III)-hydroxo species with a nearby substrate radical on either the C^3 or C^4 position of the substrate. A subsequent OH rebound to the substrate (via transition state $\mathbf{TS2}_{reb}$) leads to the mono-hydroxylated products (**IM2**) that are bound to an iron(II) species. The C–H bond that is activated in each step in the first substrate hydroxylation mechanism is identified with C3R, C3S and C4S in subscript after the label of the structure for the pathways leading to 3*R*-hydroxyarginine, 3*S*-hydroxyarginine and 4*S*-hydroxyarginine.



Scheme 4.2. Reaction mechanisms of arginine hydroxylation by the iron(IV)-oxo species of OrfP as studied in this work.

The reaction mechanism as described in Scheme 4.2 was calculated for the large cluster model **A** of OrfP for substrate hydroxylation leading to 3*S*-hydroxyarginine, 3*R*-hydroxyarginine and 4*S*-hydroxyarginine products. The potential energy landscape with key optimized geometries is shown in Figure 4.5. As can be seen from Figure 4.5 the lowest enthalpy of activation ($\Delta E^{\ddagger}$+ZPE) for hydrogen atom abstraction is from the pro-*S* C^3–H position via ^{5}**TS1**$_{\rm HA,C3S}$ with a value of $\Delta E^{\ddagger}$+ZPE = 12.1 kcal mol^{-1}. Close in enthalpy of activation is the pro-*R* C^3–H hydrogen atom abstraction barrier via ^{5}**TS1**$_{\rm HA,C3R}$ at $\Delta E^{\ddagger}$+ZPE = 15.5 kcal mol^{-1}. The hydrogen atom abstraction barrier from the C^4–H position is much higher in energy than the one from the C^3–H positions, i.e. $\Delta E^{\ddagger}$+ZPE = 23.2 kcal mol^{-1}. Based on these hydrogen atom abstraction barriers, a selective hydrogen atom abstraction from the C^3–H position is predicted with the pro-*S* channel dominant, although small amounts of pro-*R* cannot be excluded. These calculations therefore match experimental observation [30] that singly hydroxylated arginine at the C^3-position is formed. We also calculated the ^{5}**TS1**$_{\rm HA,C3S}$ and

^{5}**TS1**$_{\mathrm{HA,C4S}}$ structures at UB3LYP/BS3 level of theory and find ΔE‡+ZPE energies at UB3LYP-D3/BS2//UB3LYP/BS3 of 7.9 and 23.4 kcal mol^{-1}, respectively. As such these barriers predict the same trends as seen with UB3LYP/BS1 optimized structures and hence we continued with this method only.

Interestingly, when entropy and thermal corrections are added to the enthalpy, the pro-*R* hydrogen atom abstraction barrier becomes the lowest energy pathway: ΔG‡ = 11.6 kcal mol^{-1} for the pro-*R* C^3–H pathway, while for the pro-*S* C^3–H pathway ΔG‡ = 13.5 kcal mol^{-1} is found. Consequently, both hydrogen atoms on the C^3 atom of L-arginine can be abstracted by the iron(IV)-oxo species and this should give a mixture of 3*R*-hydroxyarginine and 3*S*-hydroxyarginine products. These products have indeed been observed experimentally [30]. To understand why entropy and thermal corrections reverse the ordering of ^{5}**TS1**$_{\mathrm{HA,C3S}}$ and ^{5}**TS1**$_{\mathrm{HA,C3R}}$, we analyzed the structures in more detail. The transition state geometries (^{5}**TS1**$_{\mathrm{HA,C3S}}$, ^{5}**TS1**$_{\mathrm{HA,C3R}}$ and ^{5}**TS1**$_{\mathrm{HA,C4S}}$) are shown in Figure 4.5. The two C^3–H hydrogen atom abstraction barriers are relatively central with similar C^3–H and O–H distances. In particular, the ^{5}**TS1**$_{\mathrm{HA,C3R}}$ has C^3–H and O–H distances of 1.265 and 1.272 Å, respectively, while those distances are 1.302 and 1.233 Å for ^{5}**TS1**$_{\mathrm{HA,C3S}}$. By contrast, the C^4–H hydrogen atom abstraction barrier is more product-like with considerably longer C^4–H distance than O–H distance: 1.358 versus 1.187 Å. The interesting difference between the three structures relates to the Fe–O–C angle in the transition states. Thus, ^{5}**TS1**$_{\mathrm{HA,C3S}}$ has a large Fe–O–C^3 angle of 167°, whereas in the ^{5}**TS1**$_{\mathrm{HA,C3R}}$ structure it is 139° and in ^{5}**TS1**$_{\mathrm{HA,C4S}}$ the angle is 127°.

We also optimized the ^{5}**TS1**$_{\mathrm{HA,C3S}}$ and ^{5}**TS1**$_{\mathrm{HA,C4S}}$ structures with UB3LYP/BS3; however, very little changes in the optimized geometries with respect to UB3LYP/BS1 are seen. For both structures the C–H and O–H distances are within 0.01Å and only the Fe–O is shortened by up to 0.04Å. As such the basis set has little influence on the optimized structures and BS1 and BS3 give qualitative similar results. Typically in hydrogen atom abstraction transition states by nonheme iron(IV)-oxo species the quintet spin state pathways shows approach of the substrate from the top and an almost linear Fe–O–C bond angle [90,91]. Despite this large difference in oxidant approach between ^{5}**TS1**$_{\mathrm{HA,C3S}}$ and ^{5}**TS1**$_{\mathrm{HA,C3R}}$ based on the Fe–O–C^3 angle; actually an overlay of the two optimized geometries does not show major differences. However, there are major differences in vibrational entropy between the two structures, whereby ^{5}**TS1**$_{\mathrm{HA,C3S}}$ has a 12.1 cal mol^{-1} K^{-1} larger vibrational entropy than ^{5}**TS1**$_{\mathrm{HA,C3R}}$. Moreover, when all small vibrations with magnitude smaller than 50 cm^{-1} are removed from the equation, the entropy difference is reversed and ^{5}**TS1**$_{\mathrm{HA,C3R}}$ has a larger vibrational entropy by 16.0 cal mol^{-1} K^{-1}. This corresponds to a free energy stabilization of 4.8 kcal mol^{-1} and reverses the individual barrier heights. The large vibrational entropy contribution of ^{5}**TS1**$_{\mathrm{HA,C3R}}$ is unrealistic and probably the result of a gas-phase model, where the vibrational contributions are overestimated. Thus, previously we did a comparative study on experimental and computational enthalpy and free energies of activation for oxygen atom transfer reactions and found the entropy to be overestimated by as much as 50% in model [92]. Moreover, the enthalpy values predict the experimental product distributions correctly and appear to be more realistic. As such, we will focus on ΔE+ZPE values only.

To test whether replacement of the transferring hydrogen atom by deuterium would have an effect on the barrier heights, we re-evaluated the vibrational frequencies by replacing the pro-*R* and pro-*S* hydrogen atom on the C^3-position by deuterium. Thus, with a deuterium atom on the pro-*R* position and a hydrogen atom at the pro-*S* position the ^{5}**TS1**$_{\mathrm{HA,C3R}}$ and ^{5}**TS1**$_{\mathrm{HA,C3S}}$ barriers change to $\Delta G^‡$ = 12.8 and 13.5 kcal mol^{-1}, respectively. Consequently, the order of the hydrogen atom abstraction barriers does not change, but the energy gap narrows to within 1 kcal mol^{-1}. The reverse situation with a deuterium atom on the pro-*S* position and a hydrogen atom at the pro-*R* position gives free energies of activation of $\Delta G^‡$ = 11.7 and 14.7 kcal mol^{-1} and widens the energy gap. As the hydrogen atom abstraction step is rate-determining for the first hydroxylation cycle of L-Arg by OrfP, we calculated the kinetic isotope effects (KIEs) using the Eyring and Wigner methods. Both C3R and C3S pathways give a KIE$_{\mathrm{Eyring}}$ of about 8, which rises to 14 – 15 when quantum mechanical tunneling corrections are added. These are typical values of hydrogen atom abstraction kinetic isotope effects that are commonly seen in hydrogen atom abstraction reactions by nonheme iron(IV)-oxo complexes in enzymatic and biomimetic model complexes [93-95].

After the hydrogen atom abstraction transition states all structures relax to a radical intermediate **IM1**; however, the two structures with a radical on C^3 (^{5}**IM1**$_{\mathrm{C3S}}$ and ^{5}**IM1**$_{\mathrm{C3R}}$) are very wide apart with ^{5}**IM1**$_{\mathrm{C3S}}$ lower in energy than ^{5}**IM1**$_{\mathrm{C3R}}$ by 7.9 kcal mol^{-1}. By contrast, the ^{5}**IM1**$_{\mathrm{C4S}}$ is within 1.1 kcal mol^{-1} of ^{5}**IM1**$_{\mathrm{C3S}}$. Therefore, the ordering of the radical intermediates ^{5}**IM1** are different from those of the transition states and probably is the result of the tight substrate binding and positioning that affect the kinetics dramatically.

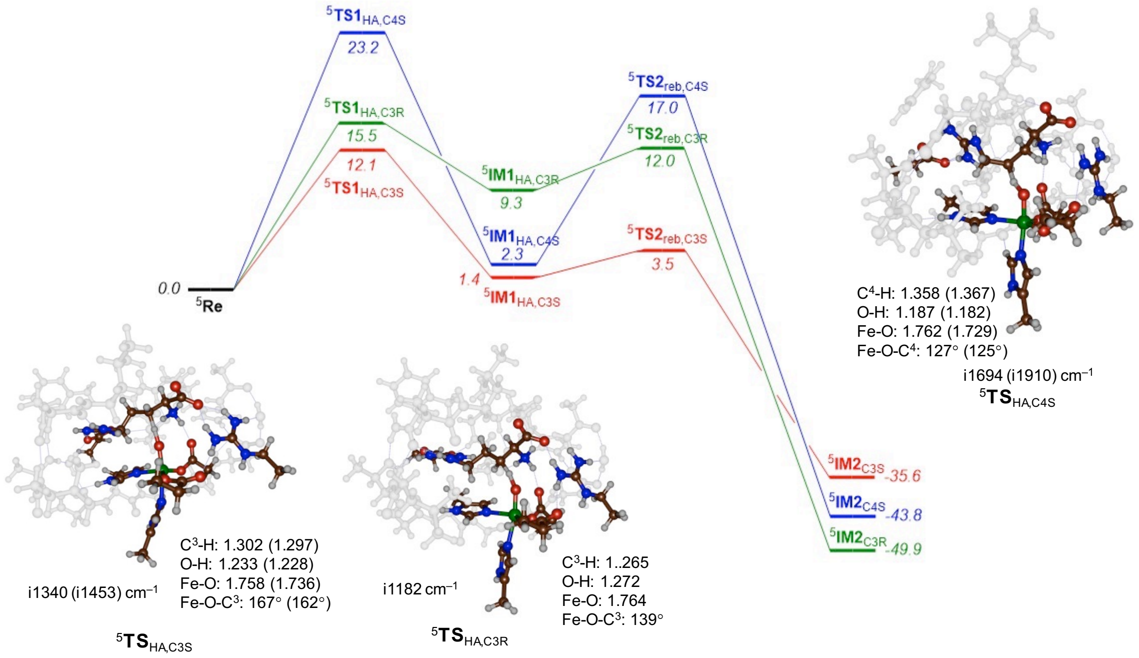
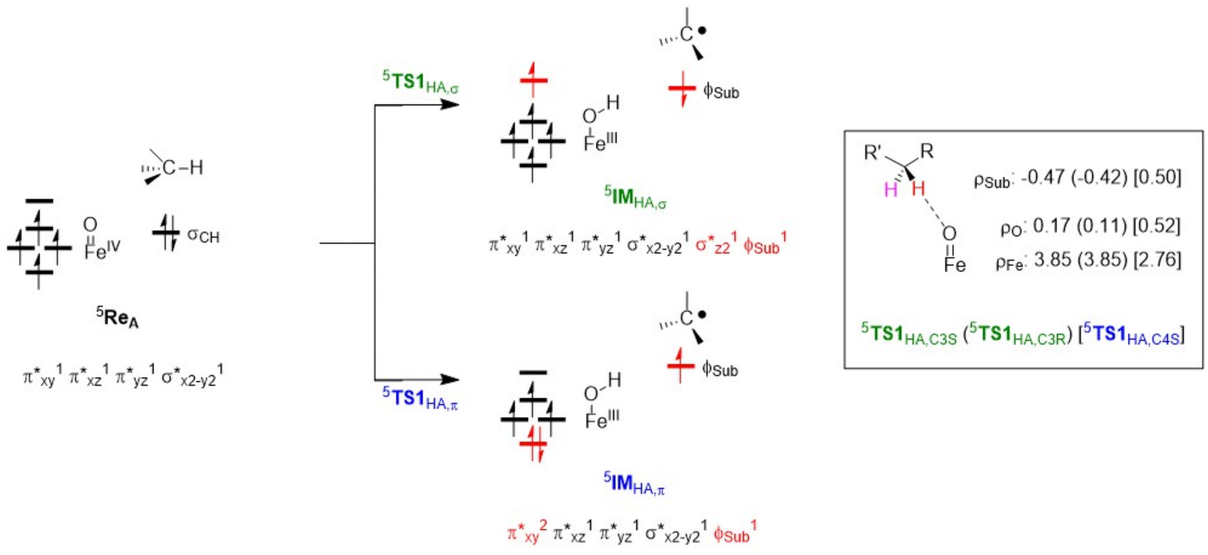


Figure 4.5. Potential energy landscape for L-arginine hydroxylation at the C^3 and C^4 positions with energies (in kcal mol^{-1}) relative to ^{5}**Re**$_A$ calculated at UB3LYP-D3/BS2//UB3LYP/BS1. Optimized geometries of the transition states are given with bond lengths in angstroms, the bond angle in degrees and the imaginary frequency in cm^{-1}. UB3LYP/BS3 optimized structures are given in parenthesis.

Next, an OH rebound step takes place from the **IM1** intermediates. The pathway from ^{5}**IM1**$_{C3}$ gives small rebound transition states of less than 3 kcal mol^{-1} to give alcohol products with large exothermicity. This small rebound barrier will imply that ^{5}**IM1**$_{C3}$ has a short lifetime and collapse to products without rearrangement and stereochemical scrambling of products. On the other hand, the rebound barrier ^{5}**TS2**$_{reb,C4S}$ is 14.7 kcal mol^{-1} above ^{5}**IM1**$_{C4S}$ and hence the radical intermediate ^{5}**IM1**$_{C4S}$ will have a finite lifetime. Interestingly, the alcohol product complexes (**IM2**) have different ordering than the **IM1** states, whereby the most stable structure is ^{5}**IM2**$_{C3R}$. Overall, the DFT modelling on cycle 1 of OrfP shows that a mixture of 3*S*-hydroxyarginine and 3*R*-hydroxyarginine may be expected as those pathways have competing reaction barriers and rate constants.

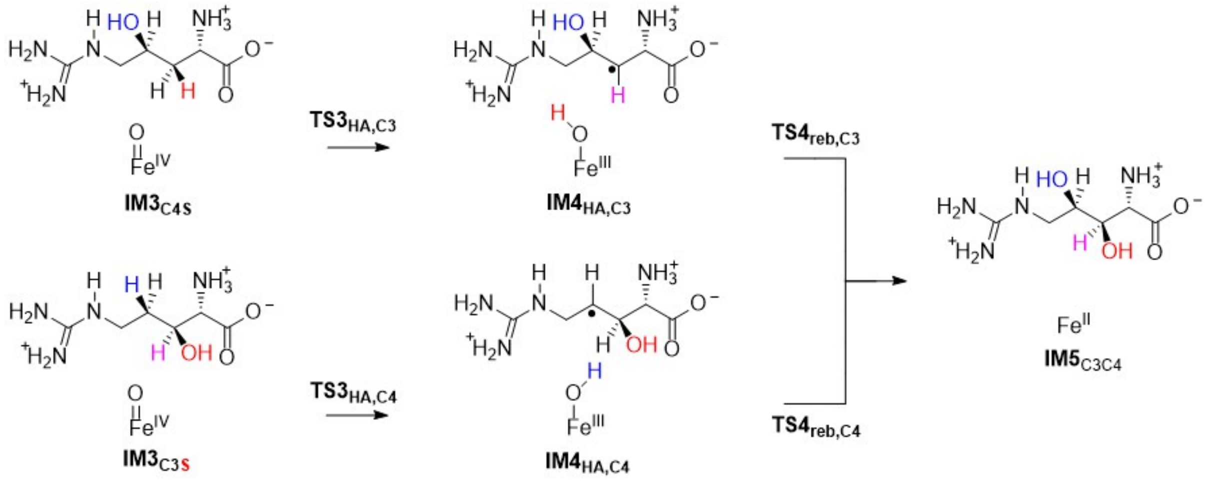


C–H
σCH
O
FeIV
5ReA
π*xy1 π*xz1 π*yz1 σ*x2-y21
5TS1HA,σ
5TS1HA,π
C•
ϕSub
O–H
FeIII
5IMHA,σ
π*xy1 π*xz1 π*yz1 σ*x2-y21 σ*z21 ϕSub1
5IMHA,π
π*xy2 π*xz1 π*yz1 σ*x2-y21 ϕSub1
R'
R
H
H
O
Fe
ρSub: -0.47 (-0.42) [0.50]
ρO: 0.17 (0.11) [0.52]
ρFe: 3.85 (3.85) [2.76]
5TS1HA,C3S (5TS1HA,C3R) [5TS1HA,C4S]

Figure 4.6. Electron transfer processes for hydrogen atom abstraction and OH rebound via $^{5}\sigma$ (top) and $^{5}\pi$ (bottom) pathways and group spin densities (ρ) as obtained from the optimized transition state geometries.

To understand the different substrate to oxidant angles (Fe–O–C angle), we show the possible electron transfer pathways in the hydrogen atom abstraction and OH rebound steps in Figure 4.6. As discussed above the reactant has $\pi^*_{xy}{}^{1}$ $\pi^*_{xz}{}^{1}$ $\pi^*_{yz}{}^{1}$ $\sigma^*_{x2\text{-}y2}{}^{1}$ configuration, while the substrate C–H bond is occupied with two electrons. Upon hydrogen atom transfer, the C–H bond cleaves homolytically and a radical is left on the substrate in orbital ϕ_{Sub}. The electron of the hydrogen atom moves into the metal 3d-block and generally there are two possibilities called the $^{5}\sigma$ and $^{5}\pi$ pathways [96-100]. In the $^{5}\sigma$ pathway an electron transfer from the substrate into the virtual π^*_{z2} orbital takes place to give a radical intermediate ^{5}**IM1**$_{HA,\sigma}$ with configuration $\pi^*_{xy}{}^{1}$ $\pi^*_{xz}{}^{1}$ $\pi^*_{yz}{}^{1}$ $\sigma^*_{x2\text{-}y2}{}^{1}$ $\sigma^*_{z2}{}^{1}$ $\phi_{Sub}{}^{1}$. As the σ^*_{z2} orbital is located along the Fe–O axis the substrate will approach from the top and an almost linear Fe–O–C angle is obtained in the transition state. In the radical intermediate the five metal-type orbitals are antiferromagnetically coupled to a radical on the substrate and are exchange stabilized. The alternative mechanism for substrate hydroxylation is via the $^{5}\pi$-pathway, where the initial electron transfer from the substrate into the singly occupied π^*_{xy} orbital takes place. This then gives an intermediate with electronic configuration $\pi^*_{xy}{}^{2}$ $\pi^*_{xz}{}^{1}$ $\pi^*_{yz}{}^{1}$ $\sigma^*_{x2\text{-}y2}{}^{1}$ $\phi_{Sub}{}^{1}$, where now all unpaired electrons are ferromagnetically coupled. The radical rebound pathways brings the $^{5}\sigma$ and $^{5}\pi$ pathways together into the same alcohol product complex ^{5}**IM2** with configuration $\pi^*_{xy}{}^{2}$ $\pi^*_{xz}{}^{1}$ $\pi^*_{yz}{}^{1}$ $\sigma^*_{x2\text{-}y2}{}^{1}$ $\sigma^*_{z2}{}^{1}$. Electron transfer into the π^*_{xy} orbital gives a side-on approach and a more bent structure (typically around 120°) is found. Based on the electron-transfer processes these pathways are called the $^{5}\sigma$ and $^{5}\pi$-pathways. To find out, whether our hydrogen atom abstraction pathways belong to the $^{5}\sigma$ or $^{5}\pi$-type, we analyzed the molecular orbitals and unpaired spin density of the three-hydrogen atom abstraction transition states $^{5}\mathbf{TS1}_{\mathrm{HA}}$.

Group spin densities of $^{5}\mathbf{TS1}_{\mathrm{HA,C3S}}$, $^{5}\mathbf{TS1}_{\mathrm{HA,C3R}}$ and $^{5}\mathbf{TS1}_{\mathrm{HA,C4S}}$ transition states are given in Figure 4.6. The spin density for the C^{3}–H hydrogen atom abstraction is large on the iron (ρ_{Fe} = 3.85 for both), while negative spin density is accumulating on the substrate. Consequently, these spin densities characterize both C^{3}–H hydrogen atom abstraction barriers as $^{5}\sigma$ pathway structures. By contrast, the $^{5}\mathbf{TS1}_{\mathrm{HA,C4S}}$ transition state has positive spin density on the substrate (ρ_{Sub} = 0.50), while the spin on iron is only ρ_{Fe} = 2.76. Therefore, the $^{5}\mathbf{TS1}_{\mathrm{HA,C4S}}$ transition state is of the $^{5}\pi$-pathway rather than of $^{5}\sigma$. In an attempt to find the $^{5}\sigma$-pathway structures, some molecular orbitals of the $^{5}\pi$-pathway transition states were swapped; however, during the SCF convergence this electronic configuration was not stable and relaxed to the $^{5}\sigma$ electronic configuration and geometry instead. Therefore, the C^{3}–H and C^{4}–H hydrogen atom abstraction pathways give differences in electronic configuration and electron transfer processes that is forced upon the system through substrate binding and positioning. These electronic and stereochemical effects push the reaction to C^{3}-hydroxylation of L-Arg selectively.



Scheme 4.3. Reaction mechanisms investigated in this work for hydroxyarginine hydroxylation by an iron(IV)-oxo species of OrfP.

4.3.2 Second oxygenation cycle of L-Arg by OrfP.

In the next set of calculations, we explored the second hydroxylation cycle of arginine. Experimental studies show that OrfP enzymes are able to incorporate two hydroxyl groups into an arginine molecule [6,30,32]. We hypothesized that after the first hydroxylation step of the substrate is completed, the hydroxyarginine is not released from the enzyme but only succinate. Subsequently, another catalytic cycle starts with binding of a new molecule of αKG and O_2 to the iron center that then react to form a second iron(IV)-oxo and succinate species. To this end, we took the optimized geometry of the 3*S*-hydroxyarginine, 3*R*-hydroxyarginine and 4*S*-hydroxyarginine bound iron(II) complexes, i.e. ^{5}**IM2**$_{\mathrm{C3S}}$, ^{5}**IM2**$_{\mathrm{C3R}}$ and ^{5}**IM2**$_{\mathrm{C4S}}$, and replaced the iron(II) group by iron(IV)-oxo to form ^{5}**IM3**$_{\mathrm{C3S}}$, ^{5}**IM3**$_{\mathrm{C3R}}$ and ^{5}**IM3**$_{\mathrm{C4S}}$, which kept the overall charge the same. Thereafter, the hydroxylation of the pro-*R* C^4–H group of 3*S*-hydroxyarginine in ^{5}**IM3**$_{\mathrm{C3S}}$ and the hydroxylation of the pro-*R* C^3–H group in 4*S*-hydroxyarginine was studied to form the 3*R*,4*R*-dihydroxyarginine product **IM5**$_{\mathrm{C3C4}}$ (Scheme 4.3). Note that the dihydroxylation changes the stereochemistry on atom C^3 from 3*S* in the mono-hydroxylated species to 3*R* in the dihydroxylated arginine. The second hydroxylation step starts with a hydrogen atom abstraction via ^{5}**TS3**$_{\mathrm{HA}}$ to form a radical intermediate ^{5}**IM4**$_{\mathrm{HA}}$. A radical rebound via transition state ^{5}**TS4**$_{\mathrm{reb}}$ gives the dihydroxylated arginine product complex ^{5}**IM5**$_{\mathrm{C3C4}}$. Figure 4.7 shows the potential energy landscape for the second hydroxylation cycle starting from these ^{5}**IM3**$_{\mathrm{B}}$. In addition to these two pathways we also investigated the hydrogen atom abstraction from the C^3–H group in 3*R*-hydroxyarginine.

The iron(IV)-oxo species for the second reaction cycle for 3*S*-hydroxyarginine and 3*R*-hydroxyarginine bound complexes (^{5}**IM3**$_{\mathrm{B,C3S}}$ and ^{5}**IM3**$_{\mathrm{B,C3R}}$) are shown in Figure 4.7. They have similar features as the reactant species of cycle 1 with an Fe–O distance of 1.656 and 1.650 Å, respectively, and the same electronic configuration of $\pi^*_{xy}{}^1\ \pi^*_{xz}{}^1\ \pi^*_{yz}{}^1\ \sigma^*_{x2-y2}{}^1$. The only difference is a hydrogen bonding interaction between the alcohol group of 3*S*-hydroxyarginine and the oxo group, which may have elongated the Fe–O bond slightly. Nevertheless, the C^4–H group is positioned close to the oxo group at a distance of 3.159 Å ^{5}**IM3**$_{\mathrm{B,C3S}}$. In contrast, in ^{5}**IM3**$_{\mathrm{B,C3R}}$ the C^3–H group is much further away and positioned at a distance of 5.019 Å. Therefore, we decided to explore hydrogen atom abstraction from the C^3–H position instead: the C^3H–O distance is 3.985 Å in ^{5}**IM3**$_{\mathrm{B,C3R}}$. Thereafter, the hydrogen atom abstraction by the iron(IV)-oxo species ^{5}**IM3**$_{\mathrm{B,C3S}}$ from the C^4–H position was calculated and a barrier of $\Delta E^{\ddagger}$+ZPE = 11.6 kcal mol^{-1} is found. This is much lower in energy than the hydrogen atom abstraction from the C^4–H position obtained for L-arginine, i.e. 23.2 kcal mol^{-1} see Figure 4.5. An analysis of the group spin densities of ^{5}**TS3**$_{HA,C4}$ gives a spin of 3.93 on iron and a spin of −0.42 on the C^4 atom of the substrate and an orbital occupation of $\pi^*_{xy}{}^1$ $\pi^*_{xz}{}^1$ $\pi^*_{yz}{}^1$ $\sigma^*_{x2-y2}{}^1$ $\sigma^*_{z2}{}^1$ $\phi_{Sub}{}^1$.

Consequently, ^{5}**TS3**$_{HA,C4}$ has a $^5\sigma$ electronic configuration, while ^{5}**TS1**$_{HA,C4}$ had a $^5\pi$ configuration. These differences are probably the result of the hydrogen bonding interaction from the hydroxo group of substrate to the oxo group that constraints the substrate approach and withdraws electron density. Indeed in previous work we showed that hydrogen bonding interactions to an iron(IV)-oxo species influence reaction barriers [90,91]. Geometrically, the transition state ^{5}**TS3**$_{HA,C4}$ is relatively central with almost equal C–H and O–H distances, i.e. 1.245 Å (C^4–H bond) and 1.296 Å (O–H bond). The C^4–O–Fe angle, however, is somewhat bent (140°) while for the first hydrogen atom abstraction transition states (**TS1**$_{HA}$ above) have an angle close to 180°. Consequently, there is considerable constraint in the structure probably due to hydrogen bonding interactions in the substrate binding pocket.

After the transition state, the system relaxes to a radical intermediate via an almost thermoneutral process (ΔE+ZPE = −1.2 kcal mol^{-1} with respect to ^{5}**IM3**$_{B,C3S}$). A subsequent rebound barrier of 6.1 kcal mol^{-1} above ^{5}**IM4**$_{HA,C4}$ leads to the dihydroxylated arginine products ^{5}**IM5**$_{C3C4}$ with large exothermicity. The alternative pathway starting from ^{5}**IM3**$_{B,C4S}$ to form ^{5}**IM5**$_{C3C4}$ was also studied. Thus, a transition state of $\Delta E^{\ddagger}$+ZPE = 12.0 kcal mol^{-1} leads to a radical intermediate that is 2.6 kcal mol^{-1} more stable than ^{5}**IM3**$_{B,C3S}$. A small OH rebound barrier of 4.3 kcal mol^{-1} above ^{5}**IM4**$_{HA,C3}$ then leads to the 3*R*,4*R*-dihydroxyarginine product. The transition state geometry of ^{5}**TS3**$_{HA,C3}$ is given in Figure 4.6 and has the transferring the hydrogen atom almost midway between the donor and acceptor groups: the C^3–H distance is 1.311 Å and the H–O distance is 1.220 Å. The C^3–O–Fe angle is 166° and therefore much larger than the corresponding angle in ^{5}**TS3**$_{HA,C4}$ of 140°.

These reaction barriers are low in energy and of the same order of magnitude as the first hydrogen atom abstraction from the pro-*S* position of the C^3–H group in Figure 4.5 above. Therefore, the dihydroxylated product can be formed from either 3*S*-hydroxyarginine or 4*S*-hydroxyarginine with similar rate constants and reaction barriers. However, as shown in Figure 4.5 above, the first hydroxylation step is highly selective and will give dominant 3*S*-hydroxyarginine through the first oxygen atom transfer cycle.

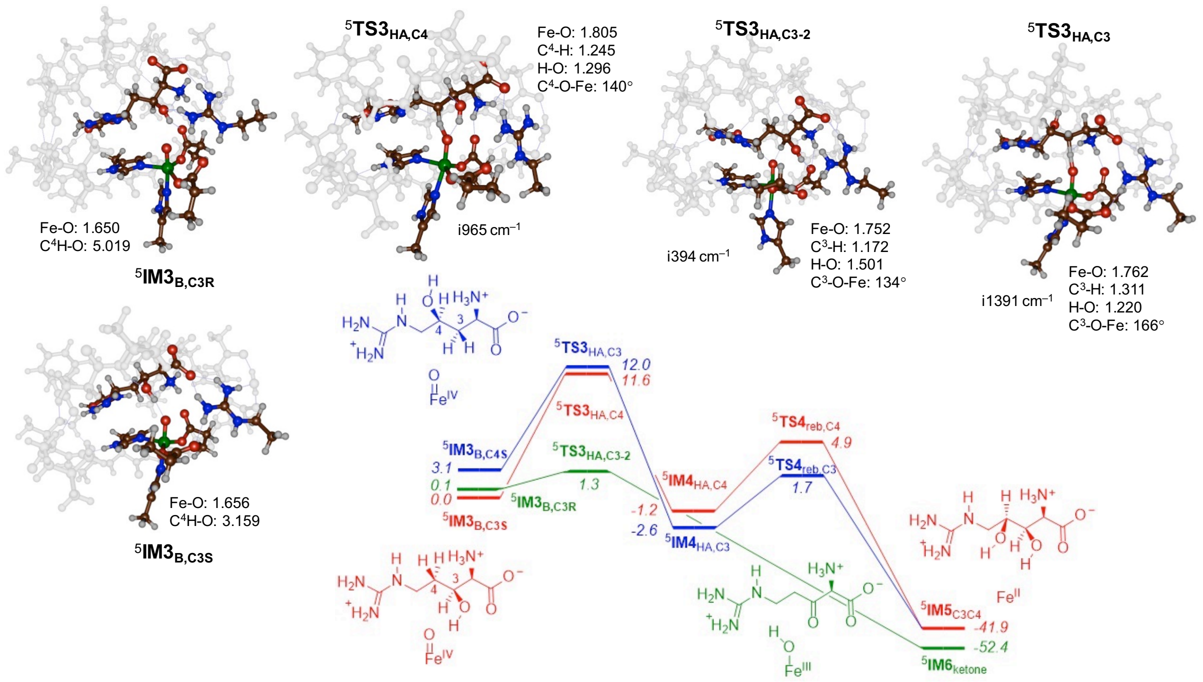


Figure 4.7. Potential energy landscape for 3-hydroxyarginine and 4-hydroxyarginine hydroxylation to form the dihydroxoarginine and β-ketoarginine products. Energies calculated at UB3LYP-D3/BS2//UB3LYP/BS1. Outside parenthesis are ΔE+ZPE values, while free energies are in parenthesis. Optimized geometries of the transition states are given with bond lengths in angstroms, angles in degrees and the imaginary frequency in cm^{-1}.

Finally, we explored activation of 3*R*-hydroxyarginine in a second reaction cycle. As the C^4–H group is located far away, due to the positioning of 3*R*-hydroxyarginine in the substrate binding pocket we instead studied the C^3–H abstraction pathway via transition state ^{5}**TS3**$_{\mathrm{HA,C3\text{-}2}}$. A barrier of only 1.3 kcal mol^{-1} with respect to ^{5}**IM3**$_{\mathrm{B,C3S}}$ is found, which is not surprising as a weak tertiary C–H bond is broken. This transition state is early with a short C^3–H distance of 1.172 Å and a long O–H distance of 1.501 Å. The barrier has a small imaginary frequency of i394 cm^{-1} with a dominant C^3–H–O stretch vibration. However, the transition states also show movement for the OH group of the substrate in the direction of the carboxylate group of Glu$_{156}$. Indeed, after the hydrogen atom abstraction barrier the system does not relax to a radical intermediate but a rapid second hydrogen atom transfer takes place from the substrate OH group to the Glu$_{156}$ group, which desaturates the C^3–O bond of substrate and form the β-keto-arginine product (^{5}**IM6**$_{\mathrm{ketone}}$). This product cannot be formed from 3*S*-hydroxyarginine and 4*S*-hydroxyarginine as the tertiary C–H bond points away from the iron-oxo group and hence these groups are not accessible for the iron(IV)-oxo species. It is evident from our calculations, therefore, that 3*R*-hydroxyarginine most likely leads to desaturation of the substrate through a very efficient and low-energy reaction process in the first reaction cycle and consequently, substrate binding and positioning should block this potential pathway. As such, substrate positioning is crucial in OrfP enzymes and the substrate binding pocket is evolved to maximize the yield of products and minimize the amount of by-products. In particular, the tertiary C–H bonds of the mono-hydroxylated arginine should point away from the metal center so that no desaturation pathways become accessible.

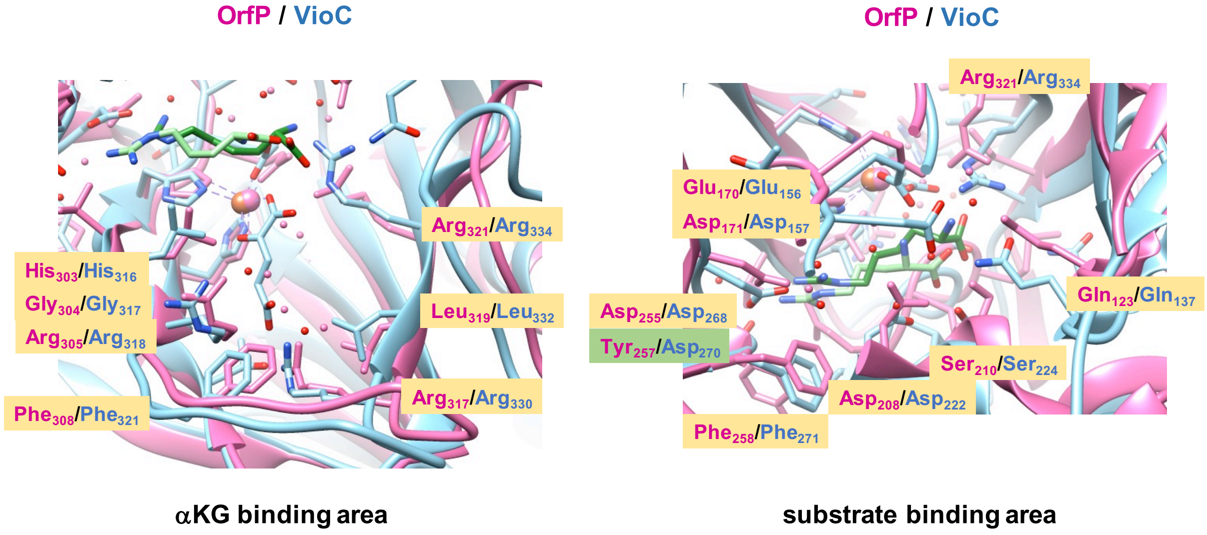


Figure 4.8. Overlay of the pdb files of OrfP [30] and VioC [23] with the focus on the αKG binding pocket (left) and L-Arg binding pocket (right).

4.4 Discussion

To understand the details of the dihydroxylation mechanism of L-Arg by OrfP enzymes, we analyzed the thermochemical properties of the oxidant and substrate in more detail and made a comparison with analogous enzymes. We compared the structures of two arginine activating nonheme iron dioxygenases, namely OrfP and VioC. These two nonheme iron dioxygenases react differently with L-arginine as a substrate, whereby VioC selectively hydroxylates it at the C^3-position and OrfP performs the dihydroxylation to form 3*R*,4*R*-dihydroxyarginine. A structural comparison between the OrfP and VioC crystal structure coordinates (4M2E versus 6ALM pdb files) [23,30] is given in Figure 4.8 as an overlay of the active site regions of the two enzymes. Both enzymes utilize αKG, dioxygen and L-Arg on a nonheme iron center that is bound to the protein through a 2-His/1-Glu facial triad coordination. The overlay on the left focuses on the αKG binding area, while the one on the right zooms into the L-Arg binding area. As can be seen from Figure 4.8, the two enzymes have an almost identical αKG binding loop that starts from the axial histidine ligand (His_{303} in OrfP and His_{316} in VioC) and the peptide chain continues with HGRXXFQXRYDGXDRWLKR. Hence in this loop of 19 amino acids only four residues (labeled as X) do not match between the two proteins and all of those amino acids point away from the αKG binding pocket. Clearly, the αKG binding loop is highly conserved and may show a similar amino acid chain in most αKG dependent nonheme iron dioxygenases. Indeed, it has been reported that a HX13R loop from the axial ligand reaches a conserved Arg residue that binds and positions αKG in the substrate binding pocket [102]. The overlay on the left-hand-side of Figure 4.8 indeed puts the residues of this loop in virtually the same position in both enzymes. In OrfP the Arg_{317} forms a salt bridge with αKG, while the analogous residue Arg_{330} has that role in VioC. Interestingly, the next Arg in this chain (Arg_{321} in OrfP/Arg_{334} in VioC) locks the carboxylate group of the substrate in a salt bridge.

Another conserved region between the OrfP and VioC structures relates to the His and Glu iron ligands (His_{168} and Glu_{170} for OrfP and His_{154} and Glu_{156} for VioC) with a peptide region WHTEDAFXPY, whereby the Asp residue (Asp_{171} for OrfP and Asp_{157} for VioC) forms a hydrogen bonding interaction with the NH_3^+ group of L-Arg. By contrast the guanidinium group of L-Arg forms a salt bridge with Asp_{255} (OrfP) and Asp_{268} (VioC). In addition, the substrate binding pocket is aligned with the conserved residues Asp_{208}/Asp_{222}, Ser_{210}/Ser_{224} and Phe_{258}/Phe_{271}, for OrfP and VioC, respectively. The only difference seen from the overlay of the structures appears to be the position of a substrate binding pocket Gln residue (Gln_{123} in OrfP and Gln_{137} in VioC) that has shifted inside in OrfP. In addition, the replacement of Tyr_{257} (OrfP) with Asp in VioC on the edge of the substrate binding pocket is observed. The latter is highlighted in green in Figure 4.8 and implicates that VioC has an additional hydrogen bonding interaction to the substrate guanidium group that is missing in OrfP. The overlay of the pdb structures of VioC and OrfP shown in Figure 4.8; therefore, implies that their active site features, and αKG and substrate binding environments are highly alike. As such, the structures do not give a clear reason why different reaction products are obtained in the two reaction mechanisms. The analysis and comparison of the crystal structure coordinates of OrfP and VioC; however, implicates that both enzymes will preferentially activate L-Arg on the C^3-

position as that group appears to be closest to the metal center. Indeed, our lowest energy reaction pathways for the first hydroxylation cycle in OrfP give the lowest barriers for hydrogen atom abstraction from the C^3–H position leading to 3*S*-hydroxyarginine products. Consequently, both VioC and OrfP are C^3-activating nonheme iron enzymes of L-Arg substrate due to careful substrate positioning in the substrate binding pocket. To gain further insight into the differences of the L-Arg activation mechanisms of VioC and OrfP, we created an active site cluster model of VioC with 3*S*-hydroxyarginine bound and optimized its geometry at UB3LYP/BS1: ^{5}**IM2**$_{\mathrm{C3S,VioC}}$. An overlay of the ^{5}**IM2**$_{\mathrm{C3S,VioC}}$ structure with the OrfP optimized geometry of 3*S*-hydroxyarginine, i.e. ^{5}**IM2**$_{\mathrm{C3S}}$, is shown in Figure 4.9. As can be seen from Figure 4.9, most of the protein residues are in a similar position between the two structures. In particular, the first coordination sphere ligands, i.e. the 2-His/1-Glu coordinated ligands and the succinate are in virtually the same position. Clearly, the differences in reactivity are not the result of differences in the coordination environment of the metal and hence come from the second coordination sphere and product release.

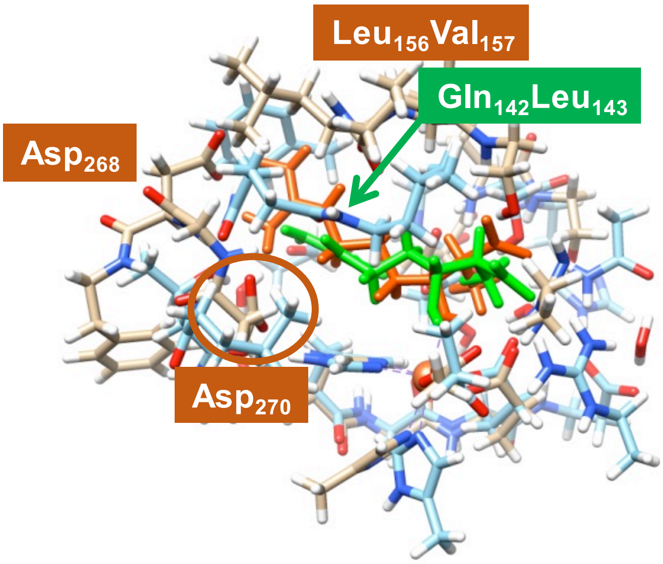


Figure 4.9. Overlay of UB3LYP/BS1 optimized geometries of 3S-hydroxyarginine bound iron(II) complexes of OrfP and VioC. Protein backbone of VioC in grey with 3S-hydroxyarginine in orange, while the protein backbone for OrfP is in light blue with the 3S-hydroxyarginine in green. Key differences between the two structures are highlighted in orange for VioC and green for OrfP.

As discussed in Figure 4.8 above, VioC has an extra carboxylic acid group in the substrate binding pocket, namely Asp_{270}, where OrfP has a Tyr residue. The optimized geometry of 3*S*-hydroxyarginine in the VioC model as a result is positioned differently and the guanidinium group is twisted with respect of the OrfP structure. This points the hydroxo group more towards the iron atom and may prevent further dioxygen binding to the iron center. In contrast, in OrfP the guanidinium group of 3*S*-hydroxyarginine only forms a salt bridge with Asp_{255}, which points the hydroxyarginine tail down, while the other terminus is pointed slightly up. This means the hydroxyl group is slightly further from the metal center and a gap has appeared where O_2 can be inserted to trigger a new catalytic cycle. The tight substrate positioning in VioC may prevent this. In addition to the extra carboxylic acid group in VioC, the row of amino acids Leu_{156}-Val_{157} is located slightly higher in the substrate binding pocket than the corresponding residues (Gln_{142}-Leu_{143}) in OrfP. Therefore, the binding pocket in OrfP is tighter and smaller in OrfP than it is in VioC and consequently product release will be slowed down. These seemingly small differences between the two enzymes determine the product release mechanism and enable a second oxidation cycle in OrfP, which does not happen in VioC.

Next, we did a thermochemical analysis of the substrate, the intermediates and products and focused on the C–H bond strengths of the various aliphatic positions of L-Arg and the hydroxyarginine isomers in the structure in the protein. To this end, we calculated the bond dissociation energy (BDE) of various C–H bonds of L-Arg by calculating an isolated L-Arg molecule, a H-atom and the substrate with one hydrogen atom removed, Equation 4.4. [101,102] The energy difference between these three structures is then the $\mathrm{BDE1}_{\mathrm{CH}}$ for that particular position. In addition, the C–H bond dissociation energies of several C–H bonds in the hydroxyarginine structures from the geometries of the **IM2** intermediates were calculated (Equation 4.5): $\mathrm{BDE2}_{\mathrm{CH}}$.

$$\mathrm{L-Arg}\ \rightarrow [\mathrm{L-Arg-H^{\blacksquare}}] + \mathrm{H^{\blacksquare}} + \mathrm{BDE1_{CH}} \tag{4.4}$$

$$\mathrm{ArgOH}\ \rightarrow [\mathrm{ArgOH-H^{\blacksquare}}] + \mathrm{H^{\blacksquare}} + \mathrm{BDE2_{CH}} \tag{4.5}$$

The $\mathrm{BDE1}_{\mathrm{CH}}$ and $\mathrm{BDE2}_{\mathrm{CH}}$ for L-Arg, 3*S*-hydroxyarginine, 4*S*-hydroxylarginine and 3*R*-hydroxyarginine were calculated for various C–H bonds at the UB3LYP/6-311++G** level of theory with solvent corrections included and are summarized in Figure 4.10. As can be seen the highlighted three C–H bond strengths of L-Arg are within 3.3 kcal mol^{-1} from each other, whereby the $BDE1_{C4S\text{-}H}$ and $BDE1_{C3R\text{-}H}$ are the lowest in energy at 94.0/94.1 kcal mol^{-1}, while the $BDE1_{C3S\text{-}H}$ is slightly higher in energy at 97.3 kcal mol^{-1}, respectively. Therefore, if the aliphatic C–H abstraction reaction from L-Arg by the enzyme is governed by the C–H bond strength of the substrate only, the reaction should proceed with dominant hydrogen atom abstraction from the pro-*R* C^3 and pro-*S* C^4 position of the substrate with minor amounts of pro-*S* C^3 hydroxylation. The DFT calculations shown above in Figure 4.5 on the enzymatic model, in contrast, show that the lowest barrier is obtained for C^3–H hydrogen atom abstraction from the pro-*S* C^3 position. Therefore, the 3*S*-hydroxyarginine is predicted to be the dominant product in the first reaction cycle based on the DFT cluster calculations even though the pro-*S* C^3–H bond is not the weakest C–H bond in the substrate. Thus, under ideal substrate approach, i.e. without substrate perturbation from the protein, the C^4-hydroxylation should be the dominant product. Clearly, substrate positioning and the tightness of the substrate binding pocket, guides the reaction to the pro-*S* C^3–H bond selectively. The C^4–H pathway, by contrast, appears to be higher in energy in the calculations on the enzymatic system as a result of access to a higher energy potential energy landscape with $^5\pi$ configuration. In addition, electrostatic interactions from the protein may destabilize its pathway. These wide differences in hydrogen atom abstraction barriers must result from the substrate binding and positioning that affects the accessibility of the substrate by the oxidant.

Technically, the first hydrogen atom abstraction step in Figure 4.5 should correlate with the energy to break the C–H bond in the substrate minus the energy to form the O–H bond in the iron(III)-hydroxo intermediate [101,102]. Previously, for a small cluster model complex we calculated a $BDE_{FeO–H}$ value of 93.0 kcal mol^{-1}[103]. Based on the difference between $BDE1_{CH}$ and $BDE_{FeO–H}$, we would predict a reaction enthalpy from reactants to $\mathbf{IM1}_{HA}$ of 4.3, 1.1 and 1.0 kcal mol^{-1} for the pro-*S* C^3, pro-*S* C^4 and pro-*R* C^3 pathways, respectively. These values are close to the $\mathbf{IM1}_{HA,C3S}$ and $\mathbf{IM1}_{HA,C4S}$ energies and shows that the optimized structures of the radical intermediates have limited disruption through the protein environment. However, the kinetics is strongly affected by the shape and size of the protein pocket and hence the transition states ($\mathbf{TS1}_{HA}$) follow a different ordering than the radical intermediates ($\mathbf{IM1}_{HA}$) and cover a wider energy range.

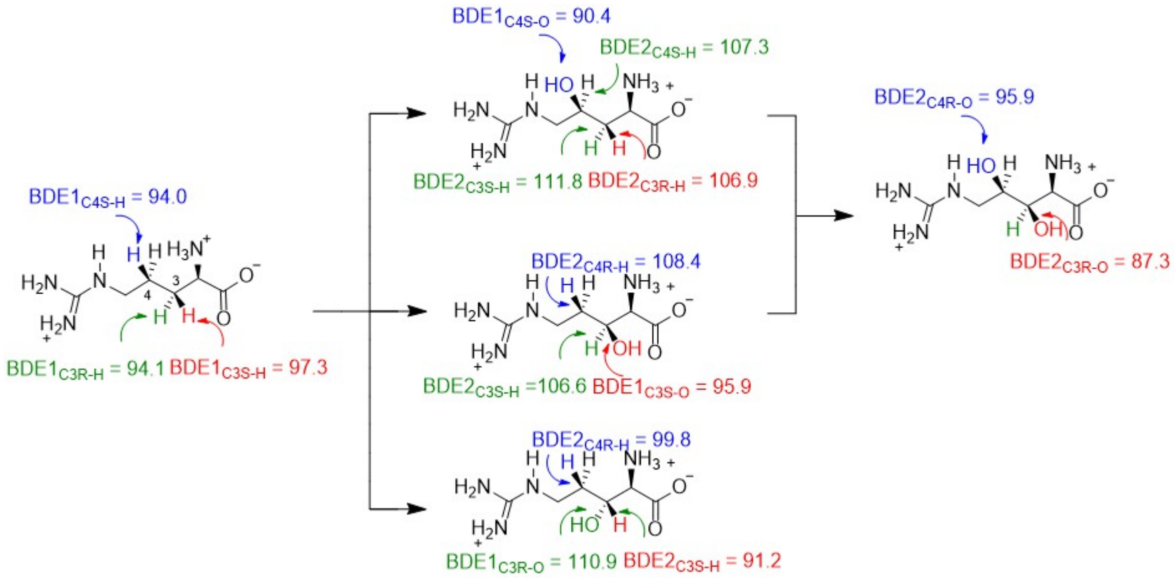

BDE1C4S-O = 90.4
BDE2C4S-H = 107.3
BDE2C4R-O = 95.9
BDE2C3S-H = 111.8 BDE2C3R-H = 106.9
BDE1C4S-H = 94.0
BDE2C3R-O = 87.3
BDE2C4R-H = 108.4
BDE1C3R-H = 94.1 BDE1C3S-H = 97.3
BDE2C3S-H =106.6 BDE1C3S-O = 95.9
BDE2C4R-H = 99.8
BDE1C3R-O = 110.9 BDE2C3S-H = 91.2

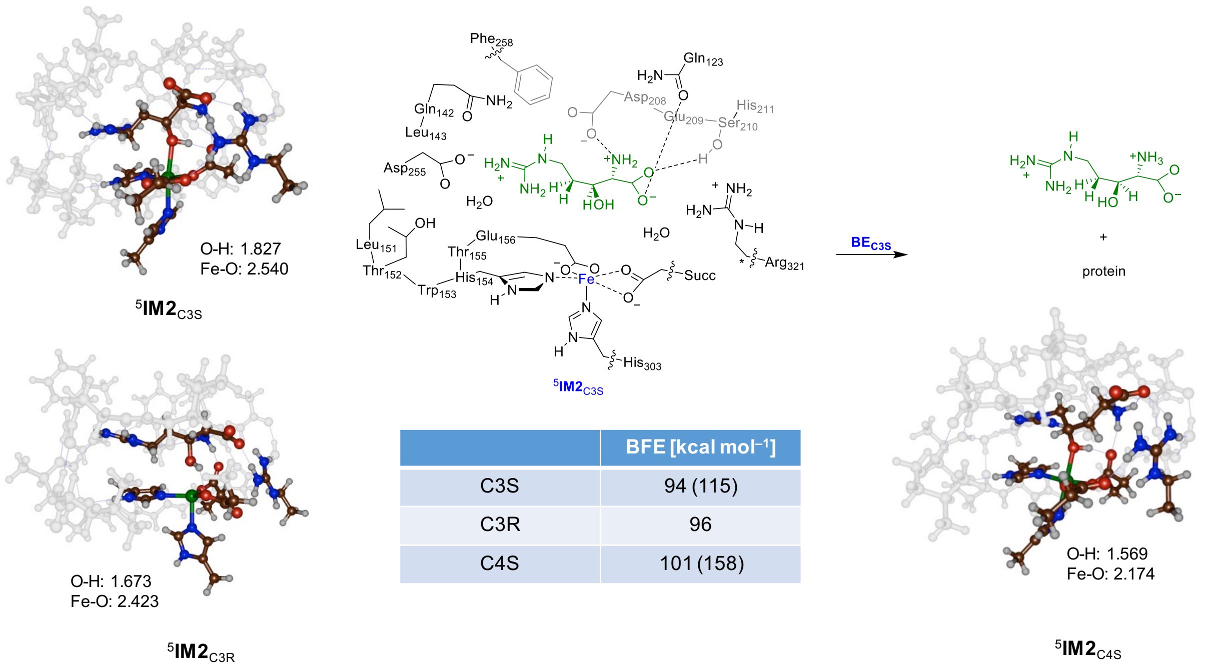
Figure 4.10. UB3LYP/6-311++G**+ZPE calculated C–H bond dissociation energies for various bonds in L-Arg. Values in kcal mol^{-1}.

We also calculated the strength of various possible C–OH bonds that are formed, designated BDE1$_{\text{C-O}}$, via a similar procedure as the BDE1$_{\text{CH}}$ values, but where we keep the geometry of the hydroxyarginine as in the **IM2** structures. The three C–O bond strengths in 3*S*-hydroxyarginine, 3*R*-hydroxyarginine and 4*S*-hydroxyarginine are 95.9, 110.9 and 90.4 kcal mol^{-1}, respectively. Consequently, based on the thermodynamics of the reaction there are many C–H bonds in the substrate of almost equal strength. Under ideal conditions, where the substrate approach is unperturbed; therefore, a mixture of products will be formed. However, substrate positioning reduces the number of reaction products and guides the reaction to C^3–H activation selectively.

In a final set of calculations, we explored the C–H bond strengths of various C–H bonds in the hydroxyarginine complexes, where we keep the hydroxyarginine as in the geometry of the **IM2** structures. These BDE2$_{\text{CH}}$ values are given in Figure 4.10 as well. The C^3–H BDE2$_{\text{CH}}$ values in 4*S*-hydroxyarginine are 111.8 (pro-*S*) and 106.9 (pro-*R*) kcal mol^{-1}, whereas the tertiary C–H bond strength at the C^4-position is 107.3 kcal mol^{-1}. Thus, the weakest C–H bonds for 4*S*-hydroxyarginine are the pro-*R* C^3 and C^4 C–H bonds. As such it is important that the C^4–H bond points away from the reaction center as its abstraction will not lead to 3*R*,4*R*-dihydroxyarginine products. Indeed, our optimized geometry shows this bond to point upwards. For 3*S*-hydroxyarginine we calculated C–H bond energies for the pro-*R* C^4 position of 108.4 kcal mol^{-1}, while the tertiary C^3–H bond has a strength of $BDE2_{CH}$ = 106.6 kcal mol^{-1}. Also for 3*S*-hydroxyarginine it is crucial that the C^3–H bond points away from the reaction center and also in this case our optimized geometry shows it to point away from the iron atom. Finally, the 3*R*-hydroxyarginine structure (bottom structure in Figure 4.11) has a very weak tertiary pro-*S* C^3–H bond of $BDE2_{C3S\text{-}H}$ = 91.2 kcal mol^{-1} and a well stronger pro-*R* C^4–H bond of $BDE2_{C4R\text{-}H}$ = 99.8 kcal mol^{-1}. Clearly, the weakest bond will lead to the 3,3,-dihydroxylated product or give desaturation to form β-ketoarginine and hence substrate positioning needs to avoid bringing this hydrogen atom close to the active site. Consequently, positioning of 3*S*-hydroxyarginine and 4*S*-hydroxyarginine in the OrfP binding pocket gives orientations where the weak tertiary C–H bond is oriented away from the reaction center and it will be unlikely to be activated, while that is not possible for 3*R*-hydroxyarginine bound.

Finally, we estimated the binding free energy (BFE) of 3*S*-hydroxyarginine, 3*R*-hydroxyarginine and 4*S*-hydroxyarginine in the substrate binding pockets of OrfP and VioC. To this end we took the optimized geometries ^{5}**IM2**$_{C3S}$, ^{5}**IM2**$_{C3R}$ and ^{5}**IM2**$_{C4S}$ and split the system into hydroxyarginine and protein and then did a single point frequency calculation of each of the fragments to estimate the binding free energy of the hydroxyarginine in the binding pocket. The largest binding free energy (BFE) for OrfP is found for 4*S*-hydroxyarginine with a value of 101 kcal mol^{-1}, while it is 96 kcal mol^{-1} for 3*R*-hydroxyarginine and 94 kcal mol^{-1} for 3S-hydroxyarginine. These binding free energies implicate that 4*S*-hydroxyarginine will be the strongest bound and it will be difficult to release it from the substrate binding pocket, whereas 3*S*- and 3*R*-hydroxyarginine form a slightly weaker interaction with the protein pocket. Therefore, the binding free energies of the singly hydroxylated products do not give an explanation for the dihydroxylation process in OrfP. Optimized geometries of the ^{5}**IM2**$_{C3S}$, ^{5}**IM2**$_{C3R}$ and ^{5}**IM2**$_{C4S}$ structures are shown in Figure 4.11. In both ^{5}**IM2**$_{C3S}$ and ^{5}**IM2**$_{C3R}$ the Fe–O bond between the metal and the product is long (>2.4Å), which is much longer than that typically seen for a covalent bond and hence will be a weak intermolecular interaction.

The structure with hydroxyarginine the strongest bound, i.e. ^{5}**IM2**$_{C4S}$, has a very short Fe–O interaction of only 2.174 Å. It also has the shortest O–H distance of the hydroxyl group of hydroxyarginine with a neighboring oxygen atom donor: in this case a distance of 1.569 Å to the carboxylate of Glu_{156} is found. Thanks to these short distances the 4*S*-hydroxyarginine will bind stronger than 3*S*- and 3*R*-hydroxyarginine.



	BFE [kcal mol^{-1}]
C3S	94 (115)
C3R	96
C4S	101 (158)

Figure 4.11. UB3LYP/BS1 optimized geometries of singly hydroxylated Arg complexes **IM2** with bond lengths in angstroms. Also given are calculated binding free energies (BFE) of singly hydroxylated Arg in the binding pocket of OrfP (VioL) in kcal mol^{-1}.

To understand substrate binding better we also calculated a VioC model with 3*S*-hydroxyarginine and 4*S*-hydroxyarginine bound. The corresponding BFE values are 115 and 158 kcal mol^{-1}, respectively. These values implicate the hydroxyarginine binds stronger in the VioC structure than in the OrfP structure, which would contradict the product release seen in those enzymes. Probably, the larger binding energy for VioC with respect to OrfP is due to the extra carboxylic acid group in the binding area, i.e. Asp$_{270}$ that forms a strong link with the substrate. However, the size of the empty substrate binding pocket we measure from the **IM2** optimized structures in OrfP are 1161 and 1187 Å^3 for ^{5}**IM2**$_{C3S}$ and ^{5}**IM2**$_{C4S}$, respectively, while the corresponding values for the VioC structures are 1322 and 1291 Å^3. Therefore, the product binding pocket in VioC is larger and gives the substrate and product more flexibility and mobility enable the hydroxyarginine to break free and escape, while the tight substrate/product binding pocket in OrfP locks the hydroxyarginine in and prevents it from escaping. It appears, therefore, that a single additional amino acid in VioC, namely Asp$_{270}$, can pull the hydroxyarginine product away from the iron center and enable its release from the substrate binding pocket. This product-release mechanism appears to be missing in OrfP. It would be interesting to see whether the Tyr257Asp or Tyr257Glu mutations in OrfP would indeed enable release of monohydroxylated arginine from the substrate binding pocket and affect the reactivity.

4.5 Conclusion

The work presented here represents a computational study into the dihydroxylating nonheme iron dioxygenase OrfP. Using large active site cluster models we calculated the mechanism on all low-lying spin states. We show that the reaction proceeds by two consecutive hydroxylation reactions by an iron(IV)-oxo species. The first cycle has a rate-determining hydrogen atom abstraction from the C3S-position of substrate and is followed by a small rebound barrier to give 3*S*-hydroxyarginine with a small preference over 3*R*-hydroxyarginine, while the 4*S*-hydroxyarginine pathway is well higher in energy. The second cycle then binds αKG and oxygen to form another iron(IV)-oxo species. Also, the second cycle has a rate-determining hydrogen atom abstraction step with similar barriers for the pathways starting from 3*S*-hydroxyarginine and 4*S*-hydroxyarginine to form the dihydroxylated product. Interestingly, the calculations show that 3*R*-hydroxyarginine in a second cycle would – via a small reaction barrier – be converted into β-ketoarginine through a desaturation step rather than lead to hydroxylation. Overall the calculations reveal that the reaction happens through negative catalysis, where a low energy pathway, i.e. the breaking of the C^4–H bond, is avoided in favor of the breaking of a stronger bond. This selectivity is the result of substrate positioning in a very tight binding pocket that guides substrate and oxidant to the C^3–H group for substrate hydroxylation in cycle 1. Our calculations highlight the function of an active site Asp residue as a hinge to lift the monohydroxylated product out of the binding pocket of VioC, which is missing in OrfP.

4.6 References

[1] Miyanaga, F.A. and Eguchi, T., Biosynthesis of natural products containing β-amino acids. *Nat. Prod. Rep.* **2014**, *31*, 1056−1073.

[2] Wu, L.F., Meng, S. and Tang, G.L., Ferrous iron and α-ketoglutarate-dependent dioxygenases in the biosynthesis of microbial natural products. *Biochem. Biophys. Acta* **2016**, *1864*, 453–470.

[3] Berlinck, R.G.S., Bertonha A.F., Takaki M. and Rodriguez, J.P.G., The chemistry and biology of guanidine natural products. *Nat. Prod. Rep.* **2017**, *34*, 1264−1301.

[4] Gao, S.S., Naowarojna N., Cheng, R., Liu, X. and Liu, P., Recent examples of α-ketoglutarate-dependent mononuclear non-haem iron enzymes in natural product biosynthesis. *Nat. Prod. Rep.* **2018**, *35*, 792−837.

[5] Hellwig, M., The chemistry of Protein oxidation in food. *Angew. Chem. Int. Ed.* **2019**, *58*, 16742−16763.

[6] Hedges, J.B. and Ryan, K.S., Biosynthetic pathways to nonproteinogenic α-amino acids. *Chem. Rev.* 2020, *120*, 3161−3209.

[7] Solomon, E.I., Brunold, T.C., Davis, M.I., Kemsley, J.N., Lee, S.K., Lehnert, N., Neese, F., Skulan, A.J., Yang, Y.S. and Zhou, J., Geometric and electronic structure/function correlations in non-heme iron enzymes. *Chem. Rev.* **2000**, *100*, 235–349.

[8] Bugg, T.D.H., Oxygenases: mechanisms and structural motifs for O2 activation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 550−555.

[9] Ryle, M.J. and Hausinger, R.P., Non-heme iron oxygenases. *Curr. Opin. Chem. Biol.* 2002, *6*, 193−201.

[10] Costas, M., Mehn, M.P., Jensen, M.P. and Que Jr, L., Dioxygen activation at mononuclear nonheme iron active sites: Enzymes, models, and intermediates. *Chem. Rev.* **2004**, *104*, 939–986.

[11] Abu-Omar, M.M., Loaiza, A. and Hontzeas, N., Reaction mechanisms of mononuclear non-heme iron oxygenases. *Chem. Rev.* **2005**, *105*, 2227–2252.

[12] Krebs, C., Fujimori, D.G., Walsh, C.T. and Bollinger Jr, J.M., Non-heme Fe(IV)–oxo intermediates. *Acc. Chem. Res.* **2007**, *40*, 484–492.

[13] de Visser, S.P. and Kumar, D., (Eds.) Iron-containing enzymes: Versatile catalysts of hydroxylation reactions in nature, Royal Society of Chemistry Publishing, Cambridge (UK), **2011**.

[14] White, M.D. and Flashman, E., Catalytic strategies of the non-heme iron dependent oxygenases and their roles in plant biology. *Curr. Opin. Chem. Biol.* **2016**, *31*, 126–135.

[15] Schofield, C.J. and Zhang, Z., Structural and mechanistic studies on 2-oxoglutarate-dependent oxygenases and related enzymes. *Curr. Opin. Struc. Biol.* **1999**, *9*, 722–731.

[16] Gorres, K. and Raines, R.T., Prolyl 4-hydroxylase. *Crit. Rev. Biochem. Mol. Biol.* **2010**, *45*, 106–124.

[17] McDonough, M.A., Li, V., Flashman, E., Chowdhury, R., Mohr, C., Lienard, B.M., Zondlo, J., Oldham, N.J., Clifton, I.J., Lewis, J., McNeill, L.A., Kurzeja, R.J., Hewitson, K.S., Yang, E., Jordan, S., Syed, R.S. and Schofield, C.J., Cellular oxygen sensing: Crystal structure of hypoxia-inducible factor prolyl hydroxylase (PHD2). *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9814–9819.

[18] de Visser, S.P., Mechanistic insight on the activity and substrate selectivity of nonheme iron dioxygenases. *Chem. Record.* **2018**, *18*, 1501–1516.

[19] Baud, D., Saaidi, P.L., Monfleur, A., Harari, M., Cuccaro, J., Fossey, A., Besnard, M., Debard, A., Mariage, A., Pellouin, V., Petit, J.L., Salanoubat, M., Weissenbach, J., de Berardinis, V. and Zaparucha, A., Synthesis of mono- and dihydroxylated amino acids with new α-ketoglutarate-dependent dioxygenases: biocatalytic oxidation of C-H bonds. *ChemCatChem*. **2014**, *6*, 3012–3017.

[20] Carter II, J.H., Du Bus, R.H., Dyer, J.R., Floyd, J.C., Rice, K.C. and Shaw, P.D., Biosynthesis of viomycin. I. Origin of alpha, beta-diaminopropionic acid and serine. *Biochemistry* **1974**, *13*, 1227–1233.

[21] Yin, X. and Zabriskie, T.M., VioC is a non-heme iron, α-ketoglutarate-dependent oxygenase that catalyzes the formation of 3S-hydroxy-L-arginine during viomycin biosynthesis. *ChemBioChem*. **2004**, *5*, 1274–1277.

[22] Helmetag, V., Samel, S.A., Thomas, M.G., Marahiel, M.A. and Essen, L.O., Structural basis for the erythro-stereospecificity of the L-arginine oxygenase VioC in viomycin biosynthesis. *FEBS J.* **2009**, *276*, 3669–3682.

[23] Mitchell, A.J., Dunham, N.P., Martinie, R.J., Bergman, J.A., Pollock, C.J., Hu, K., Allen, B.D., Chang, W.C., Silakov, A., Bollinger Jr, J.M., Krebs, C. and Boal, A.K., Visualizing the Reaction Cycle in an Iron(II)- and 2-(Oxo)-glutarate-Dependent Hydroxylase. *J. Am. Chem. Soc*. **2017**, *139*, 13830–13836.

[24] Dunham, N.P., Mitchell, A.J., Del Río Pantoja, J.M., Krebs, C., Bollinger Jr, J.M. and Boal, A.K., *Biochemistry* **2018**, *57*, 6479–6488.

[25] Dunham, N.P., Chang, W.C., Mitchell, A.J., Martinie, R.J., Zhang, B., Bergman, J.A., Rajakovich, L.J., Wang, B., Silakov, A., Krebs, C., Boal, A. K. and Bollinger Jr, J.M., Two distinct mechanisms for C-C desaturation by iron(II)- and 2-(Oxo)glutarate-dependent oxygenases: importance of α-heteroatom assistance. *J. Am. Chem. Soc.* **2018**, *140*, 7116–7126.

[26] Zwick III, C.R., Sosa, M.B. and Renata, H., Characterization of a citrulline 4-hydroxylase from nonribosomal peptide GE81112 biosynthesis and engineering of its substrate specificity for the chemoenzymatic synthesis of enduracididine. *Angew. Chem. Int. Ed*. **2019**, *58*, 18854–18858.

[27] Martinez, S., Fellner, M., Herr, C.Q, Ritchie, A., Hu, J. and Hausinger, R.P., Structures and mechanisms of the non-heme Fe(II)- and 2-oxoglutarate-dependent ethylene-forming enzyme: Substrate binding creates a twist. *J. Am. Chem. Soc*. **2017**, *139*, 11980–11988.

[28] Zhang, Z., Smart, T.J., Choi, H., Hardy, F., Lohans, C.T., Abboud, M.I., Richardson, M.S.W., Paton, R.S., McDonough, M.A. and Schofield, C. J., Structural and stereo-electronic insights into oxygenase-catalyzed formation of ethylene from 2-oxoglutarate. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 4667–4672.

[29] Li, M., Martinez, S., Hausinger, R.P. and Emerson, J.P., Thermodynamics of Iron(II) and substrate binding to the ethylene-forming enzyme. *Biochemistry* **2018**, *57*, 5696–5705.

[30] Guo, Z., Li, J., Qin, H., Wang, M., Lv, X., Li, X. and Chen, Y., Biosynthesis of the carbamoylated D-gulosamine moiety of streptothricins: Involvement of a guanidino-*N*-glycosyltransferase and an *N*-Acetyl-D-gulosamine deacetylase. *Angew. Chem. Int. Ed.* **2015**, *54*, 5175–5178.

[31] Masuda, Y., Maruyama, C., Kawabata, K., Hamano, Y. and Doi, T., Synthesis of (2S,3R,4R)-3,4-dihydroxyarginine and its inhibitory activity against nitric oxide synthase. *Tetrahedron.* **2016**, *72*, 5602–5611.

[32] Chang, C.Y., Lyu, S.Y., Liu, Y.C., Hsu, N.S., Wu, C.C., Tang, C.F., Lin, K.H., Ho, J.Y., Wu, C.J., Tsai, M.D. and Li, T.L., Biosynthesis of streptolidine involved two unexpected intermediates produced by a dihydroxylase and a cyclase through unusual mechanisms. *Angew. Chem. Int. Ed.* **2014**, *53*, 1943–1948.

[33] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., The protein data bank. *Nucl. Acids Res*. **2000**, *28*, 235–242.

[34] Bruijnincx, P.C.A., van Koten, G. and Klein Gebbink, R.J.M., Mononuclear non-heme iron enzymes with the 2-His-1-carboxylate facial triad: recent developments in enzymology and modeling studies. *Chem. Soc. Rev*. **2008**, *37*, 2716–2744.

[35] Kal, S. and Que Jr, L., Dioxygen activation by nonheme iron enzymes with the 2-His-1-carboxylate facial triad that generate high-valent oxoiron oxidants. *J. Biol. Inorg. Chem*. **2017**, *22*, 339–365.

[36] Iyer, S.R., Chaplin, V.D., Knapp, M.J. and Solomon, E.I., O_2 activation by nonheme Fe^{II} α-ketoglutarate-dependent enzyme variants: Elucidating the role of the facial triad carboxylate in FIH. *J. Am. Chem. Soc*. **2018**, *140*, 11777−11783.

[37] Proshlyakov, D.A., Henshaw, T.F., Monterosso, G.R., Ryle, M.J. and Hausinger, R.P., Direct detection of oxygen intermediates in the non-heme Fe enzyme taurine/alpha-ketoglutarate dioxygenase. *J. Am. Chem. Soc*. **2004**, *126*, 1022−1023.

[38] Bollinger Jr, J.M., Price, J.C., Hoffart, L.M., Barr, E.W. and Krebs, C., Mechanism of Taurine: α-Ketoglutarate Dioxygenase (TauD) from *Escherichia coli*. *Eur. J. Inorg. Chem*. **2005**, 4245–4254.

[39] Price, J.C., Barr, E.W., Tirupati, B., Bollinger Jr, J.M. and Krebs C., The first direct characterization of a high-valent iron intermediate in the reaction of an alpha-ketoglutarate-dependent dioxygenase: a high-spin FeIV complex in taurine/alpha-ketoglutarate dioxygenase from Escherichia coli. *Biochemistry* **2003**, *42*, 7497−7508.

[40] Riggs-Gelasco, P.J., Price, J.C., Guyer, R.B., Brehm, J.H., Barr, E.W., Bollinger Jr, J.M. and Krebs, C., EXAFS Spectroscopic Evidence for an Fe═O Unit in the Fe(IV) intermediate observed during oxygen activation by taurine:α-ketoglutarate dioxygenase. *J. Am. Chem. Soc*. **2004**, *126*, 8108–8109.

[41] Yan, W., Song, H., Song, F., Guo, Y., Wu, C.H., Her A.S., Pu Y., Wang, S., Naowarojna, N., Weitz, A., Hendrich, M.P., Costello, C.E., Zhang, L., Liu, P. and Zhang, Y.J., Endoperoxide formation by an α-ketoglutarate-dependent mononuclear non-haem iron enzyme. *Nature* **2015**, *527*, 539–543.

[42] Galonić Fujimori, D., Barr, E.W., Matthews, M.L., Koch, G.M., Yonce, J.R., Walsh, C.T., Bollinger Jr, J.M., Krebs, C. and Riggs-Gelasco, P.J., Experimental correlation of substrate position with reaction outcome in the aliphatic halogenase, SyrB2. *J. Am. Chem. Soc*. **2007**, *129*, 13408−13409.

[43] Blomberg, M.R.A., Borowski, T., Himo, F., Liao, R.Z. and Siegbahn, P.E.M., Quantum chemical studies of mechanisms for metalloenzymes. *Chem. Rev*. **2014**, *114*, 3601–3658.

[44] de Visser, S.P., Quesne, M.G., Martin, B., Comba, P. and Ryde, U., Computational modelling of oxygenation processes in enzymes and biomimetic model complexes. *Chem. Commun*. **2014**, *50*, 262–282.

[45] Quesne, M.G., Borowski, T. and de Visser, S.P., Quantum mechanics/molecular mechanics modeling of enzymatic processes: Caveats and breakthroughs. *Chem. Eur. J.* **2016**, *22*, 2562−2581.

[46] Pickl, M., Kurakin, S., Cantú Reinhard, F.G., Schmid, P., Pöcheim, A., Winkler, C.K., Kroutil, W., de Visser, S.P. and Faber, K., Mechanistic studies of fatty acid activation by CYP152 peroxygenases reveal unexpected desaturase activity. *ACS Catal*. **2019**, *9*, 565–577.

[47] Grosdidier, A., Zoete, C. and Michielin, O., SwissDock, A protein-small molecule docking web service based on EADock DSS. *Nucl. Acids Res*. **2011**, *39*, 270−277.

[48] Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G. S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E., UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem*. **2004**, *25*, 1605–1612.

[49] Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.

[50] Case, D.A., Ben-Shalom, I.Y., Cerutti, D.S., Cheatham, T.E., III, Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Homeyer, N., et al. AMBER 2018. *University of California, San Francisco.* **2018**.

[50] Wang, J.M., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[51] Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. and Simmerling C., ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

[52] Wang, J.M., Wang, W., Kollman, P.A. and Case, D.A., Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

[53] Li, P. and Merz, K.M., MCPB.py: A Python based metal center parameter Bbuilder. *J. Chem. Inf. Model.* **2016**, *56*, 599−604.

[54] Loncharich, R.J., Brooks, B.R. and Pastor R.W., Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **1992**, *32*, 523–535.

[55] Lin, Y., Pan, D., Li, J., Zhang, L. and Shao, X., Application of Berendsen barostat in dissipative particle dynamics for nonequilibrium dynamic simulation. *J. Chem. Phys.* **2017**, *146*, 124108.

[56] 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., et al., Fox, Gaussian, Inc., Wallingford CT, 2010.

[57] Quesne, M.G., Latifi, R., Gonzalez-Ovalle, L.E., Kumar, D. and de Visser, S.P., Quantum mechanics/molecular mechanics study on the oxygen binding and substrate hydroxylation step in AlkB repair enzymes. *Chem. Eur. J.* **2014**, *20*, 435–446.

[58] Timmins, A., Saint-André, M. and de Visser, S.P., Understanding how prolyl-4-hydroxylase structure steers a ferryl oxidant toward scission of a strong C–H bond. *J. Am. Chem. Soc*. **2017**, *139*, 9855−9866.

[59] Becke, A.D., Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys*. **1993**, *98*, 5648–5652.

[60] Lee, C., Yang, W. and Parr, R.G., Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

[61] Hay, P.J. and Wadt, W.R., *Ab initio* effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *J. Chem. Phys*. **1985**, *82*, 270–283.

[62] Hehre, W.J., Ditchfield, R. and Pople, J.A., Self-consistent molecular orbital methods. XIV. An extended Gaussian-type basis for molecular orbital studies of organic molecules. inclusion of second row elements. *J. Chem. Phys*. **1972**, *56*, 2257–2261.

[63] Kaczmarek, M.A., Malhotra, A., Balan, G.A., Timmins A., de Visser, S.P., Quantum mechanics/molecular mechanics studies on the relative reactivities of compound I and II in cytochrome P450 enzymes. *Chem. Eur. J.* **2018**, *24*, 5293−5302.

[64] Peng, C. and Schlegel, H.B., Combining synchronous transit and quasi-newton methods to find transition states. *Isr. J. Chem*. **1993**, *33*, 449–454.

[65] Grimme, S., Antony, J., Ehrlich, S. and Krieg, H., A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys*. **2010**, *132*, 154104.

[66] Cossi, M., Scalmani, G., Rega, N. and Barone, V., New developments in the polarizable continuum model for quantum mechanical and classical calculations on molecules in solution. *J. Chem. Phys*. **2002**, *117*, 43−54.

[67] Barman, P., Upadhyay, P., Faponle, A.S., Kumar, J., Nag, S.S., Kumar, D., Sastri, C.V. and de Visser, S.P., Deformylation reaction by a nonheme manganese(III)-peroxo complex via initial hydrogen-atom abstraction. *Angew. Chem. Int. Ed.* **2016**, *55*, 11091–11095.

[68] Mukherjee, G., Lee, C.W.Z., Nag, S.S., Cantú Reinhard, F.G., Kumar, D., Sastri, C.V. and de Visser, S. P., Dramatic rate-enhancement of oxygen atom transfer by an iron(iv)-oxo species by equatorial ligand field perturbations. *Dalton Trans.* **2018**, *47*, 14945–14957.

[69] Borowski,T. and Bassan A., and Siegbahn, P.E.M., Bioinorganic reaction mechanisms – quantum chemistry approach. *Chem. Eur. J.* **2004**, *10*, 1031–1041.

[70] de Visser, S.P., Propene activation by the oxo-iron active species of taurine/α-ketoglutarate dioxygenase (TauD) enzyme. How does the catalysis compare to heme-enzymes? *J. Am. Chem. Soc.* **2006**, *128*, 9813–9824.

[71] Sinnecker, S., Svensen, N., Barr, E.W., Ye, S., Bollinger Jr, J.M., Neese, F. and Krebs, C., Spectroscopic and computational evaluation of the structure of the high-spin Fe(IV)-oxo intermediates in taurine: alpha-ketoglutarate dioxygenase from Escherichia coli and its His99Ala ligand variant. *J. Am. Chem. Soc.* **2007**, *129*, 6168–6179.

[72] Bushnell, E.A.C., Fortowsky, G.B. and Gauld, J.W., Model iron-oxo species and the oxidation of imidazole: insights into the mechanism of OvoA and EgtB? *Inorg. Chem.* **2012**, *51*, 13351–13356.

[73] Kulik, H.J. and Drennan, C.L., Substrate placement influences reactivity in non-heme Fe(II) halogenases and hydroxylases. *J. Biol. Chem.* **2013**, *288*, 11233–11241.

[74] Wójcik, A., Radoń, M. and Borowski, T., Mechanism of O_2 activation by α-ketoglutarate dependent oxygenases revisited. A quantum chemical study. *J. Phys. Chem. A* **2016**, *120*, 1261–1274.

[75] Song, X., Lu, J. and Lai, W., Mechanistic insights into dioxygen activation, oxygen atom exchange and substrate epoxidation by AsqJ dioxygenase from quantum mechanical/molecular mechanical calculations. *Phys. Chem. Chem. Phys.* **2017**, *19*, 20188–20197.

[76] Su, H., Sheng, X., Zhu, W., Ma, G. and Liu, Y., Mechanistic insights into the decoupled desaturation and epoxidation catalyzed by dioxygenase AsqJ involved in the biosynthesis of quinolone alkaloids. *ACS Catal.* **2017**, *7*, 5534–5543.

[77] Manna, R.N., Malakar, T., Jana, B. and Paul, A., Unraveling the crucial role of single active water molecule in the oxidative cleavage of aliphatic C–C bond of 2,4′-dihydroxyacetophenone catalyzed by 2,4′-dihydroxyacetophenone dioxygenase enzyme: A quantum mechanics/molecular mechanics investigation. *ACS Catal.* **2018**, *8*, 10043–10050.

[78] Lin, Y.T., Stańczak, A., Manchev, Y., Straganz, G.D. and de Visser, S.P., Second-coordination sphere effects on selectivity and specificity of heme and nonheme iron E

enzymes. *Chem. Eur. J.* **2020**, *26*, 2233–2242.

[79] Latifi, R., Sainna, M.A., Rybak-Akimova, E.V. and de Visse, S.P., Does hydrogen-bonding donation to manganese(IV)-oxo and iron(IV)-oxo oxidants affect the oxygen-atom transfer ability? A computational study. *Chem. Eur. J.* **2013**, *19*, 4058–4068.

[80] Sahoo, D., Quesne, M.G., de Visser, S.P. and Rath, S.P., Hydrogen-bonding interactions trigger a spin-flip in iron(III) porphyrin complexes. *Angew. Chem. Int. Ed.* **2015**, *54*, 4796–4800.

[81] Godfrey, E., Porro, C.S. and de Visser, S.P., Comparative quantum mechanics /molecular mechanics (QM/MM) and density functional theory calculations on the Oxo−Iron Species of Taurine/α-ketoglutarate dioxygenase. *J. Phys. Chem. A* **2008**, *112*, 2464–2468.

[82] Diebold, A.R., Brown-Marshall, C.D., Neidig, M.L., Brownlee, J.M., Moran, G.R. and Solomon, E.I., Activation of α-keto acid-dependent dioxygenases: Application of an $\{FeNO\}^7/\{FeO_2\}^8$ methodology for characterizing the initial steps of O_2 activation. *J. Am. Chem. Soc.* **2011**, *133*, 18148–18160.

[83] Wick, C.R., Lanig, H., Jäger, C.M., Burzlaf, N. and Clark, T., Structural insight into the prolyl hydroxylase PHD2: A molecular dynamics and DFT study. *Eur. J. Inorg. Chem.* **2012**, 4973–4985.

[84] Dong, G., Shaik, S. and Lai, W., Oxygen activation by homoprotocatechuate 2,3-dioxygenase: A QM/MM study reveals the key intermediates in the activation cycle. *Chem. Sci.* **2013**, *4*, 3624–3635.

[85] Wójcik, A., Radoń, M. and Borowski, T., Mechanism of O2 activation by α-ketoglutarate dependent oxygenases revisited. A quantum chemical study. *J. Phys. Chem. A* **2016**, *120*, 1261−1275.

[86] Song, X., Lu, J. and Lai, W., Mechanistic insights into dioxygen activation, oxygen atom exchange and substrate epoxidation by AsqJ dioxygenase from quantum mechanical/molecular mechanical calculations. *Phys. Chem. Chem. Phys.* **2017**, *19*, 20188–20197.

[87] Faponle, A.S., Seebeck, F.P. and de Visser, S.P., Sulfoxide synthase versus cysteine dioxygenase reactivity in a nonheme iron enzyme. *J. Am. Chem. Soc.* **2017**, *139*, 9259–9270.

[88] Chaturvedi, S.S., Ramanan, R., Lehnert, N.I., Schofield, C. J., Karabencheva-Christova, T.G. and Christov, C.Z., Catalysis by the non-heme Iron(II) histone demethylase PHF8 involves iron center rearrangement and conformational modulation of substrate orientation. *ACS Catal*. **2020**, *10*, 1195−1209.

[89] de Visser, S.P., What factors influence the ratio of C−H hydroxylation versus C=C epoxidation by a nonheme cytochrome P450 biomimetic? *J. Am. Chem. Soc*. **2006**, *128*, 15809−15818;

[90] Decker, A., Rohde, J.U., Klinker, E.J., Wong, S.D., Que Jr, L. and Solomon, E.I., Spectroscopic and quantum chemical studies on low-spin Fe^{IV}=O complexes: Fe−O bonding and its contributions to reactivity. *J. Am. Chem. Soc*. **2007**, *129*, 15983−15996.

[91] Cantú Reinhard, F.G., Faponle, A.S. and de Visser, S.P., Substrate sulfoxidation by an iron(IV)-oxo complex: Benchmarking computationally calculated barrier heights to experiment. *J. Phys. Chem. A* **2016**, *120*, 9805−9814.

[92] Shaik, S., Kumar, D. and de Visser, S.P., A valence bond modeling of trends in hydrogen abstraction barriers and transition states of hydroxylation reactions catalyzed by cytochrome P450 enzymes. *J. Am. Chem. Soc*. **2008**, *130*, 10128−10140.

[93] Postils, V., Sun, W., Li, X.X., Faponle, A.S., Solà, M., Wang, Y., Nam, W. and de Visser, S.P., Quantum mechanics/molecular mechanics studies on the relative reactivities of compound I and II in cytochrome P450 enzymes. *Chem. Eur. J*. **2017**, *23*, 6406–6418.

[94] Barman, P., Cantú Reinhard, F.G., Bagha, U.K., Kumar, D., Sastri, C.V. and de Visser, S.P., Hydrogen by deuterium substitution in an aldehyde tunes the regioselectivity by a nonheme manganese(III)–peroxo complex. *Angew. Chem. Int. Ed*. **2019**, *58*, 10639–10643.

[95] Kumar, D., Hirao, H., Que Jr, L. and Shaik, S., Theoretical investigation of C−H hydroxylation by (N4Py)Fe^{IV}=O^{2+}: An oxidant more powerful than P450? *J. Am. Chem. Soc*. **2005**, *127*, 8026−8027.

[96] de Visser, S.P., Can the peroxosuccinate complex in the catalytic cycle of taurine/α-ketoglutarate dioxygenase (TauD) act as an alternative oxidant? A density functional study. *Angew. Chem. Int. Ed*. **2006**, *45*, 1790−1793.

[97] Geng, C., Ye, S. and Neese, F., Analysis of reaction channels for alkane hydroxylation by nonheme iron(IV)–oxo complexes. *Angew. Chem. Int. Ed*. **2010**, *49*, 5717−5720.

[98] Ye, S. and Neese, F., Nonheme oxo-iron(IV) intermediates form an oxyl radical upon approaching the C–H bond activation transition state. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1228−1233.

[99] Bernasconi, L. and Baerends, E.J., A frontier orbital study with ab initio molecular dynamics of the effects of solvation on chemical reactivity: Solvent-induced orbital control in FeO activated hydroxylation reactions. *J. Am. Chem. Soc*. **2013**, *135*, 8857−8867.

[100] Timmins, A., Quesne, M.G., Borowski, T. and de Visser, S.P., Group transfer to an aliphatic bond: A biomimetic study inspired by nonheme iron halogenases. *ACS Catal*. **2018**, *8*, 8685−8698.

[101] Bordwell, F.G. and Cheng, J.P., Trends in substrate hydroxylation reactions by heme and nonheme iron(IV)-oxo oxidants give correlations between intrinsic properties of the oxidant with barrier height. *J. Am. Chem. Soc*. **1991**, *113*, 1736–1743.

[102] Mayer, J.M., Proton-coupled electron transfer: A reaction chemit's view. *Annu. Rev. Phys. Chem*. **2004**, *55*, 363–390.

[103] Ghafoor, S., Mansha, A. and de Visser, S.P., Selective hydrogen atom abstraction from dihydroflavonol by a nonheme iron center is the key step in the enzymatic flavonol synthesis and avoids by-products. *J. Am. Chem. Soc*. **2019**, *141*, 20278−20292.

Chapter 5. Lignin biodegradation by a cytochrome P450 enzyme: A computational study into syringol activation by GcoA

PAPER 4

Hafiz Saqib Ali[1,2], Richard H. Henchman[1,2], and Sam P. de Visser[1,3]

[1] Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[2] School of Chemistry, The University of Manchester, Oxford Road, Manchester M13 9PL, UK
[3] Department of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

* Corresponding author:

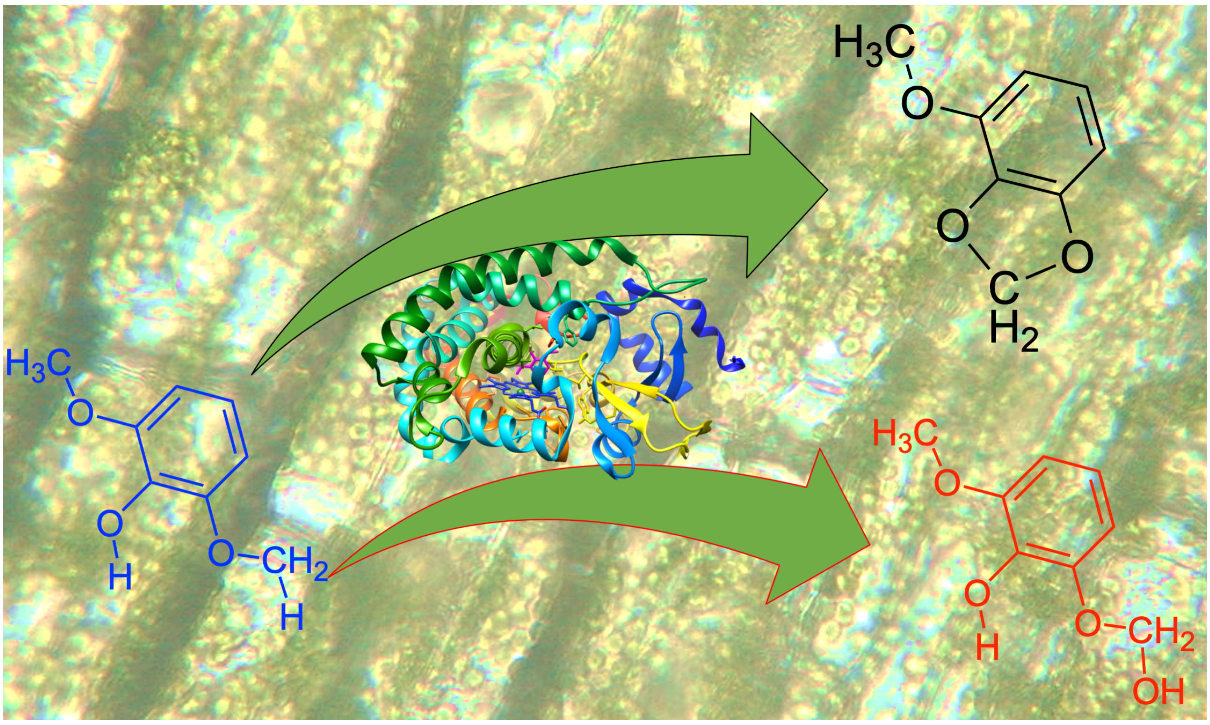
E-mail: sam.devisser@manchester.ac.uk

Conceptualization, supervision and project administration, S.P.V.; investigation and funding acquisition, H.S.A.; validation, visualization and writing—original draft preparation, H.S.A. and S.P.V.; formal analysis and writing—review and editing, H.S.A., R.H.H. and S.P.V.

Published in *Chemistry — A European Journal*, Volume 26, Issue 57, Page 13093-13102, 02 July 2020.

Abstract

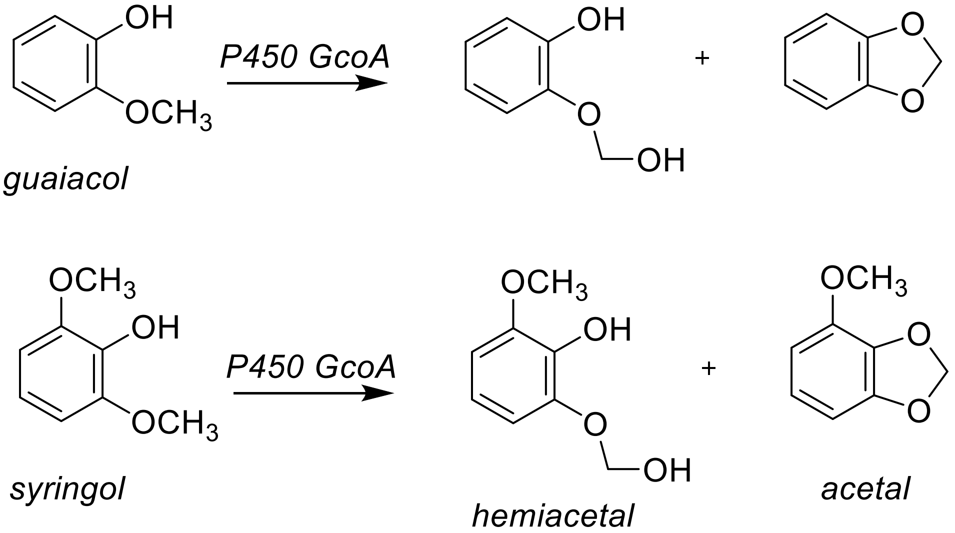
A recently characterized cytochrome P450 isozyme GcoA activates lignin components through a selective *O*-demethylation or alternatively an acetal formation reaction. These are important reactions in biotechnology and since lignin is the main component in plant cell walls it is readily available. In this work we present a density functional theory study on a large active site model of GcoA to investigate syringol activation by an iron(IV)-oxo heme cation radical oxidant (Compound I) leading to hemiacetal and acetal products. Several substrate-binding positions were tested and full energy landscapes calculated. The study shows that substrate positioning determines the product distributions. Thus, with the phenol group pointing away from the heme, an *O*-demethylation is predicted, whereas an initial hydrogen-atom abstraction of the weak phenolic O–H group would trigger a pathway leading to ring-closure to form acetal products. Predictions on how to engineer P450 GcoA to get more selective product distributions are given.



5.1 Introduction

Lignin is a complex biopolymer that makes up the cell walls and tissues in plants as well as in some fungi. It is built up from mainly aromatic and phenolic residues bridged by ether and C–C bonds and has a highly branched structure that gives it its chemical and physical strength and biological properties. Several enzymes in nature can biodegrade lignin or parts thereof, including the lignin peroxidases, which contain a heme active site and utilize H_2O_2 as an oxidant [1-3]. Currently, the agricultural and industrial sectors generate substantial amounts of lignocellulose, much of which currently goes to waste. However, lignocellulose has the potential to be converted into valuable materials or used as an energy source. Therefore, ongoing studies to find biotechnological applications of lignin degrading enzymes are being conducted to convert lignin into small aromatic compounds or drugs [4-6].

Recently, it was found that the cytochromes P450 can also participate in lignin degradation pathways [7-9]. These P450 enzymes are heme monoxygenases that utilize molecular oxygen, often as a means to hydroxylate aromatic or aliphatic substrates, although dealkylation reactions have also been reported [10-18]. In particular, the P450 isozyme CYP255A (GcoA) was found to demethylate aromatic compounds such as those originating from lignin components, including guaiacol and various alkoxybenzoates [19,20]. The work showed that engineered GcoA isozymes with enlarged substrate binding pockets, e.g. through replacement of Phe_{169} by Ala, enhanced the reactivity with these substrates. Furthermore, studies with a variety of *O*-methoxy-aromatic compounds measured product distributions as well as substrate binding affinities and constants [21]. A combined experimental and computational study looked into the mechanisms and possibilities of guaiacol activation by GcoA. Two pathways were considered, namely *O*-demethylation proposed to start with methoxy hydroxylation to form hemiacetal and ring-closure to form acetal (Scheme 5.1), whereby the former is expected to release formaldehyde to give catechol. A minimal density functional theory (DFT) cluster model was studied that did not consider the substrate-binding pocket, but nevertheless gave insights into possible reaction pathways. Recent computational studies of us showed that the second coordination sphere has important functions in substrate and oxidant positioning and hence affects regio- and chemoselectivities of enzymatic reactions [22-24]. Since, the substrate-binding pocket in GcoA is tight with various π-stacking interactions, we felt a more advanced computational study that takes the effect of the protein into consideration would give better model and more insight into the details of the reaction mechanism of GcoA enzymes, its substrate range and selectivity.



Scheme 5.1. Possible reaction products of guaiacol and syringol activation by P450 GcoA.

Moreover, acetal-type structures and particularly cyclic ones are common in biomaterials, including corticosteroids and drug molecules like paroxetine. Hence, an enzyme that could synthesize cyclic acetal-bound structures selectively would be useful in biotechnology. Therefore, we studied the reaction mechanism of the lignin fragment syringol (2-6-methoxyphenol) activated by GcoA using a large active-site cluster model of GcoA that includes much of the substrate-binding pocket.

The *O*-dealkylation of substrates by the P450s has been observed for various isozymes and, for instance, is part of the biodegradation and metabolism of drug molecules in the liver [25-27]. Computational studies established a mechanism that starts with hydroxylation of the methyl group to form a hemiacetal-like intermediate, which in solution, upon addition of protons, releases formaldehyde to complete the *O*-demethylation process [28-31]. As such, the *O*-demethylation reaction shows similarities with aliphatic hydroxylation by P450 enzymes that generally proceeds via a stepwise mechanism with an initial hydrogen-atom abstraction by Compound I (CpdI; iron(IV)-oxo heme cation radical intermediate) to form an iron(IV)-hydroxo complex that rebounds its OH group to the substrate to form alcohol products [32-35]. Experimental support for this hypothesis came from kinetic isotope effect (KIE) studies that established a large change in the rate constant when the transferring hydrogen atom was replaced by a deuterium atom [36-38].

To gain insight into the lignin biodegradation pathways by P450 isozymes, we investigated the mechanism of syringol activation by a large GcoA model structure. In particular, we focused the work on the bifurcation pathways leading to *O*-methoxy hydroxylation and acetal formation using two substrate-binding orientations. The work shows that the protein environment is important: it sets up substrate approach, guides the reaction in a certain direction and leads to different product distributions with the different substrate orientations. As the phenol O–H bond is the weakest bond in the substrate, substrate activation preferentially takes place there, but is only possible with a substrate-bound orientation that points the phenol group in the direction of the heme. Overall competing pathways to both products were identified and analyzed.

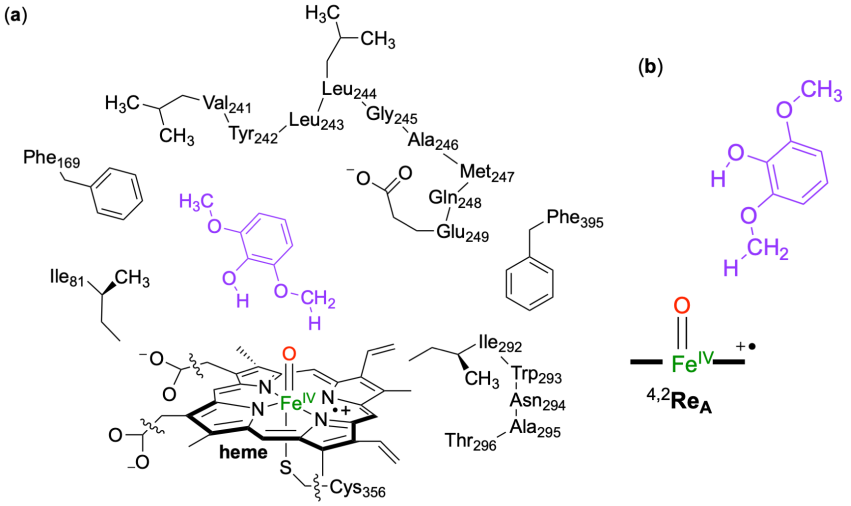
5.2 Experimental Section

The calculations reported in this work were done using density functional theory methods as implemented in Gaussian-09 software package [39]. In general, the unrestricted B3LYP hybrid density functional method [40,41] was employed in combination with a basis set containing an LANL2DZ + ECP on iron and 6-31G* on the rest of the atoms (basis set BS1) [42,43]. Full geometry optimization and frequencies were run for all structures at UB3LYP/BS1 in the gas-phase. Subsequent single-point calculations with the polarized continuum model (CPCM) were performed with a dielectric constant mimicking ethylphenylether [44], and a triple-ξ quality basis set (basis set BS2): LACV3P+ with ECP on iron and 6-311+G* on the rest of the atoms. In previous work we extensively tested and benchmarked models and methods for P450 reaction mechanisms and well reproduced experimental structures and rate constants [45-47]. These studies showed that the electronic configuration of CpdI and the general reaction mechanisms are little affected by the choice of the density functional method and basis set and most methods predict close-lying doublet and quartet spin configurations with similar hydrogen atom abstraction barriers.

5.3 Results

Focusing on lignin biodegradation by P450 isozymes, we created a large active-site cluster model of GcoA with syringol bound and studied substrate activation. Our model set-up follows previously reported procedures from our group [48-50], that start from a deposited crystal structure from the protein databank (pdb) [51], and a detailed analysis of the co-factor and substrate environment. Based on key local environmental interactions from charged residues and hydrogen bonding and stereochemical influences, we created an active site cluster model of 302 atoms as shown in Scheme 5.2 (a). The 5OMU protein databank file [21] was used for the model as it is a P450 monomer structure of GcoA with syringol bound. The residues included in our model are highlighted in Scheme 5.2 (a). We took the heme and kept all side chains except the propionate groups, which were replaced by methyl. The axial cysteinate of the heme (Cys_{356}) was included as methylmercaptate and iron(III)-heme replaced by iron(IV)-oxo heme cation radical, i.e. Compound I (CpdI). The substrate-binding pocket was described through the residues Ile_{81} (as butane), Phe_{169} and Phe_{395} (as ethylbenzene). In addition, two elaborate protein chains were included in the model, namely the chain Val_{241}-Tyr_{242}-Leu_{243}-Leu_{244}-Gly_{245}-Ala_{246}-Met_{247}-Gln_{248}-Glu_{249} and Ile_{292}-Trp_{293}-Asn_{294}-Ala_{295}-Thr_{296}. The amino acid side chains pointing away from the substrate binding pocket were replaced by Gly, namely those of Tyr_{242}, Leu_{243}, Met_{247}, Trp_{293} and Asn_{294}. The complete model was calculated in the doublet and quartet spin states. We decided to explore two different binding conformations of the substrate: model **A** with one of the methoxy groups pointing toward CpdI and model **B** that has both the phenol and one of the methoxy groups in close proximity to CpdI (Scheme 5.2 (b)). These structures were manually created and are labelled as $\mathbf{Re_A}$ and $\mathbf{Re_B}$, respectively.

DFT optimized geometries of the reactant complexes $\mathbf{Re_A}$ and $\mathbf{Re_B}$ in the doublet and quartet spin states are given in Figure 5.1. Both structures have close-lying doublet and quartet spin state configurations with three unpaired electrons in the orbitals labelled as π^*_{xz}, π^*_{yz} and a_{2u}. Thus, the metal 3d-orbitals interact with orbitals on the ligands and give the following five valence orbitals: δ_{x2-y2}, π^*_{xz}, π^*_{yz}, σ^*_{z2} and σ^*_{xy}, whereby the z-axis is defined along the S–Fe–O axis and the xy-plane is in the porphyrin plane with the axis through the Fe–N bonds. The two σ^* orbitals are virtual in CpdI, while the δ_{x2-y2} is non-bonding and doubly occupied. The singly occupied molecular orbitals of the CpdI reactant structures are shown on the left-hand-side of Figure 5.1 and represent the antibonding interactions of the metal with the oxo group (π^*_{xz} and π^*_{yz}) and a mixed porphyrin-axial ligand orbital labeled a_{2u} [51]. In the quartet spin state these three orbitals are ferromagnetically coupled, while in the doublet spin state the two π^* orbitals are antiferromagnetically coupled to the a_{2u} electron.



Scheme 5.2. (a) DFT cluster model studied in this work. Wiggly lines identify where covalent bonds were cut. (b) Substrate orientations A and B.

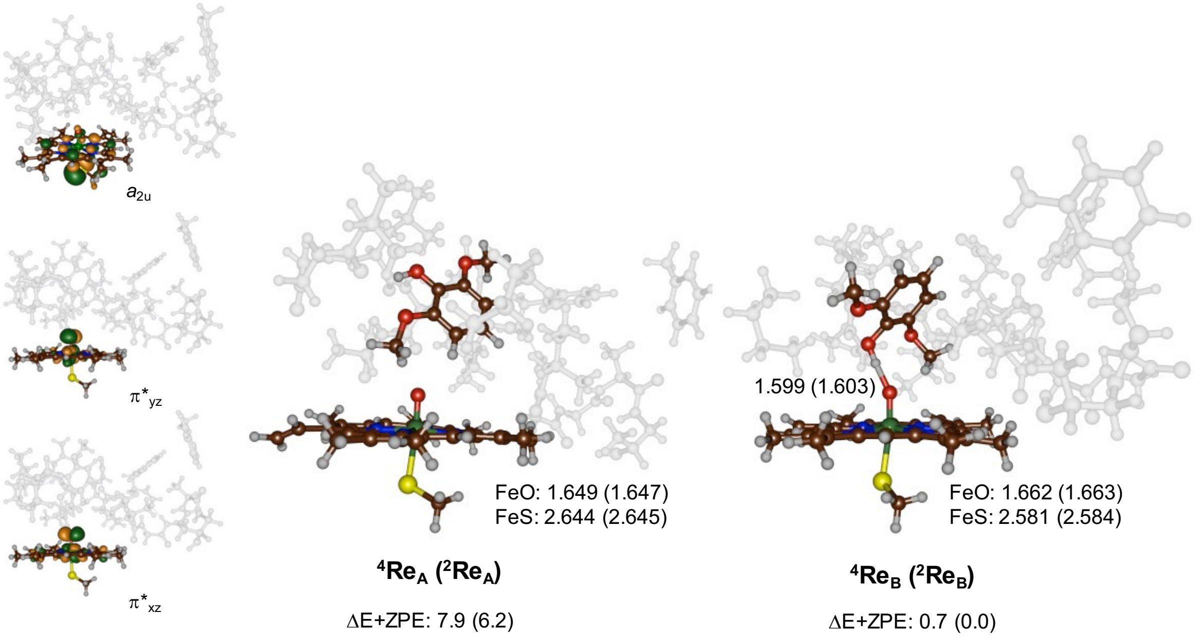
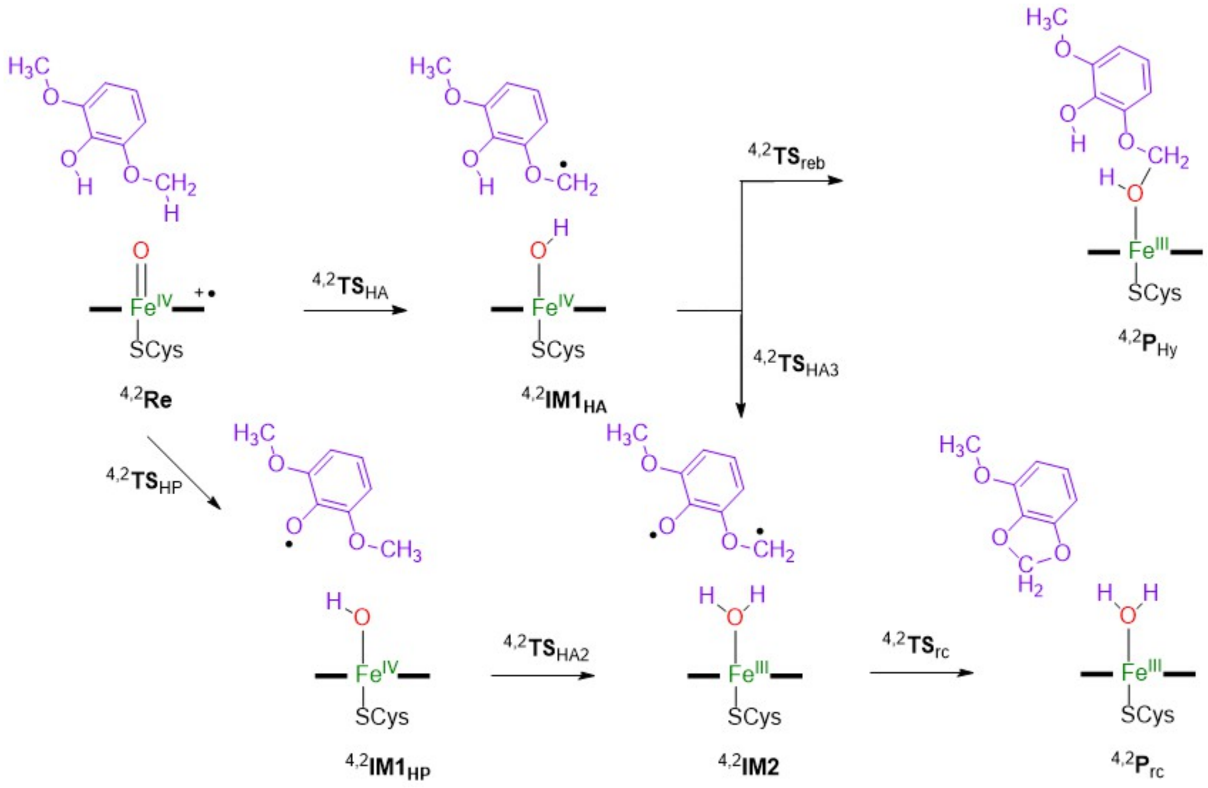


Figure 5.1. UB3LYP/BS1 optimized geometries of $^{4,2}\mathbf{Re_A}$ and $^{4,2}\mathbf{Re_B}$ with bond lengths in Angstrom (Å). Singly occupied orbitals shown on the left-hand-side for $^{2}\mathbf{Re_A}$ as an example. Relative energies (kcal mol^{-1}) are UB3LYP/BS2//UB3LYP/BS1 values with zero-point energy (ZPE) included.

As before [52-57], the doublet and quartet spin states of CpdI are close in energy as can be seen from the pairs of energies for ^{4,2}**Re$_{A}$** and ^{4,2}**Re$_{B}$**. The structures **Re$_{B}$** are the lowest in energy, probably due to the additional hydrogen bonding of the phenol group of the substrate with the oxo group of CpdI that gives these structures extra stability. Therefore, reactant configuration **Re$_{B}$** has the substrate the strongest bound and hence represents the more favorable binding orientation.

Geometrically, there are differences between the reactant complexes **Re$_{A}$** and **Re$_{B}$**, mainly due to the hydrogen bond of the phenol group of the substrate to the oxo group in **Re$_{B}$**. Thus, in **Re$_{A}$** the Fe–O distance is short (1.649 and 1.647 Å for the quartet and doublet spin states), while they are elongated to 1.662/1.663 Å for ^{4}**Re$_{B}$**/^{2}**Re$_{B}$** as a result of the hydrogen bond interaction with the phenol group at 1.60 Å. At the same time, the Fe–S bond has shortened from about 2.64 Å in **Re$_{A}$** to 2.58 Å in **Re$_{B}$**. Overall, the optimized geometries and electronic configuration matches previous studies well on CpdI models with either DFT cluster models or QM/MM [58-60].

Next we calculated the activation of syringol by CpdI using models **Re$_{A}$** and **Re$_{B}$** as starting points, whereby we give the substrate binding orientation with **A** or **B** as a subscript after the label. Details of the pathways explored with definition of the structures are given in Scheme 5.3. Firstly, we tested hydroxylation of the methoxy group of syringol for models **A** and **B** and calculated the hydrogen-atom abstraction transition state (**TS$_{HA}$**) from the methoxy C–H bond by CpdI that leads to an iron-hydroxo complex and substrate radical (**IM1$_{HA}$**). Radical rebound via **TS$_{reb}$** gives the hemiacetal product complex (**P$_{Hy}$**). Due to substrate positioning these pathways are possible for both model **A** and model **B**. However, for substrate positioning **B**, we also explored alternative pathways that involve the phenol group of the substrate. Thus, for the substrate bound in orientation **B** we investigated hydrogen-atom abstraction from the phenol group via transition state **TS$_{HP,B}$** to form the alternative radical intermediate **IM1$_{HP,B}$**. In substrate orientation **A** the phenol group points away from CpdI and hence O–H hydrogen-atom abstraction is not feasible in this orientation. From the intermediate **IM1**$_{HP,B}$ a second hydrogen-atom abstraction from the methoxy group of the substrate via transition state **TS**$_{HA2}$ was tested to give an iron(III)-water complex and a biradical on the substrate (**IM2**$_{HP,B}$). Of course, **IM2**$_{HP,B}$ can also be formed from **IM2**$_{HA,B}$ by hydrogen-atom abstraction from the phenol group in **IM1**$_{HA,B}$ via transition state **TS**$_{HA3,B}$. The biradical via a ring-closure transition state (**TS**$_{rc,B}$) leads to the acetal product complex (**P**$_{rc,B}$).



Scheme 5.3. Reaction mechanism with definition of individual structures for syringol activation by GcoA.

We first consider the substrate activation using model **A**, where only aliphatic hydroxylation of the methoxy group is possible. The calculated potential energy landscape for hydroxylation of the methoxy group of syringol by a CpdI model **A** of GcoA is given in Figure 5.2. The hydrogen-atom abstraction barriers ($^{4,2}\mathbf{TS}_{\mathrm{HA,A}}$) are relatively high in energy: 22.0 and 23.4 kcal mol^{-1} in the doublet and quartet spin states, respectively. However, these values are relative to the more stable reactant conformation $^{2}\mathbf{Re}_{\mathrm{B}}$, although relative to the reactant in the same configuration, $\mathbf{Re}_{\mathrm{A}}$, they are still 15.8 and 17.2 kcal mol^{-1} in energy. The optimized geometries of the transition states are given on the right-hand-side of Figure 5.2. Both structures have a characteristic almost linear O—H—C angle ranging from 174° – 178°, which is typical for hydrogen atom abstraction transition states [61-64]. The transition states are product-like with larger C–H than O–H distances. Generally, product-like transition states correspond to higher reaction barriers than earlier transition states [63], as confirmed from the relative energies.

The aliphatic hydrogen-atom abstraction transition states $^{4,2}\mathbf{TS_{HA,A}}$ are characterized by a large imaginary frequency i1221 cm^{-1} in the quartet spin state and i864 cm^{-1} in the doublet spin state. The imaginary frequencies in the transition states represent the C—H—O stretch vibration along the reaction coordinate. The large values for the imaginary frequency are typical of hydrogen-atom abstraction barriers and implicate a large amount of tunneling and that the reaction likely will show a large kinetic isotope effect when the transferring hydrogen atom is replaced by deuterium [65-67]. After the transition states, the system relaxes to a radical intermediate ($^{4,2}\mathbf{IM1}_{\mathrm{HA,A}}$). On both spin-state surfaces an electron transfer from the substrate into the a_{2u} orbital takes place to give an iron(IV)-hydroxo(heme) and substrate radical, whereby the substrate has up-spin radical in the quartet and down-spin in the doublet. Both π^*_{xz} and π^*_{yz} orbitals remain singly occupied in the radical intermediates $^{4,2}\mathbf{IM1}_{\mathrm{HA,A}}$.

The radical intermediates in pathway **A**, i.e. $^{4,2}\mathbf{IM1}_{\mathrm{HA,A}}$, are characterized as local minima on the potential energy surface with real frequencies only. However, the radical rebound barriers for both spin states were found to be very low in energy (< 1 kcal mol^{-1}) and hence could not be characterized. Therefore, the radical intermediates will have a short lifetime and quickly collapse to form alcohol products. Indeed, the exothermicity from radical intermediates to products $^{4,2}\mathbf{Pr}_{\mathrm{Hy,A}}$ is very large. These short radical lifetimes of the intermediate complexes also makes unlikely a possible ring-closure to form the acetal products for this substrate binding orientation and hence the reaction will be highly selective in substrate-binding position **A**. In previous work it was shown through valence bond rationalization that the doublet spin radical rebound barrier correlates with the ionization energy of the radical and the electron affinity of the iron(IV)-hydroxo complex [61,65,68,69]. In the quartet spin state the radical rebound in addition has a term for the electron excitation from the π^*_{xz} to σ^*_{z2} orbital. We calculated the ionization energy of the radical and found 166.4 kcal mol^{-1} and with the reported electron affinity of the iron(IV)-hydroxo species of 88.9 kcal mol^{-1} [70] predict a negligible rebound barrier from valence bond principles and that consequently, the rebound will be fast.

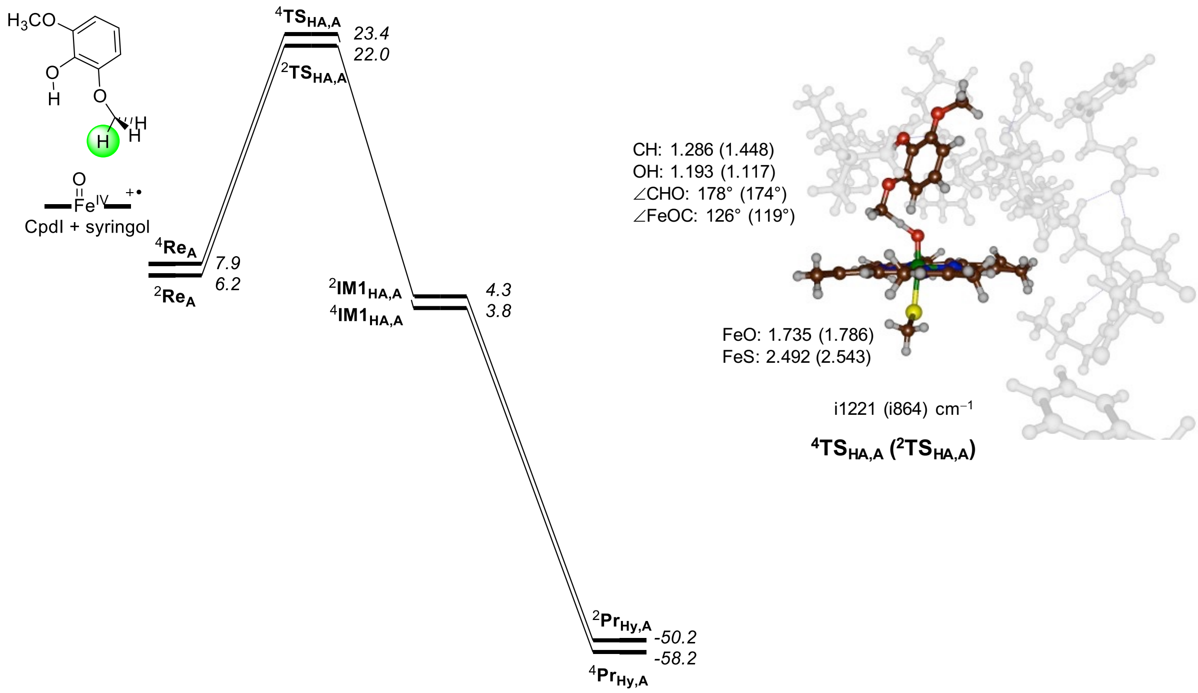


Figure 5.2. UB3LYP/BS2//UB3LYP/BS1 calculated potential energy surface for syringol activation by CpdI model **A** of GcoA. Energies contain ZPE and are given in kcal mol^{-1} relative to ^{2}**Re$_B$**. Optimized geometries of the transition states give bond lengths in angstroms, angles in degrees and the imaginary frequency in cm^{-1}.

Subsequently, the substrate activation pathways with substrate in binding position **B** were explored, and the results are in Figure 5.3. As can be seen, the lowest barriers are obtained for phenolic hydrogen-atom abstraction with a magnitude of $\Delta E^{\ddagger}$+ZPE = 0.9 and 1.6 kcal mol^{-1} in the doublet and quartet spin states. Recent work of ours on the vancomycin biosynthesis enzyme OxyB showed that two sequential phenolic hydrogen atom abstraction reactions can be performed by CpdI and CpdII (Compound II) to enable the aromatic cross-linking of glycopeptide units [71]. For the P450 OxyB system, the two hydrogen atom abstraction barriers were found to be very low in energy as the phenolic O–H bonds are very weak. The values of the hydrogen-atom abstraction barriers in GcoA are also extremely low in energy in line with the OxyB results. However, both of these sets of barriers are much lower in energy than those calculated previously for the abstraction from aliphatic C–H bonds [61-64]. For instance, using the same computational methods as used here, a hydrogen-atom abstraction barrier from the benzylic position of ethylbenzene gave a value of 12.6 kcal mol^{-1}, while for the C^5–H bond cleavage in camphor 14.5 kcal mol^{-1} was found [61]. Our aliphatic hydrogen-atom abstraction barriers from the methoxy group of syringol indeed have values of that size with ^{2}**TS$_{HA,B}$** at

13.6 kcal mol^{-1} and ^{4}**TS$_{\mathbf{HA,B}}$** at 17.4 kcal mol^{-1}. For pathway **A** the barrier height with respect to **Re$_{\mathbf{A}}$** is similar as expected because the same C–H bond is broken and same electron transfer takes place. However, since the substrate-bound complex **B** is more stable, its hydrogen-atom abstraction barriers are lower in energy. Nevertheless, the ^{2,4}**TS$_{\mathbf{HA,B}}$** and ^{2,4}**TS$_{\mathbf{HA,A}}$** structures are strikingly different. Although the substrate binding position **A** has the substrate in an upright position, its transfer of a hydrogen atom happens under an almost linear angle O–H–C of 178° and 174° for the quartet and doublet spin states, respectively. By contrast, in the ^{4,2}**TS$_{\mathbf{HA,B}}$** structures the angles are slightly more bent (169° and 165°) due to the hydrogen bond from the phenol group to the oxo that gives the substrate approach lesser flexibility. These differences in orientation also affect the C–H and O–H distances in the transition states as is seen in Figures 5.2 and 5.3.

After the aliphatic hydrogen-atom abstraction in the quartet spin state, the system relaxes to a radical intermediate (^{4}**IM1**$_{\mathrm{HA,B}}$), which is similar to the one seen for the structure in binding position **A**. However, due to additional hydrogen bonding interactions ^{4}**IM1**$_{\mathrm{HA,B}}$ is much lower in energy than ^{4}**IM1**$_{\mathrm{HA,A}}$: –5.0 kcal mol^{-1} with respect to ^{4}**Re**$_{\mathrm{B}}$. Furthermore, the reaction is followed by an almost barrierless second hydrogen-atom abstraction, namely hydrogen-atom abstraction from the phenol O–H group leads to ^{4}**IM2**$_{\mathrm{HA,B}}$ with large exothermicity. A subsequent, also barrierless ring-closure step gives the acetal products. In addition to this pathway, we attempted to calculate the OH rebound from ^{4}**IM1**$_{\mathrm{HA,B}}$ to form the hemiacetal products. However, due to hydrogen bonding interactions between the phenol group and the iron-hydroxo groups, the radical rebound is hampered. The constraint geometry scan for the radical rebound from ^{4}**IM1**$_{\mathrm{HA,B}}$ therefore, gave a barrier of at least 13.7 kcal mol^{-1}. Previously, in nonheme iron halogenases as well as in the P450 decarboxylase OleT and synthetic model complexes, we identified hydrogen bonds to an iron-hydroxo intermediate that prevented radical rebound and guided the mechanism to a side reaction [69,72,73].

Consequently, substrate positioning in GcoA enzymes is very important and determines the reaction mechanism, whereby substrate binding orientation **B** can lead to acetal products, while we do not see those products resulting from substrate binding position **A**. On the doublet spin-state surface no radical intermediate (^{2}**IM1**$_{\mathrm{HA,B}}$) could be identified and its geometry optimization fell to ^{2}**IM2**$_{\mathrm{B}}$ directly. Similarly to the high-spin this intermediate reverted to the acetal product in a barrierless fashion.

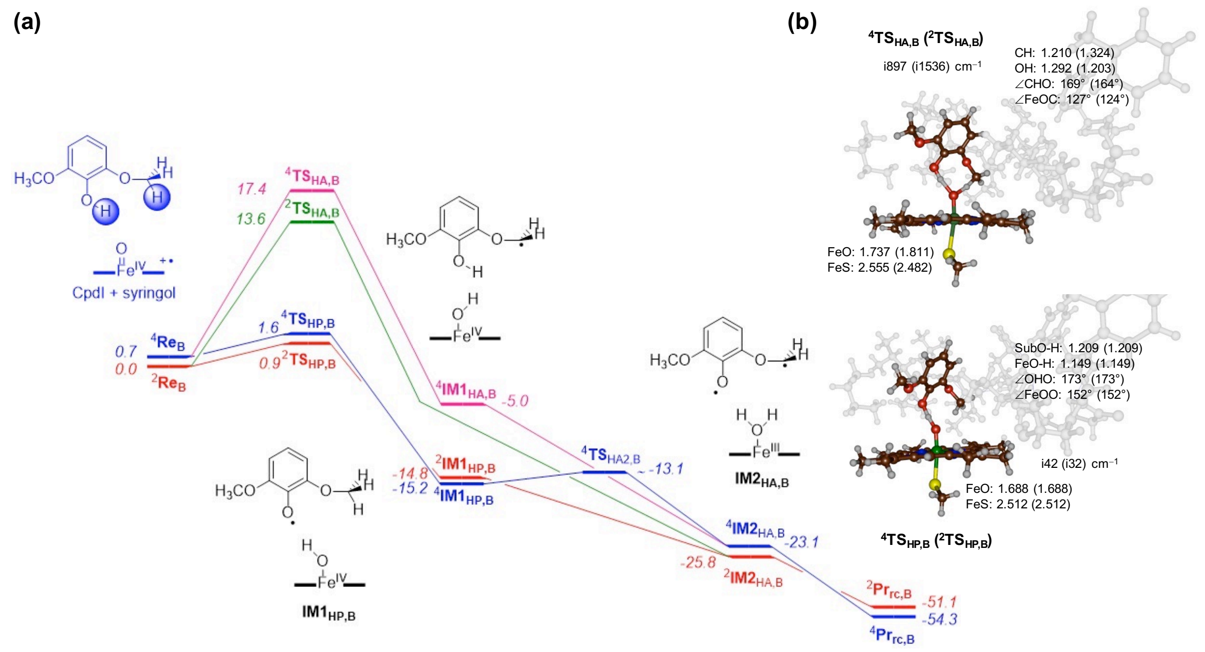


Figure 5.3. UB3LYP/BS2//UB3LYP/BS1 calculated potential energy surface for syringol activation by CpdI model **B** of GcoA. Energies contain ZPE and are given in kcal mol^{-1} relative to ^{2}**Re$_B$**. Optimized geometries of the transition states give bond lengths in angstroms, angles in degrees and the imaginary frequency in cm^{-1}.

Finally, we also tested phenol activation by our GcoA CpdI model. Phenolic hydrogen-atom abstraction is possible in substrate binding position **B** and happens through a very small transition state with an imaginary frequency of only i42 (i32) cm^{-1} for ^{4}**TS$_{HP,B}$** (^{2}**TS$_{HP,B}$**), respectively. Analysis of the imaginary frequency gives a clear hydrogen-transfer mode although part of the substrate displaces as well. Similar to the aliphatic transition states, the O—H—O angle around the transferring hydrogen atom is almost linear: 173° in both spin states. The transition states are relatively central with slightly shorter FeO–H distances than phenolic H–O distances.

After the phenolic hydrogen-atom abstraction, the system relaxes to a radical intermediate (^{4,2}**IM1**$_{HP,B}$). These structures are much lower in energy than reactants by 14.8 (15.2) kcal mol^{-1} in the doublet (quartet) spin states. As such, these radical intermediates will be quickly formed. Since, no OH rebound is possible after phenolic hydrogen-atom abstraction, we explored a second hydrogen-atom abstraction from the methoxy CH_3 group. On the doublet spin state the barrier (^{2}**TS**$_{HA2,B}$) is negligible and the system transfers to the iron(III)-water complex (^{2}**IM2**$_{HA,B}$) with an exothermicity of more than 10 kcal mol^{-1}. A geometry scan for the quartet spin pathway identified a small barrier (^{4}**TS**$_{HA2,B}$) about 2.1 kcal mol^{-1} above

^{4}**IM1**$_{\mathrm{HP,B}}$ shown in Figure 5.4. As this is a low-barrier transition state that is located close to a local minimum our TS search failed to converge and the optimization fell back to the intermediate. As discussed above the iron(III)-water complexes ^{4,2}**IM2**$_{\mathrm{HA,B}}$ quickly close the acetal ring to form the ^{4,2}**Pr**$_{\mathrm{rc,B}}$ products without much of a barrier.

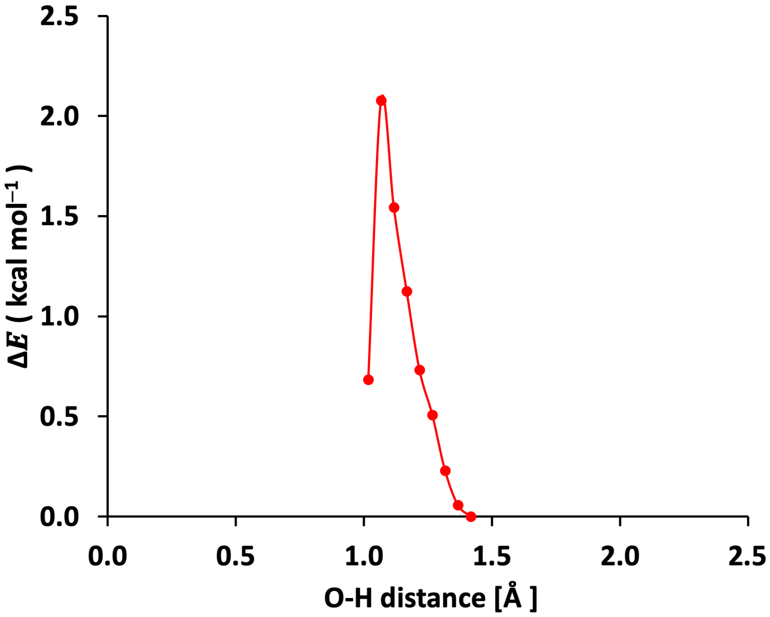
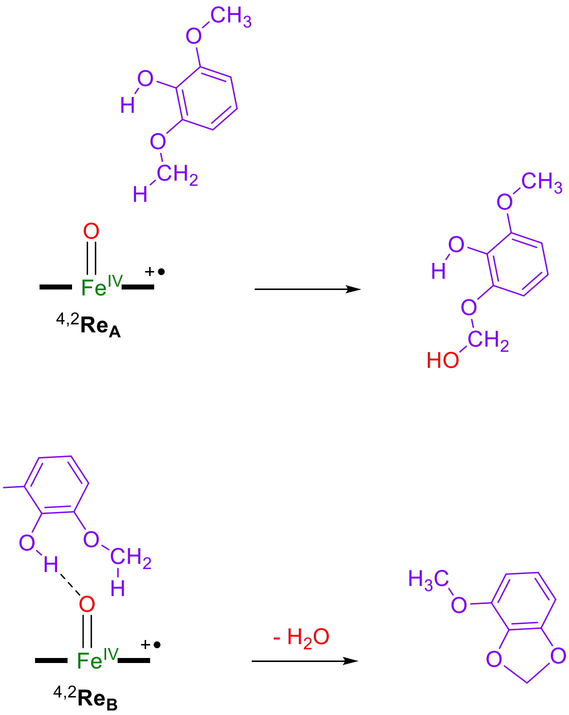


Figure 5.4. UB3LYP/BS1 calculated constraint geometry scan for hydrogen-atom abstraction from ^{4}IM1$_{\mathrm{HP}}$ as calculated in Gaussian. An estimate for the barrier height from the scan is 2.1 kcal mol^{-1} above the energy of ^{4}IM1$_{\mathrm{HP,B}}$.

Overall, the mechanism for substrate activation in binding position **B** shows that sequentially two hydrogen atoms are abstracted from the substrate, the first one from the phenol O–H group and second one from a methoxy C–H group. The first reaction barrier is rate-determining while the subsequent barriers were too small to be fully characterized. Therefore, the acetal formation will be a highly efficient and fast process and much faster than the substrate-binding and product-release steps in the protein. Consequently, the results on syringol activation by a large GcoA model shows that different products are predicted from substrate binding positions **A** and **B**.

5.4 Discussion

The work described here is focused on syringol activation by a lignin activating P450 isozyme, namely GcoA. A large active site model of 302 atoms was considered that contains the heme active site and a large part of the substrate binding pocket with the substrate in two specific binding poses. The mechanism of substrate activation leading to hemiacetal and acetal products for the two binding poses was investigated. In substrate binding position **A** a rate-determining hydrogen-atom abstraction leads to methoxy hydroxylation efficiently, see Scheme 5.4. By contrast, in binding pose **B** the weak phenolic O–H bond points toward the heme and therefore can be abstracted by CpdI easily. In particular, the phenolic O–H group has a much lower barrier for hydrogen-atom abstraction than the aliphatic C–H abstraction from the methoxy group. However, this hydrogen atom can be abstracted in a subsequent step and lead to ring-closure to form acetal products. As such, the two substrate binding poses (Scheme 5.4) lead to different product distributions for syringol activation by GcoA. To understand the key factors that determine substrate activation, we analyzed the structures in more detail.



Scheme 5.4. Products obtained for substrate activation by GcoA through substrate orientation **A** and **B**.

Firstly, we calculated the various C–H and O–H bond dissociation energies (BDE1s) of syringol substrate (SubH) using Equation 5.1. The BDE1 values were estimated from the difference in energy of the individually calculated species in the reaction, i.e. we calculated the substrate, a hydrogen-atom and the substrate with one hydrogen-atom removed from either the phenol or methoxy groups (Sub$^{\bullet}$). The reaction energy for Equation 5.1 was then evaluated for hydrogen-atom abstraction from the phenol group of syringol ($\mathrm{BDE1_{O–H}}$) and for hydrogen-atom abstraction from the C–H group of the methoxy unit ($\mathrm{BDE1_{C–H}}$), see Figure 5.5. At UB3LYP/6-311++G** level of theory we find a $\mathrm{BDE1_{O–H}}$ = 76.7 kcal mol^{-1} and a $\mathrm{BDE1_{C–H}}$ = 93.4 kcal mol^{-1}, see Figure 5.5. Therefore, the phenolic O–H bond is considerably weaker than the aliphatic C–H bond of the substrate and it should be easier to abstract the phenolic hydrogen atom than the methoxy hydrogen atom. Indeed, the potential energy landscape in Figure 5.3 for pathway **B** shows that the phenolic hydrogen-atom abstraction has a much lower barrier than the one for aliphatic C–H abstraction in line with the large differences in BDE values.

$$\mathrm{Sub - H \rightarrow Sub^{\blacksquare} + H^{\blacksquare} + BDE_{Sub-H}} \qquad (5.1)$$

For a small model complex of CpdI representing $[\mathrm{Fe^{IV}(O)(Por^{+\bullet})SCH_3}]$, Por = porphyrin without side chains, we calculated the $\mathrm{BDE_{CpdII}}$ as the energy difference between $[\mathrm{Fe^{IV}(OH)(Por)(SCH_3)}]$ and CpdI and an isolated hydrogen atom and obtained a value of $\mathrm{BDE_{CpdII}}$ = 87.4 kcal mol^{-1}. In previous work a slightly smaller model with thiolate rather than $\mathrm{SCH_3^-}$ as axial ligand was used and gave a $\mathrm{BDE_{CpdII}}$ = 88.9 kcal mol^{-1} in the gas-phase [70,74]. Therefore, the change in axial ligand from thiolate to methylmercaptate has little effect on the BDE values. The energy differences in Figure 5.2 and 5.3 between the reactant complexes and $\mathbf{IM1}_{\mathrm{HA}}$ and $\mathbf{IM1}_{\mathrm{HP}}$ should be equal to the difference in energy between the C–H/O–H bonds broken and formed. The difference in energy between $\mathrm{BDE1_{O–H}}$ and $\mathrm{BDE_{CpdII}}$ is –10.7 kcal mol^{-1}, which is close in energy to the exothermicity to form $\mathbf{IM1}_{\mathrm{HP,B}}$ from reactants. Hence, the potential energy landscape in Figure 5.3 follows for the hydrogen atom abstraction follows the strengths of the C–H and O–H bonds broken and formed.

Similar to the phenol hydrogen atom abstraction pathway, we evaluated the difference in the bond strengths that are formed and broken for the methoxy hydrogen atom abstraction. Thus, the difference in energy between $\mathrm{BDE1_{C–H}}$ and $\mathrm{BDE_{CpdII}}$ is +6.0 kcal mol^{-1}. The energy difference between $^{2}\mathbf{Re}_{\mathrm{A}}$ and $^{2}\mathbf{IM1}_{\mathrm{HA,A}}$ is –1.9 kcal mol^{-1} (Figure 5.2), while the one between $^{2}\mathbf{Re}_{\mathrm{B}}$ and $^{2}\mathbf{IM1}_{\mathrm{HA,B}}$ is –5.7 kcal mol^{-1} (Figure 5.3). These driving forces are, therefore, somewhat lower in energy than what would have been predicted based on the difference in bond dissociation energy of the bonds that are broken and formed. To understand these differences better, we display in Figure 5.5 the optimized geometries of ^{4,2}**IM1**$_{\mathrm{HA,A}}$ and ^{4,2}**IM1**$_{\mathrm{HA,B}}$. The **IM1**$_{\mathrm{HA,A}}$ radical intermediates have the substrate hydrogen bonded to the peptide carbonyl of Val$_{241}$, the amide of Gly$_{245}$, while the methoxy group hydrogen bonds to the carboxylate of Glu$_{249}$. All of these interactions are well over 1.7 Å in length and hence there is only a small stabilization effect with respect to the thermodynamic bond energy differences due to hydrogen bonding interactions. By contrast, in the **IM1**$_{\mathrm{HA,B}}$ structures the phenol OH group is close to the iron(IV)-hydroxo at a distance of 1.418 (1.591) Å in the quartet (doublet) spin state. This is a strong hydrogen bonding interaction that will stabilize these radical intermediates considerably. Indeed, the stabilization energy is much more exothermic than the difference in bond strength of the bond that is broken and formed implicates. The hydrogen bonding interactions of the protein and the iron-hydroxo species, therefore, stabilize the radical intermediates and make the reaction more exothermic. Therefore, the first hydrogen-atom abstraction barriers follow the thermodynamics of the individual hydrogen-atom abstraction processes.

$$[\mathrm{Fe(Por)(SCH_3)]O-H} \rightarrow \mathrm{CpdI} + \mathrm{H}^{\blacksquare} + \mathrm{BDE}_{\mathrm{CpdII}} \tag{5.2}$$

Subsequently, we calculated the phenolic O–H bond strength and the methoxy aliphatic C–H bond strength from the radicals as BDE2$_{\mathrm{O–H}}$ and BDE2$_{\mathrm{C–H}}$, also shown in Figure 5.4. Values of BDE2$_{\mathrm{O–H}}$ = 79.1 kcal mol^{-1} and BDE2$_{\mathrm{C–H}}$ = 95.8 kcal mol^{-1} were calculated. This implies that the second hydrogen-atom abstraction takes a similar energy to break as the first hydrogen-atom abstraction. To predict the reaction energy for the second hydrogen-atom abstraction step, we calculated the BDE2$_{\mathrm{water}}$ for the conversion of an iron(III)-water(heme) complex into an iron(IV)-hydroxo(heme) and a hydrogen atom and obtained a value of 87.5 kcal mol^{-1}.

Based on the difference in energy between BDE2$_{\mathrm{O–H}}$ and BDE2$_{\mathrm{water}}$, the aliphatic hydrogen-atom abstraction should be followed by an exothermic second hydrogen-atom abstraction from the phenol group by –8.4 kcal mol^{-1} while initial phenol activation should be followed by an endothermic aliphatic hydrogen-atom abstraction with energy of 8.3 kcal mol^{-1} (difference in energy between BDE2$_{\mathrm{C–H}}$ and BDE2$_{\mathrm{water}}$). As a matter of fact, the reaction energy from ^{4}**IM1**$_{\mathrm{HA,B}}$ to ^{4}**IM2**$_{\mathrm{HA}}$ is highly exothermic (by 18.1 kcal mol^{-1}) in line with the difference in energy of the BDE values. Moreover, it explains why the second barrier has a negligible hydrogen-atom abstraction barrier. The energy difference between ^{2}**IM1**$_{HP}$ and ^{2}**IM2**$_{HA}$ is –11.0 kcal mol^{-1}, which is somewhat lower than the energy predicted based on BDE values and shows that the product is highly stabilized through local hydrogen bonds.

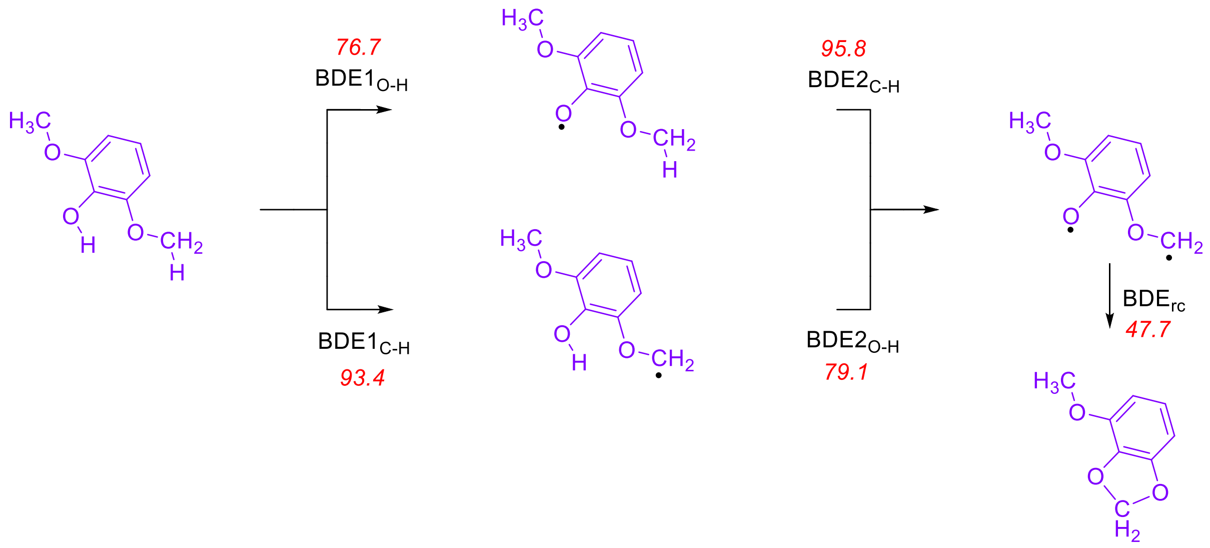
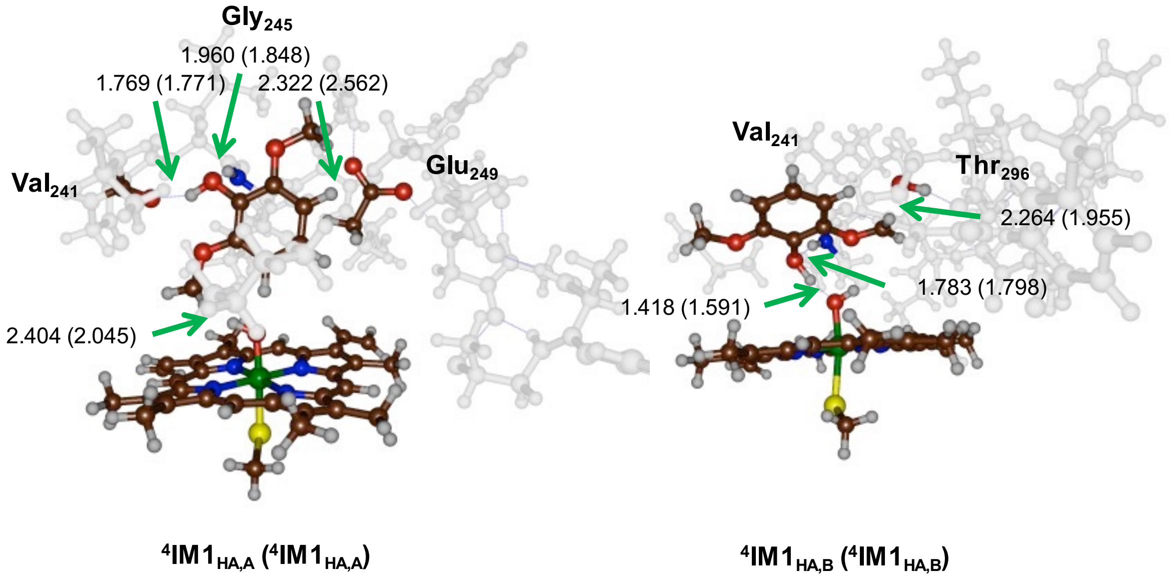


Figure 5.5. UB3LYP/6-311++G** calculated bond dissociation and formation energies (kcal mol^{-1}) including ZPE corrections in the substrate syringol.



^{4}IM1$_{HA,A}$ (^{4}IM1$_{HA,A}$)

^{4}IM1$_{HA,B}$ (^{4}IM1$_{HA,B}$)

Figure 5.6. Optimized geometries of ^{4,2}**IM1**$_{HA,A}$ and ^{4,2}**IM1**$_{HA,B}$ as obtained at UB3LYP/BS1 with bond lengths in angstroms.

The biradical system was calculated in the triplet and open shell singlet spin states and the energy to close the ring to form acetal products gave a BDE$_{rc}$ of 47.7 kcal mol^{-1}. However, several hydrogen bonds are lost between the bound water molecule and the product complex upon ring-closure, so that the stabilization energy for the ring-closure is much lower than this. Therefore, an energy difference of ΔE+ZPE = 28.5 kcal mol^{-1} is calculated between $^{4}\mathbf{IM2}_{HA,B}$ and $^{4}\mathbf{Pr}_{rc,B}$ in line with the energy difference to close the acetal ring and the cost of breaking several short hydrogen bonds.

From the calculations it is clear that when the phenol group of substrate is accessible by CpdI, a hydrogen-atom abstraction from the O–H group will take place as its O–H bond is much weaker than aliphatic C–H bonds such as those of the methoxy group. If hemiacetal is the preferred product, however, the substrate should be positioned with the phenol group pointing away from CpdI, while at the same time the methoxy group points to CpdI.

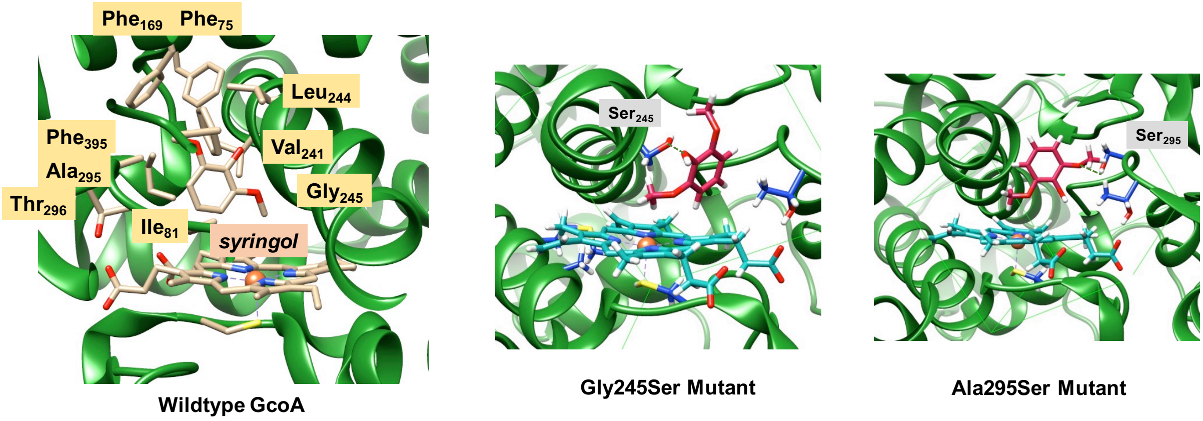


Figure 5.7. Extract of the substrate-bound GcoA structure as taken from the 5OMU pdb file and two predicted mutants, namely Gly245Ser and Ala295Ser that position substrate differently.

In order to gain insight into probable product distributions based on substrate positioning in the enzyme, we analyzed the substrate binding pocket in more detail. Figure 5.7 displays the active site structure and key residues in the substrate binding pocket of GcoA. As can be seen, the substrate binding pocket is aligned with mostly aromatic and aliphatic amino acid residues including Phe$_{75}$, Ile$_{81}$, Phe$_{169}$, Val$_{241}$, Leu$_{244}$, Ala$_{295}$ and Phe$_{395}$. Therefore, few polar interactions are available to position the substrate in a specific orientation. Probably, substrate positioning in GcoA enzymes is not important as long as the lignin degradation pathways proceed and the selectivity of the enzyme seems limited.

To make GcoA more substrate and regioselective, we decided to create two in silico mutants, whereby an additional hydroxyl group in the substrate binding pocket is included that can position the substrate better and tighter. To this end, we took the 5OMU pdb file, removed the substrate and created the Gly245Ser and Ala295Ser mutants. Subsequently, using Autodock [75] syringol was docked into the substrate binding pocket. The lowest energy syringol bound conformation of the two mutants is shown in Figure 5.7. As can be seen, the Gly245Ser mutant gives a hydrogen bond between the Ser_{245} and phenolic O–H group of the substrate. This positions the methoxy group close to the heme and the phenol group away from the heme. We predict that the Gly245Ser mutant, therefore, will give predominantly methoxy hydroxylation or *O*-demethylation products. In contrast, the Ala295Ser mutant has the substrate bound with a hydrogen bond between Ser_{295} and the oxygen atom of one of the methoxy groups. This structure has the other methoxy group and the phenol group both pointing towards the heme and are likely positioned to convert substrate into acetal products.

The GcoA wildtype structure, however, has a tight and closed substrate binding pocket where the substrate is locked in by bulky aromatic residues such as those of Phe_{75}, Phe_{169} and Phe_{395}. Therefore, GcoA should only be able to bind relatively small substrates such as syringol. Actually, experimental studies showed only activity with lignin monomers, which indicates that the substrate binding pocket is closed and only accessible to small substrates. In particular, slow guaiacol and even slower syringol activation by GcoA was observed [20]. Furthermore, mutations of Phe_{169} by Ala enabled syringol activation with better turnover numbers, but only *O*-demethylation products were obtained. Clearly, substrate binding in wildtype GcoA positions the substrate with the phenol group away from the heme center and drives the reaction via pathway **A** to give predominantly methoxy hydroxylation followed by deformylation. Based on the structural analysis in this work, it is clear that acetal products from syringol activation in GcoA will require further mutations to position the substrate better and enhance its selectivity. This also could be done by opening the substrate binding pocket so that longer lignin molecules or components can be inserted into the heme active site that will enable its oxidation.

5.5 Conclusion

In this work a computational study is presented on lignin activation by the cytochrome P450 isozyme GcoA. We tested several substrate-binding orientations and spin-state structures. The work shows that syringol activation should predominantly lead to acetal products through two sequential hydrogen atom abstraction steps from the phenol and methoxy groups followed by radical coupling to close the acetal ring. We then analyzed P450 structures and give suggestions on how to engineer the P450 and give higher contribution of hemiacetal and acetal products. Overall, the work shows that the P450s are efficient oxidants and should be able to activate and degrade lignin molecules easily. The fact that this does not happen regularly in nature reflects the point that the substrate binding pocket is accessible to small substrates only and it will require some protein engineering to make it bind lignin strands.

5.6 Reference

[1] Ahmad, M., Roberts, J.N., Hardiman, E.M., Singh, R., Eltis, L.D. and Bugg, T.D.H., Identification of DypB from rhodococcus jostii RHA1 as a lignin peroxidase. *Biochemistry* **2011**, *50*, 5096–5107.

[2] Falade, A.O., Nwodo, U.U., Iweriebor, B.C., Green, E., Mabinya, L.V. and Okoh, A.I., Lignin peroxidase functionalities and prospective applications. *Microbiol. Open* **2017**, *6*, e00394.

[3] Brown, M.E. and Chang, M.C.Y., Exploring bacterial lignin degradation. *Curr. Opin. Chem. Biol.* **2014**, *19*, 1–7.

[4] Li, X. and Zheng, Y., Biotransformation of lignin: Mechanisms, applications and future work. *Biotechnol. Pro.* **2020**, *36*, e2922.

[5] Granja-Travez, R.S., Persinoti, G.F., Squina, F.M. and Bugg, T.D.H., Functional genomic analysis of bacterial lignin degraders: diversity in mechanisms of lignin oxidation and metabolism. *Appl. Microbiol. Biotechnol.* **2020**, *104*, 3305–3320.

[6] Timofeevski, S.L., Nie, G., Reading, N.S. and Aust, S.D., Substrate specificity of lignin peroxidase and a S168W variant of manganese peroxidase. *Arch. Biochem. Biophys.* **2000**, *373*, 147–153.

[7] Ichinose, H., Cytochrome P450 of wood-rotting basidiomycetes and biotechnological applications. *Biotechnol. Appl. Biochem*. **2013**, *1*, 71–81.

[8] García-Hidalgo, J., Ravi, K., Kuré, L.L., Lidén, G. and Gorwa-Grauslund, M., Vanillin production in *pseudomonas*: Whole-genome sequencing of *pseudomonas* sp. strain 9.1 and reannotation of *pseudomonas putida* CalA as a vanillin reductase. *AMB Expr*. **2019**, *9*, 34–44.

[9] Park, H., Park, G., Jeon, W., Ahn, J.O., Yang, Y.H. and Choi, K.Y., Whole-cell biocatalysis using cytochrome P450 monooxygenases for biotransformation of sustainable bioresources (fatty acids, fatty alkanes, and aromatic amino acids). *Biotechnol. Adv*. **2020**, *40*, 107504.

[10] Sono, M., Roach, M.P., Coulter, E.D. and Dawson, J.H., Heme-containing oxygenases. *Chem. Rev*. **1996**, *96*, 2841–2888.

[11] Handbook of Porphyrin Science, (Eds.: K. M. Kadish, K. M. Smith, R. Guilard), World Scientific Publishing Co., New Jersey, **2010**.

[12] Iron-containing enzymes: Versatile catalysts of hydroxylation reaction in nature (Eds.: S.P. de Visser, D. Kumar), RSC Publishing, Cambridge (UK), **2011**.

[13] Ortiz de Montellano, P.R., Hydrocarbon hydroxylation by cytochrome P450 enzymes. *Chem. Rev*. **2010**, *110*, 932–948;

[14] Meunier, B., de Visser, S.P. Shaik, S., Mechanism of oxidation reactions catalyzed by cytochrome p450 enzymes. *Chem. Rev*. **2004**, *104*, 3947–3980.

[15] Huang, X. and Groves, J.T., Oxygen activation and radical transformations in heme proteins and metalloporphyrins. *Chem. Rev*. **2018**, *118*, 2491-2553.

[16] Girvan, H.M. and Munro, A.W., Applications of microbial cytochrome P450 enzymes in biotechnology and synthetic biology. *Curr. Opin. Chem. Biol.* **2016**, *31*, 136-45.

[17] Denisov, I.G., Makris, T.M., Sligar, S.G. and Schlichting, I., Structure and chemistry of cytochrome P450. *Chem. Rev.* **2005,** *105*, 2253-77.

[18] Green, M.T., C-H bond activation in heme proteins: The role of thiolate ligation in cytochrome P450. *Curr. Opin. Chem. Biol*. **2009**, *13*, 84–88.

[19] Nakano, C., Horinouchi, S. and Ohnishi, Y., Characterization of a novel sesquiterpene cyclase involved in (+)-caryolan-1-ol biosynthesis in *Streptomyces griseus. J. Biol. Chem.* **2011**, *286*, 27980–27987.

[20] Machovina, M.M., Mallinson, S.J.B., Knott, B.C., Meyers, A.W., Garcia-Borràs, M., Bu, L., Gado, J.E., Oliver, A., Schmidt, G.P., Hinchen, D.J., Crowley, M.F., Johnson, C.W., Neidle, E.L., Payne, C.M., Houk, K.N., Beckham, G. T., McGeehan, J.E. and DuBois, J.L., Enabling microbial syringol conversion through structure-guided protein engineering. *Proc. Natl. Acad. Sci. USA* **2019**,*116*, 13970-13976.

[21] Mallinson, S.J.B., Machovina, M.M., Silveira, R.L., Garcia-Borràs, M., Gallup, N., Johnson, C.W., Allen, M.D., Skaf, M.S., Crowley, M.F., Neidle, E.L., Houk, K.N., Beckham, G.T.; DuBois, J.L. and McGeehan, J.E., A promiscuous cytochrome P450 aromatic O-demethylase for lignin bioconversion. *Nature Commun.* **2018,** *9*, 2487.

[22] Timmins, A., Saint-André, M. and de Visser, S.P., Understanding how prolyl-4-hydroxylase structure steers a ferryl oxidant toward scission of a strong C–H bond. *J. Am. Chem. Soc*. **2017,** *139*, 9855-9866.

[23] Pickl, M., Kurakin, S., Cantú Reinhard, F.G., Schmid, P., Pöcheim, A., Winkler, C. K., Kroutil, W., de Visser, S.P. and Faber, K., Mechanistic studies of fatty acid activation by CYP152 peroxygenases reveal unexpected desaturase activity. *ACS Catal.* **2019**, *9*, 565-577.

[24] de Visser, S.P., Second-coordination sphere effects on selectivity and specificity of heme and nonheme iron enzymes. *Chem. Eur. J.* **2020**, *26*, 5308-5327.

[25] Kramlinger, V.M., Rojas, A.M., Kanamori, T. and Guengerich, F.P., Introduction: Metals in biology: α-ketoglutarate/iron-dependent dioxygenases. *J. Biol. Chem*. **2015**, *290*, 20200–20210.

[26] Podgorski, M.N., Coleman, T., Chao, R.R., De Voss, J.J., Bruning, J.B. and Bell, S.G., Investigation of the requirements for efficient and selective cytochrome P450 monooxygenase catalysis across different reactions. *J. Inorg. Biochem*. **2020**, *203*, 110913.

[27] Taxak, N., Patel, B. and Bharatam, P.V., Carbene generation by cytochromes and electronic structure of heme-iron-porphyrin-carbene complex: A quantum chemical study. *Inorg. Chem.* **2013,** *52*, 5097-5109.

[28] Schyman, P., Lai, W., Chen, H., Wang, Y. and Shaik, S., The directive of the protein: How does cytochrome P450 select the mechanism of dopamine formation? *J. Am. Chem. Soc.* **2011,** *133*, 7977-7984.

[29] Oláh, J., Mulholland, A.J. and Harvey, J.N., Understanding the determinants of selectivity in drug metabolism through modeling of dextromethorphan oxidation by cytochrome P450. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 6050.

[30] Rydberg, P., Ryde, U. and Olsen, L., Sulfoxide, sulfur, and nitrogen oxidation and dealkylation by cytochrome P450. *J. Chem. Theory Comput*. **2008**, *4*, 1369–1377.

[31] Li, D., Wang, Y., Yang, C. and Han, K., Theoretical study of N-dealkylation of N-cyclopropyl-N-methylaniline catalyzed by cytochrome P450: insight into the origin of the regioselectivity. *Dalton Trans.* **2009**, 38, 291-297.

[32] Ogliaro, F., Harris, N., Cohen, S., Filatov, M., de Visser, S.P. and Shaik, S., A model “rebound” mechanism of hydroxylation by cytochrome P450: Stepwise and effectively concerted pathways, and their reactivity patterns. *J. Am. Chem. Soc.* **2000,** *122*, 8977-8989.

[33] Kamachi, T., Shestakov, A.F. Yoshizawa, K., How heme metabolism occurs in heme oxygenase: Computational study of oxygen-donation ability of the oxo and hydroperoxo species. *J. Am. Chem. Soc.* **2004,** *126*, 3672-3673.

[34] Olsen, L., Rydberg, P., Rod, T.H. and Ryde, U., Prediction of activation energies for hydrogen abstraction by cytochrome P450. *J. Med. Chem.* **2006**, *49*, 6489-6499.

[35] Kaczmarek, M.A., Malhotra, A., Balan, G.A., Timmins, A. and de Visser, S.P., Nitrogen reduction to ammonia on a biomimetic mononuclear iron centre: Insights into the nitrogenase enzyme. *Chemistry.* **2018,** *24*, 5293-5302.

[36] Groves, J.T., Avaria-Neisser, G.E., Fish, K.M., Imachi, M. and Kuczkowski, R.L., Hydrogen-deuterium exchange during propylene epoxidation by cytochrome P-450. *J. Am. Chem. Soc.* **1986,** *108*, 3837-3838.

[37] Rittle, J. and Green, M.T., Cytochrome P450 compound I: Capture, characterization, and C-H bond activation kinetics. *Science* **2010,** *330*, 933-937.

[38] Takahashi, A., Kurahashi, T. and Fujii, H., Redox potentials of oxoiron(IV) porphyrin π-cation radical complexes: Participation of electron transfer process in oxygenation reactions. *Inorg. Chem.* **2011,** *50*, 6922-6928.

[39] 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.,et al., Fox, Gaussian, Inc., Wallingford CT, 2010.

[40] Becke, A.D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988,** *38*, 3098-3100.

[41] Lee, C., Yang, W. and Parr, R.G., Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev.B* **1988**, *37*, 785-789.

[42] Hay, P.J. and Wadt, W.R., Ab initio effective core potentials for molecular calculations. potentials for the transition metal atoms Sc to Hg. *J. Chem. Phys.* **1985**, *82*, 270-283.

[43] Hehre, W.J., Ditchfield, R. and Pople, J.A., Self—consistent molecular orbital methods. XII. Further extensions of gaussian—type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972,** *56*, 2257-2261.

[44] Tomasi, J., Mennucci, B. and Cammi, R., Quantum mechanical continuum solvation models. *Chem. Rev.* **2005,** *105*, 2999-3093.

[45] Cantú Reinhard, F.G., Faponle, A.S., de Visser, S.P., Substrate sulfoxidation by an iron(IV)-oxo complex: Benchmarking computationally calculated barrier heights to experiment. *J. Phys. Chem. A* **2016,** *120*, 9805-9814.

[46] Cheaib, K., Mubarak, M.Q.E., Sénéchal-David, K., Herrero, C., Guillot, R., Clémancey, M., Latour, J.M., de Visser, S.P., Mahy, J.P., Banse, F. and Avenier, F., Selective formation of an Fe(IV)-O or an Fe(III) OOH intermediate from iron(II) and H(2) O(2):Controlled heterolytic versus homolytic oxygen-oxygen bond cleavage by the second coordination sphere. *Angew. Chem. Int. Ed.* **2019**, *58*, 854–858.

[47] Barman, P., Cantú Reinhard, F.G., Bagha, U.K., Kumar, D., Sastri, C.V. and de Visser, S.P., Hydrogen by deuterium substitution in an aldehyde tunes the regioselectivity by a nonheme manganese(III)–peroxo complex. *Angew. Chem. Int. Ed.* **2019**, *58*, 10639–10643.

[48] Quesne, M.G., Borowski, T. and de Visser, S.P., Quantum mechanics/molecular mechanics modeling of enzymatic processes: Caveats and breakthroughs. *Chem. Eur. J.* **2016**, *22*, 2562-2581.

[49] Ghafoor, S., Mansha, A. and de Visser, S. P., Selective hydrogen atom abstraction from dihydroflavonol by a nonheme iron center is the key step in the enzymatic flavonol synthesis and avoids by-products. *J. Am. Chem. Soc.* **2019,** *141*, 20278-20292.

[50] Lin, Y.T., Stańczak, A., Manchev, Y., Straganz, G.D. and de Visser, S.P., Can a mononuclear iron(III)-superoxo active site catalyze the decarboxylation of dodecanoic acid in undA to produce biofuels? *Chem. Eur. J.* **2020,** *26*, 2233-2242.

[51] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., The Protein Data Bank. *Nucl. Acids. Res.* **2000,** *28*, 235-242.

[52] Ghosh, A., First-principles quantum chemical studies of porphyrins. *Acc. Chem. Res.* **1998,** *31*, 189-198.

[53] Green, M.T., Evidence for sulfur-based radicals in thiolate compound I intermediates. *J. Am. Chem. Soc.* **1999**, *121*, 7939-7940.

[54] de Visser, S.P., Ogliaro, F., Sharma, P.K. and Shaik, S., What factors affect the regioselectivity of oxidation by cytochrome P450? A DFT study of allylic hydroxylation and double bond epoxidation in a model reaction. *J. Am. Chem. Soc.* **2002,** *124*, 11809-11826.

[55] Bathelt, C.M., Zurek, J., Mulholland, A.J. and Harvey, J.N., Electronic structure of compound I in human isoforms of cytochrome P450 from QM/MM modeling. *J. Am. Chem. Soc.* **2005,** *127*, 12900-12908.

[56] Radoń, M., Broclawik, E. and Pierloot, K., DFT and ab Initio study of iron-oxo porphyrins: May they have a low-lying iron(V)-oxo electromer? *J. Chem. Theory Comput.* **2011,** *7*, 898-908.

[57] Quesne, M.G., Senthilnathan, D., Singh, D., Kumar, D., Maldivi, P., Sorokin, A. B. and de Visser, S.P., Origin of the enhanced reactivity of μ-nitrido-bridged diiron(IV)-oxo porphyrinoid complexes over cytochrome P450 compound I. *ACS Catal.* **2016,** *6*, 2230-2243.

[58] Shaik, S., Kumar, D., de Visser, S.P., Altun, A. and Thiel, W., Theoretical perspective on the structure and mechanism of cytochrome P450 enzymes. *Chem. Rev.* **2005,** *105*, 2279-328.

[59] Porro, C.S., Sutcliffe, M.J. and de Visser, S.P., Quantum mechanics/molecular mechanics studies on the sulfoxidation of dimethyl sulfide by compound I and compound 0 of cytochrome P450: Which is the better oxidant? *J. Phys. Chem. A* **2009,** *113*, 11635-11642.

[60] Cantú Reinhard, F.G., Lin, Y.T., Stańczak, A. and de Visser, S.P., Bioengineering of cytochrome P450 OleTJE: How does substrate positioning affect the product distributions? *Molecules* **2020,** *25*, 2675.

[61] Shaik, S., Kumar, D. and de Visser, S.P., A valence bond modeling of trends in hydrogen abstraction barriers and transition states of hydroxylation reactions catalyzed by cytochrome P450 enzymes. *J. Am. Chem. Soc.* **2008,** *130*, 10128-10140.

[62] Shaik, S., Cohen, S., Wang, Y., Chen, H., Kumar, D. and Thiel, W., P450 enzymes: Their structure, reactivity, and selectivity—modeled by QM/MM calculations. *Chem. Rev.* **2010,** *110*, 949-1017.

[63] Latifi, R., Bagherzadeh, M. and de Visser, S.P., Origin of the correlation of the rate constant of substrate hydroxylation by nonheme iron(IV)–oxo complexes with the bond-dissociation energy of the C-H bond of the substrate. *Chem. Eur. J.* **2009,** *15*, 6651–6662.

[64] de Visser, S.P., Trends in substrate hydroxylation reactions by heme and nonheme iron(IV)-oxo oxidants give correlations between intrinsic properties of the oxidant with barrier height. *J. Am. Chem. Soc.* **2010,** *132*, 1087-1097.

[65] Ji, L., Faponle, A.S., Quesne, M.G., Sainna, M.A., Zhang, J., Franke, A., Kumar, D., van Eldik, R., Liu, W. and de Visser, S.P., Drug metabolism by cytochrome P450 enzymes: What distinguishes the pathways leading to substrate hydroxylation over desaturation? *Chem. Eur. J.* **2015,** *21*, 9083-9092.

[66] Barman, P., Upadhyay, P., Faponle, A.S., Kumar, J., Nag, S.S., Kumar, D., Sastri, C.V. and de Visser, S.P., Deformylation reaction by a nonheme manganese(III)–peroxo complex via initial hydrogen-atom abstraction. *Angew. Chem. Int. Ed.* **2016**, *55*, 11091–11095.

[67] Cantú Reinhard, F.G., Barman, P., Mukherjee, G., Kumar, J., Kumar, D., Sastri, C.V. and de Visser, S.P., Keto-enol tautomerization triggers an electrophilic aldehyde deformylation reaction by a nonheme manganese(III)-peroxo complex. *J. Am. Chem. Soc.* **2017,** *139*, 18328-18338.

[68] Shaik, S., Cohen, S., de Visser, S.P., Sharma, P.K., Kumar, D., Kozuch, S., Ogliaro, F. and Danovich, D., The "rebound controversy": An overview and theoretical modeling of the rebound step in C−H hydroxylation by cytochrome P450. *Eur. J. Inorg. Chem.* **2004,** *2004*, 207-226.

[69] Faponle, A.S., Quesne, M.G. and de Visser, S.P., Origin of the regioselective fatty-acid hydroxylation versus decarboxylation by a cytochrome P450 peroxygenase: What drives the reaction to biofuel production? *Chem. Eur. J.* **2016,** *22*, 5478-5483.

[70] de Visser, S.P. and Tan, L.S., Hydroxylation reactions catalyzed by cytochrome P450 enzymes. *J. Am. Chem. Soc.* **2008,** *130*, 10128-10140.

[71] Ali, H.S., Henchman, R.H. and de Visser, S.P., Cross-linking of aromatic phenolate groups by cytochrome P450 enzymes: A model for the biosynthesis of vancomycin by OxyB. *Org. Biomol. Chem.* **2020,** *18*, 4610-4618.

[72] Timmins, A., Fowler, N.J., Warwicker, J., Straganz, G.D. and de Visser, S.P., Does substrate positioning affect the selectivity and reactivity in the hectochlorin biosynthesis halogenase? *Front. Chem.* **2018,** *6,* 513.

[73] Latifi, R., Sainna, M.A., Rybak-Akimova, E.V. and de Visser, S.P., Does hydrogen-bonding donation to manganese(IV)–oxo and iron(IV)–oxo oxidants affect the oxygen-atom transfer ability? A computational study. *Chem. Eur. J.* **2013,** *19*, 4058-4068.

[74] Ogliaro, F., de Visser, S.P., Cohen, S., Kaneti, J. and Shaik, S., The experimentally elusive oxidant of cytochrome P450: A theoretical "trapping" defining more closely the "real" species. *Chem. Bio. Chem.* **2001,** *2*, 848-851.

[75] Trott, O. and Olson, A.J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. comput. Chem.* **2010,** *31,* 455-461.

Chapter 6. Energy-Entropy Method Using Multiscale Cell Correlation to Calculate Binding Free Energies in the SAMPL8 Host-Guest Challenge

PAPER 5

Hafiz Saqib Ali[1,2], Arghya Chakravorty, Jas Kalayan[1,2], Sam P. de Visser[1,3] and Richard H. Henchman[1,2]

[1] Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester M1 7DN, UK

[2] School of Chemistry, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

[3] Department of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

* Corresponding author:

E-mail: henchman@manchester.ac.uk

Conceptualization, supervision and project administration, R.H.H. investigation and funding acquisition, H.S.A.; methodology and software, A.C., J.K. and R.H.H.; validation, visualization and writing—original draft preparation, H.S.A., and R.H.H.; formal analysis and writing—review and editing, H.S.A., A.C., J.K, S.P.V and R.H.H.

Published in *Journal of Computer-Aided Molecular Design*, May 2021, (accepted/in press).

Abstract

Free energy drives a wide range of molecular processes such as solvation, binding, chemical reactions and conformational change. Given the central importance of binding, a wide range of methods exist to calculate it, whether based on scoring functions, machine-learning, classical or electronic structure methods, alchemy, or explicit evaluation of energy and entropy. Here we present a new energy-entropy (EE) method to calculate the host-guest binding free energy directly from molecular dynamics (MD) simulation. Entropy is evaluated using Multiscale Cell Correlation (MCC) which uses force and torque covariance and contacts at two different length scales. The method is tested on a series of seven host-guest complexes in the SAMPL8 (Statistical Assessment of the Modeling of Proteins and Ligands) "Drugs of Abuse" Blind Challenge. The EE-MCC binding free energies are found to agree with experiment with an average error of 0.9 kcal mol^{-1}. MCC makes clear the origin of the entropy changes, showing that the large loss of positional, orientational, and to a lesser extent conformational entropy of each binding guest is compensated for by a gain in orientational entropy of water released to bulk, combined with smaller decreases in vibrational entropy of the host, guest and contacting water.

6.1 Introduction

The accurate prediction of binding between molecules in solution is a key question in theoretical and computational chemistry. It has relevance to much of chemistry but also more broadly to fields such as biology, pharmacology, chemical engineering and environmental science. Under ambient conditions, binding is governed by the change in Gibbs free energy $\Delta G = -RT \ln K$ where RT is the gas constant times temperature and K is the equilibrium constant, which is the ratio of probability of the bound form relative to the unbound form at equilibrium for a given concentration of the molecules involved, typically 1 M.

Many methods have been developed to calculate binding free energy, which feature the typical trade-off of speed versus accuracy [1-4]. At the faster end are scoring functions which are parametrised to reproduce known binding data, being made ever more accurate by using larger data sets and machine-learning methods that resolve the optimal model form at the cost of providing molecular insight [5-10]. Simulation methods using classical potentials can determine the free energy difference from the relative probability of the bound and free states, whether this be by brute-force sampling, biased simulations such as metadynamics [11] or umbrella sampling [12,13], or alchemical methods such as free energy perturbation [14,15] or thermodynamic integration [16,17] which utilise shorter, unphysical binding paths by varying the molecules' interacting Hamiltonian rather than their positions. Combining these methods with more accurate electronic-structure methods is not yet achievable when simulating ensembles of solvated systems for multiple states along a path. However, they or regular force fields can be used in the energy-entropy (EE) class of methods which evaluate the free energy of the bound and free states separately and directly from the system energy and entropy and get the binding free energy from their difference. These are sometimes referred to as "end-point" methods but this is somewhat of a misnomer, given that no reference is required to a path or its end.

EE methods are more approximate and limited than other methods because calculating the entropy requires knowing the probability distribution of all quantum states of a system involving both solutes and solvent alike. This goes beyond the usual analyses of flexibility in MD simulations that typically look at distributions in only one or a few coordinates. The evaluation of a system's energy from the force-field Hamiltonian is much more straightforward, subject to getting converged values and to all the approximations inherent in the force field or electronic-structure method used. To make EE methods faster and more practical, they often employ an implicit-solvent model to give a solvation free energy [18], as is done in the widely used Molecular Mechanics/Poisson-Boltzmann Surface Area (MM/PBSA) method and its Generalised-Born variant (MM/GBSA) [19-22]. In addition to the approximations in the solvation model, such as the choice of dielectric constant, surface or surface tension parameters, their application to explicit-solvent simulations brings about an inconsistency between the Hamiltonian used for sampling and that used for free-energy evaluation. A frequent approximation is to apply normal mode analysis to a minimised configuration under the assumption of a Gaussian distribution [19,23], but this requires expensive matrix diagonalization for every minimum considered [24], and even then these minimized configurations are only approximately representative of thermalised ensembles. Consequently, the entropy contribution is often neglected [25], justified by the assumption that it is constant and therefore unimportant for relative binding calculations. A widely used method that does use thermalised ensembles is quasiharmonic analysis [26] based on coordinate covariance. Its assumption of a single Gaussian probability distribution permits a simple implementation, but this is known to over-estimate entropy [27,28]. Alternative ways to use ensembles beyond the approximation is to integrate Boltzmann factors over minima numerically [28,29], which is limited to a small number of minima, dihedral distributions [30], mutual information expansions [31] and the minimal spanning tree (MIST) variant [32], whose slowly converging nature limits their accuracy. For all non-Gaussian kinds of method, their classical formulation means they are limited to soft degrees of freedom, such as dihedrals or non-bonded interactions, and if applied to covalent bonds would give unphysically negative entropies, although one recent method includes dihedral correlations in a mutual-information manner supplemented by normal mode analysis [33,34]. Moreover, as mentioned earlier, many methods use a continuum treatment of solvent. Treating the solutes and solvent differently leads to formulations that allocate the ideal-gas translational and rotational entropy to the binding molecules, which obscures understanding the entropy loss of binding, with a larger proportion being assigned to the binding molecules and a corresponding entropy gain to the release of excluded-volume solvent [35]. Explicit solvent entropy has often been considered in binding, often to the exclusion of other entropic contributions and mostly in the context of inhomogeneous solvation theory [36-38]. Various other binding studies with combination terms have appeared such as molecular mechanics energy with the 3D Reference Interaction Site Model [39], continuum and explicit solvent [40], or inhomogeneous solvation theory with the loss of translational and rotational entropy [41] or with dihedral entropy [42].

To address the above-mentioned deficiencies of EE methods, we adapt the Multiscale Cell Correlation (MCC) method [43-45] to calculate the free energy of binding. MCC has been developed progressively, first in the context of cell theory for liquids [46,47] and solutions [45,48,49], later to account for correlations in flexible molecules [50], and most recently at multiple length scales [43,44,51]. A key feature of the theory is that it is applied to all molecules in the system equally, which makes it readily extendable, such as to large flexible molecules in solution. We calculate the free energy of the unbound and bound molecules in water from the energy and entropy in molecular dynamics (MD) simulations, building off earlier work addressing the change in molecular rigid-body translational and rotational entropy of binding [25,35]. We apply MCC to a series of host-guest complexes in the SAMPL8 "Drugs of Abuse" Blind Challenge (Statistical Assessment of the Modeling of Proteins and Ligands). Binding free energy has been a long-running quantity to calculate in the SAMPL Blind Challenges [52-55]. The SAMPL8 challenge involves the prediction of the binding free energies of seven drug molecules to the drug-carrier molecule cucurbit[8]uril (CB8), which are illustrated in Figure 6.1.

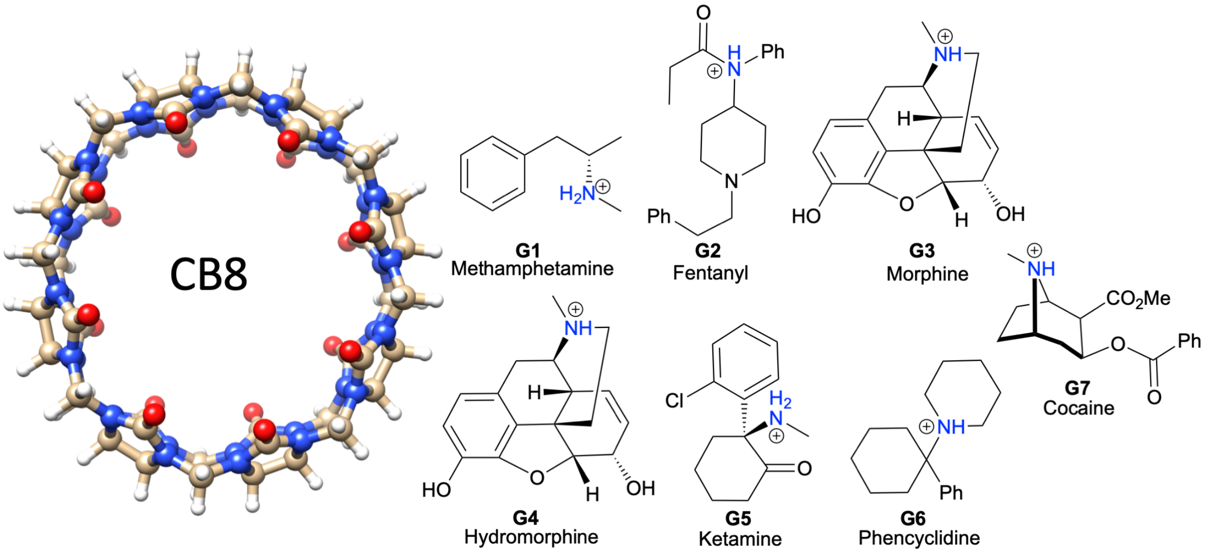


Figure 6.1. Chemical structures of the host CB8 and guests G1 to G7.

As well as giving reasonable agreement with experiment binding free energy, MCC is able to explain these values by showing how the entropy change is distributed over all molecules in the system.

6.2 Theory

6.2.1 Free Energy Theory

The standard binding free energy ($\Delta G^{\circ}_{\text{bind}}$) of the host and guest molecules to form the host-guest complex in aqueous solution at the standard-state 1 M ligand concentration is determined from the Gibbs free energies G calculated directly from simulations of the host in water, the guest in water, the host-guest complex in water, and bulk water:

$$\Delta G^{\circ}_{\text{bind}} = \left(G_{\text{complex}} + G_{\text{water}}\right) - \left(G_{\text{host}} + G^{\circ}_{\text{guest}}\right) \tag{6.1}$$

as illustrated in Figure 6.2. In energy-entropy methods, G is evaluated from the enthalpy H and entropy S using $G = H - TS$ where T is temperature. The pressure-volume PV terms is omitted to allow the approximation $H \approx E$, where E is the system energy, being small, on the order of 3 J mol^{-1} for the solutions studied here and even then almost entirely cancelling in the binding difference.

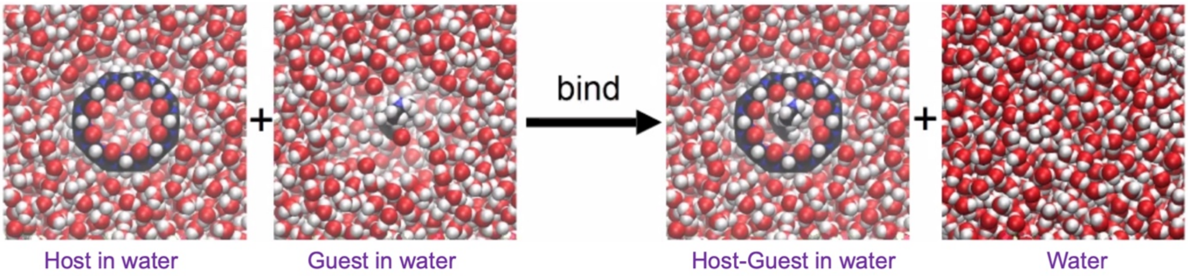


Figure 6.2. The four systems simulated to calculate the binding free energy by the EE method.

In MCC, S is calculated in a multiscale fashion in terms of cells of correlated units. It is the sum of four different kinds of term S_{ij}^{kl}

$$S = \sum_{i}^{\text{molecule}} \sum_{j}^{\text{level}} \sum_{k}^{\text{motion}} \sum_{l}^{\text{minima}} S_{ij}^{kl} \tag{6.2}$$

First, S is partitioned over each kind of molecule i, whether host, guest or water. Second, for the molecules studied here, S has two levels of hierarchy j: molecule (M) and united atom (UA). Third, at each level, S is classified according to the type of motion k: translational or rotational. Fourth, for each motion, S is divided into vibrational and topographical terms l, which arise from the discretization of the potential energy surface into energy minima.

6.2.2 Molecular Entropy

An important feature of MCC is that the same entropy theory is applied to all molecules in the system. However, only the entropy of water molecules in the first hydration shell of the host and guest is considered. This is because this reduces statistical noise that scales with the number of water molecules, and because the entropy of the remaining water molecules is assumed to remain unchanged upon binding. Similarly, the entropy of the single neutralizing Cl^- ion is neglected in the calculations. To ensure balanced stoichiometry in binding, the number of bulk water molecules N_{WB} that contributes to G_{water} in Equation 6.1 is chosen to ensure that

$$N_{\text{WB}} + N_{\text{WS,H-G}} = N_{\text{WS,H}} + N_{\text{WS,G}} \tag{6.3}$$

where $N_{\text{WS,H}}$, $N_{\text{WS,G}}$, and $N_{\text{WS,H-G}}$ are the number of water molecules around the host, guest and host-guest complex, respectively. Later, N_{WB} is partitioned into water released into bulk from the host or guest, $N_{\text{WS,H-G}}$ is partitioned into water nearest the host or guest, and $N_{\text{WS,H}}$ and $N_{\text{WS,G}}$ are in turn partitioned into water released into bulk and water staying with the host or guest, respectively.

6.2.3 Entropy for Each Level and Motion

The axes of each molecule are taken as its principal axes with the origin at the molecular center of mass. All molecules considered here, treated as non-linear rigid bodies, have three translational and three rotational degrees of freedom. At the united-atom level, each united atom is defined as each heavy atom and its bonded hydrogen atoms. A united atom has three translational degrees of freedom and a number of rotational degrees of freedom depending on the number of hydrogens and resulting geometry: 3 for non-linear (>1 hydrogen), 2 for linear (one hydrogen) and 0 for a point (no hydrogens). Its origin is taken as the heavy atom and the axes are defined with respect to the covalent bonds to the bonded hydrogens [43]. Note that it was necessary to use the reference frame of the host-guest complex when evaluating the entropy of the bound host at the united-atom level because this ensured a consistent alignment of the host with the guest.

6.2.4 Vibrational Entropy

The vibrational entropy is evaluated in the harmonic approximation for the quantum hormonic oscillator:

$$S_{\mathrm{vib}} = k_{\mathrm{B}} \sum_{\mathrm{i}=1}^{N_{\mathrm{vib}}} \left(\frac{hv_{\mathrm{i}}/k_{\mathrm{B}}T}{\mathrm{e}^{hv_{\mathrm{i}}/k_{\mathrm{B}}T}-1} - \ln\left(1 - \mathrm{e}^{hv_{\mathrm{i}}/k_{\mathrm{B}}T}\right) \right) \tag{6.4}$$

where k_{B} is Boltzmann's constant, N_{vib} is the number of vibrations, T is temperature, h is Planck's constant, and v_{i} are the vibrational frequencies which are calculated from the eigenvalues λ_i of the appropriate covariance matrix

$$v_{\mathrm{i}} = \frac{1}{2\pi}\sqrt{\frac{\lambda_{\mathrm{i}}}{k_{\mathrm{B}}T}} \tag{6.5}$$

At the molecular level, $N_{\mathrm{vib}} = 6$, corresponding to the *xyz* directions. Two covariance matrices are constructed, one from the mass-weighted forces for translation and one from the moment-of-inertia-weighted torques for rotation, each of these for the whole molecule with forces and torques halved in the mean-field approximation [43-47]. Their associated entropies are termed "transvibrational" and "rovibrational". For transvibration at the united-atom level, $N_{\mathrm{vib}} = 3N - 6$ where N is the number of united atoms in the molecule, and the six lowest-frequency motions have been removed to avoid duplication of transvibrational and rovibrational entropy at the molecular level. For rovibration at the united-atom level, N_{vib} is summed over the number of rotational degrees of freedom of each united atom. Covariance matrices are constructed as before but over all united atoms in the molecule and with halved torques in the mean-field approximation for weakly correlated degrees of freedom.

6.2.5 Topographical Entropy

For the topographical entropy at the molecular level, the translational term is known as the "positional" entropy and the rotational term is known as the "orientational" entropy. The positional entropy at the standard 1 M concentration is evaluated as

$$S_{\mathrm{pos}}^{\circ} = k_{\mathrm{B}} \ln \frac{1}{x_{\mathrm{aq}}^{\circ}} \tag{6.6}$$

where x_{aq}° is the mole fraction of the molecule. For a solute when dilute, this is taken as 1/55.5, where 55.5 is the number of water molecules in the standard volume 1661 Å^3, while for the solvent water $x_{\mathrm{aq}}^{\circ} \approx 1$. The orientational entropy for a molecule in solution is evaluated as

$$S_{\mathrm{or}} = k_{\mathrm{B}} \ln \frac{N_{\mathrm{c}}^{(3/2)} \pi^{1/2} p_{\mathrm{corr}}}{\sigma} \tag{6.7}$$

where N_{c} is the coordination number of the molecule, σ is the symmetry number, and p_{corr} is the probability that the neighboring molecules are oriented suitably for each solute, $p_{\mathrm{corr}} = 1$ while for water $p_{\mathrm{corr}} = 0.25$ to account for hydrogen-bond correlation [43]. For a molecule in solution, N_{c} is the number of solvent molecules in the first hydration shell of the solute calculated using RAD [56]. For a guest bound to the CB8 host,

$$S_{\mathrm{or}} = k_{\mathrm{B}} \ln \frac{\sigma_{\mathrm{host}}}{\sigma_{\mathrm{guest}}} \tag{6.8}$$

where σ_{host}, the symmetry number of the CB8, equals 16, given its 8-fold and 2-fold rotational axes. At the united-atom level, the topographical entropy is known as the "conformational" entropy, with the translational term corresponding to dihedrals involving heavy atoms. It is calculated from the probability distribution of each set of unique conformations for all conformations having dihedrals of united atoms using

$$S_{\mathrm{conf}} = -k_{\mathrm{B}} \sum_{\mathrm{i}=1}^{N_{\mathrm{conf}}} p_{\mathrm{i}} \ln p_{\mathrm{i}} \tag{6.9}$$

where p_{i} and N_{conf} are the probability and number of each set of conformations, respectively. Each conformation is defined adaptively whereby the dihedral is assigned to the nearest peak in the dihedral distribution calculated using a histogram with 30° bin widths [51]. The united-atom rotational topographical term is ignored because it corresponds to dihedrals involving exclusively hydrogens at one end and is either zero by symmetry, as in methyl groups, or small due to strong correlation with the solvent, as for hydroxyl groups. An additional entropic contribution to binding of $-0.5\ k_{\mathrm{B}}$ ln2 was included for guest G5 (Figure 6.1) to account for the shift from half protonated when unbound to fully protonated when bound as pointed out in the SAMPL8 instructions.

6.3 Methodology

6.3.1 System Preparation

The structural coordinates of the host and guest molecules were taken from the SAMPL8 Github website. All guests were built with their amino nitrogen in the protonated state; for guests G3, G4 and G7, the S stereochemistry was taken for G3 and G4 and the R stereochemistry for G7. The starting structure for the host-guest complex was taken as the lowest docked energy in docking of each guest molecule to the host using the AutoDock Vina software [57]. Amber Tools 19 [58] was used to create the topology and coordinate files of each system. The second-generation General AMBER Force Field (GAFF2) [59] with AM1-BCC partial charges as implemented in Antechamber [60] was used for the host and guest, TIP3P [61] was used for water, and the Joung and Cheatham parameters were used for the one chloride ion [62], which was added to neutralize the +1 charge of the guest. Four kinds of MD simulation were set up: (i) 1500 water molecules, (ii) the host molecules solvated in 1500 water molecules, (iii) each guest molecule in 1500 water molecules, and (iv) each host-guest complex in 1500 water molecules. All simulation boxes were cubic with side ~36 Å.

6.3.2 Molecular Dynamics Simulation Protocol

The simulations were performed with the GROMACS 2018.4 software package [63]. The topology and coordinate files for each system were converted from AMBER into GROMACS format using the GROMACS ParmEd tool because the entropy code used later does not yet work with AMBER trajectories. For equilibration, each system was minimized for 500 steps of steepest-descent minimization and heated gradually from 0 to 300 K for 100 ps of NVT molecular dynamics simulation using the V-rescale thermostat [64], followed by 100 ps of NPT simulation using the Parrinello-Rahman barostat [65] with a 2 ps time constant and the isothermal compressibility of water 4.5×10^{-5} bar^{-1}. The long-range electrostatic interactions were calculated using the Partial Mesh Ewald (PME) method with the Verlet cutoff-scheme, the non-bonded cutoff was 10 Å with periodic boundary conditions, and the time step was 2 fs. Data collection under the same conditions was run for 100 ns of MD simulation, with forces and coordinates saved every 100 ps to give 1000 frames for analysis. Entropies were calculated using MCC [44,45] with additional terms for binding [48]. Calculation of all entropy terms was performed with two separate python codes, one code for the solutes (Github) and an in-house code for the solvent, each reading in the force, coordinate and topology files for each simulation. Four simulation were needed for each binding calculation as shown in Figure 6.2 and each MD simulation was run in triplicate with slightly different starting structures, yielding ΔG of binding via equation 6.1.

6.3.3 Error Analysis

The standard error of the mean (SEM) for *G*, *H* and *S* are calculated from the standard deviation σ of the values from those derived from the three separate simulations

$$\mathrm{SEM} = \frac{\sigma}{\sqrt{n}} \tag{6.11}$$

where *n* = 3 is the number of simulations. The mean average error (MAE) with respect to experiment is

$$\mathrm{MAE} = \frac{\Sigma|\Delta G - \Delta G_{\mathrm{expt}}|}{n} \tag{6.12}$$

where *n* = 7 is the number of molecules.

6.4 Results and Discussion

The calculated binding Gibbs free energies together with SEM error bars are plotted in Figure 6.3 versus experiment.

The values of the EE-MCC and experimental [66] binding Gibbs free energies are listed in Table 6.1, together with the Δ*H* and *T*Δ*S* components. The MAE for Δ*G* averaged unsigned error over all molecules is 0.9 kcal mol^{-1} and for Δ*H* and *T*Δ*S* they are 2.0 and 1.8 kcal mol^{-1}, respectively. Evidently, there is some correlation between the enthalpy and entropy that brings about a lower error in the binding Gibbs free energy than in these two components, particularly for compounds G2, G4 and G5 which have larger but compensating errors in Δ*H* and *T*Δ*S*.

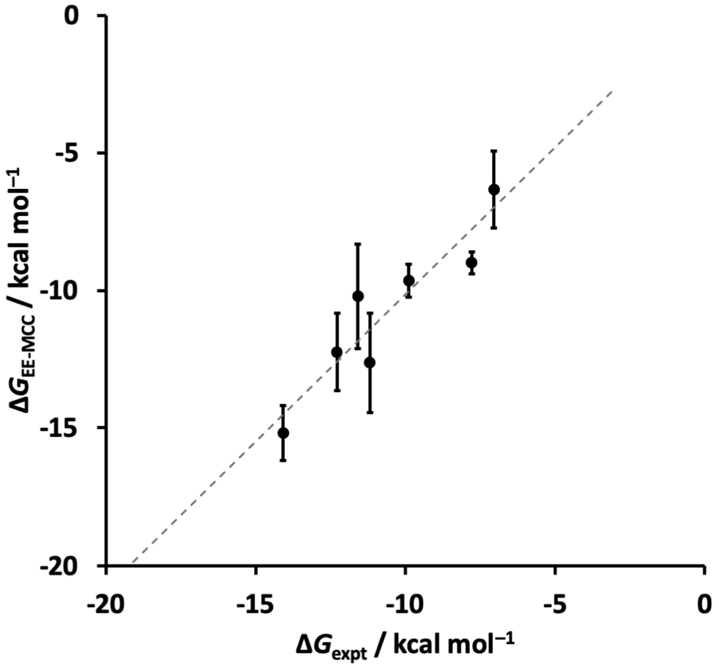


Figure 6.3. EE-MCC Gibbs free energies of binding (error bars given by the SEM) versus experiment.

Table 6.1. Predicted Binding Free Energies, Enthalpies and Entropies versus Experiment [66]

Guest	ΔG / kcal mol^{-1}		ΔH / kcal mol^{-1}		$T\Delta S$ / kcal mol^{-1}	
	EE-MCC	Expt	EE-MCC	Expt	EE-MCC	Expt
G1	–6.3	–7.1	–7.6	–7.8	–1.3	–0.8
G2	–9.6	–9.9	–5.0	–10.8	4.6	–0.9
G3	–10.2	–11.6	–11.9	–13.6	–1.7	–2.0
G4	–12.6	–11.2	–11.7	–15.8	1.0	–4.6
G5	–12.2	–12.3	–14.0	–17.3	–1.7	–5.0
G6	–15.33	–14.1	–14.4	–14.9	1.0	–0.8
G7	–9.0	–7.9	–11.5	–8.3	–2.5	–0.3

6.4.1 Entropy Components with MCC

MCC yields the entropy of the system and its decomposition over molecules, level, motion and minima according to equation 6.2. In figure 6.3 we show plots for the changes in vibrational and topographical entropy components upon binding for the host and guest at molecule and united-atom levels of hierarchy.

The host entropy, which is all vibrational, decreases for all guests but only by a small amount. The contributions are slightly larger at the united-atom level and the rovibrational term is sometimes weakly positive. The positional and orientational entropy of the host is taken not to change, taken as defining the reference frame for the binding process. The decrease in entropy of the guest is much larger because it comprises the loss of positional and orientational entropy, the former constant for all guests at 1 M concentration and the latter dependent on the size of the molecule via the number of first-shell water molecules. There is a smaller but moderate decrease in conformational entropy of up to 15 J K^{-1} mol^{-1} for the more flexible guests G7, G1, G2 and G5 which have more freely rotating dihedrals. The guests have only a small decrease in vibrational entropy, as for the host, with the occasional tiny increase at the united-atom level. The total guest entropy losses of 60-75 J K^{-1} mol^{-1} are similar to the values of 71-73 J K^{-1} mol^{-1} from an earlier study on protein-ligand systems with comparatively sized ligands that only considered the molecule-level entropy.

The corresponding changes in water entropy are show in Figure 6.5. There is a fairly sizeable decrease in rovibrational entropy for water around the host upon binding, with the exception of G6 which has a slight increase, possibly because its cationic nitrogen is fully buried inside the host and so cannot constrain water molecules. The changes in water's transvibrational and orientational entropy are smaller and either higher or lower, depending on the guest. The changes in water hydrating the guest are smaller, given that the guest has little solvent exposure when bound; in most cases the decrease is transvibrational or orientational, with some increase rovibrational. For water released into bulk from either host or guest, there is a large gain in orientational entropy for all guests, consistent with the larger number of hydrogen-bonding neighbours of a water molecule in bulk. There is a larger contribution from water around the guest because the guest becomes more buried and releases more water molecules. Water released from the host is seen to gain a small amount of transvibrational entropy, while the vibrational terms change little for the guest.

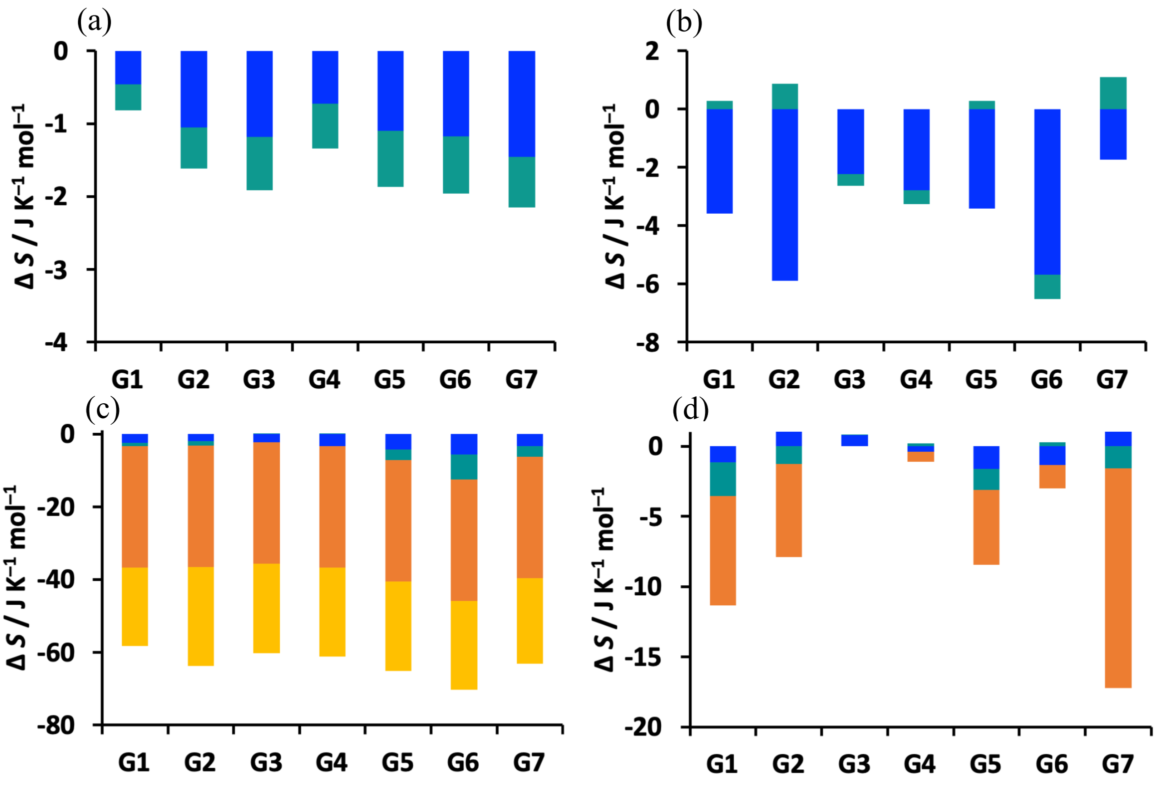


Figure 6.4. Binding entropy components for the (a) host at molecular level, (b) host at united-atom level, (c) guest at molecular level, and (d) guest at united-atom level. The components are transvibrational (blue), rovibrational (turquoise), positional/conformational (orange), and orientational (yellow).

Table 6.2. Entropy Components of Unbound and Bound Host and Associated Water (J K^{-1} mol^{-1})

	H	H-G1	H-G2	H-G3	H-G4	H-G5	H-G6	H-G7
$S_{\mathrm{H,M}}^{\mathrm{transvib}}$	70	70	69	69	69	68	69	69
$S_{\mathrm{H,M}}^{\mathrm{rovib}}$	74	73	73	73	73	73	74	73
$S_{\mathrm{H,UA}}^{\mathrm{transvib}}$	662	659	656	660	659	661	658	658
$S_{\mathrm{H,UA}}^{\mathrm{rovib}}$	159	159	160	158	158	159	158	160
$S_{\mathrm{H,UA}}^{\mathrm{conf}}$	0	0	0	0	0	0	0	0
$S_{\mathrm{WS}}^{\mathrm{transvib}}$	4079	3566	3440	3505	3526	3583	3555	3507
$S_{\mathrm{WS}}^{\mathrm{rovib}}$	1515	1310	1253	1275	1290	1312	1232	1270
$S_{\mathrm{WS}}^{\mathrm{conf}}$	251	648	632	638	652	661	647	635
N_{WS}	87.4	76.4	73.7	75.2	75.4	76.9	76.2	75.3

$S_{\mathrm{WB}}^{\mathrm{transvib}}$		516	644	573	563	491	525	569
$S_{\mathrm{WB}}^{\mathrm{rovib}}$		190	238	211	208	181	194	210
$S_{\mathrm{WB}}^{\mathrm{conf}}$		123	153	136	134	117	125	135
N_{WB}		11.0	13.7	12.2	12.0	10.5	11.2	12.1

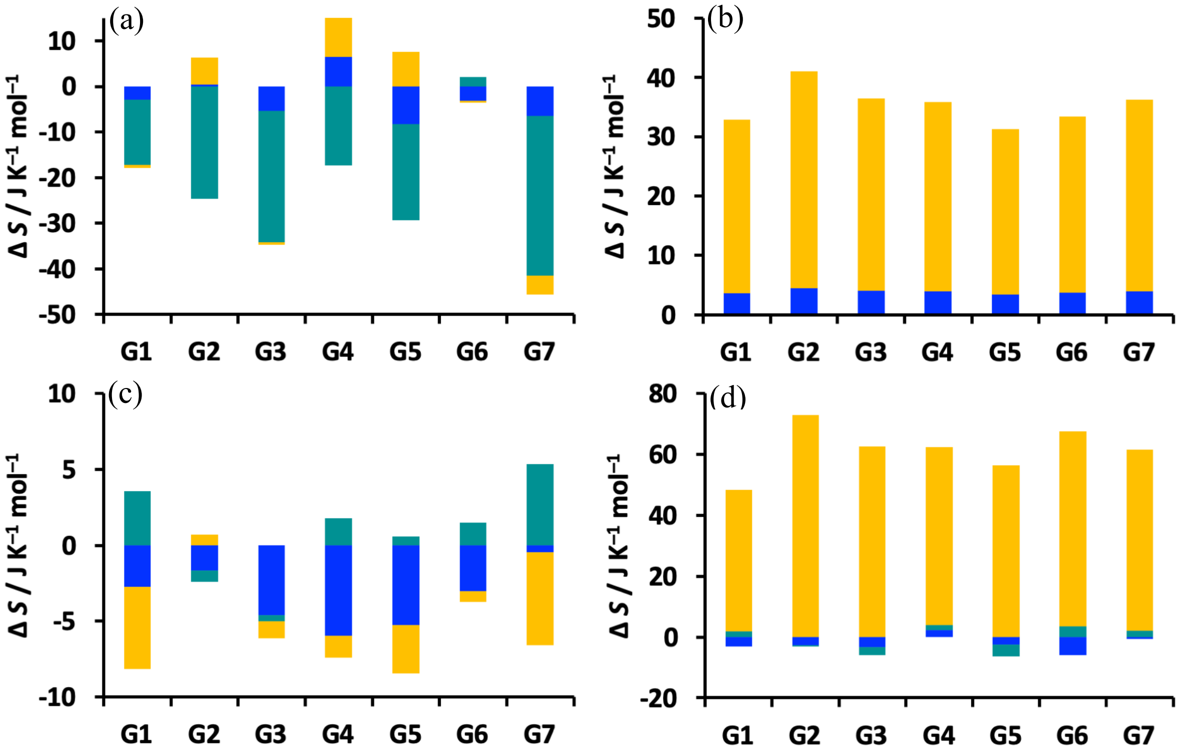


Figure 6.5. Changes in binding entropy components for the (a) water staying in the hydration shell of the host (WS), (b) water released from the host into bulk water (WB), (c) water staying in the hydration shell of the guest (WS), and (d) water released from the guest into bulk water (WB). Coloring is as in Figure 6.3.

The corresponding entropy components for all contributing species when unbound or bound are shown in Tables 6.2 and 6.3, together with the number of contributing water molecules, either staying bound in the hydration shell of the host or guest (WS) or being released into bulk (WB).

Table 6.3. Entropy Components of Unbound and Bound Guests and Associated Water (J K^{-1} mol^{-1})

	Component	G1	G2	G3	G4	G5	G6	G7
Unbound Guest	$S_{\mathrm{G,M}}^{\mathrm{transvib}}$	62	68	66	67	66	67	68
	$S_{\mathrm{G,M}}^{\mathrm{rovib}}$	59	67	61	62	65	68	64
	$S_{\mathrm{G,M}}^{\mathrm{pos}}$	33	33	33	33	33	33	33
	$S_{\mathrm{G,M}}^{\mathrm{or}}$	45	50	48	48	46	48	49
	$S_{\mathrm{G,UA}}^{\mathrm{transvib}}$	41	137	96	95	84	81	119
	$S_{\mathrm{G,UA}}^{\mathrm{rovib}}$	74	118	86	87	72	99	96
	$S_{\mathrm{G,UA}}^{\mathrm{conf}}$	20	29	0	2	16	7	17
	$S_{\mathrm{WS}}^{\mathrm{transvib}}$	1144	1786	1458	1460	1335	1496	1663
	$S_{\mathrm{WS}}^{\mathrm{rovib}}$	419	658	540	538	496	545	610
	$S_{\mathrm{WS}}^{\mathrm{conf}}$	72	126	97	278	254	277	311
	N_{WS}	24.3	37.9	30.9	31.1	28.4	31.7	35.4
Bound Guest	$S_{\mathrm{G,M}}^{\mathrm{transvib}}$	60	66	64	64	62	63	65
	$S_{\mathrm{G,M}}^{\mathrm{rovib}}$	58	64	61	61	62	62	62
	$S_{\mathrm{G,M}}^{\mathrm{pos}}$	0	0	0	0	0	0	0
	$S_{\mathrm{G,M}}^{\mathrm{or}}$	23	23	23	23	23	23	23
	$S_{\mathrm{G,UA}}^{\mathrm{transvib}}$	40	138	96	95	82	80	121
	$S_{\mathrm{G,UA}}^{\mathrm{rovib}}$	72	116	86	84	70	99	95
	$S_{\mathrm{G,UA}}^{\mathrm{conf}}$	12	22	0	1	10	6	1
	$S_{\mathrm{WB}}^{\mathrm{transvib}}$	1031	1395	1239	1242	1201	1248	1190
	$S_{\mathrm{WB}}^{\mathrm{rovib}}$	377	514	459	458	446	454	437
	$S_{\mathrm{WB}}^{\mathrm{conf}}$	198	258	231	236	228	231	223
	N_{WB}	21.9	29.6	26.3	26.5	25.5	26.4	25.3
	$S_{\mathrm{WS}}^{\mathrm{transvib}}$	110	390	214	212	129	245	472
	$S_{\mathrm{WS}}^{\mathrm{rovib}}$	45	144	81	82	51	92	179
	$S_{\mathrm{WS}}^{\mathrm{conf}}$	16	73	40	40	22	45	82
	N_{WS}	2.4	8.3	4.7	4.6	2.9	5.3	10.1

These numbers are consistent with the trends in Figure 6.4 and 6.5. Their most insightful revelation is the magnitudes of the entropies involved. Clearly, most of the entropy is in the solvent water, and the size of this entropy term scales near linearly with the number of water molecules in the first hydration shell. The contributions from the host and guest molecules for their respective unbound cases are much smaller at only about 14% and 14-20%, respectively. Most of the host entropy, 85%, is at the united atom level, and of that, 80% is transvibrational and the rest rovibrational while at the molecule level these two terms are comparable in size, as seen in earlier work [34,43,44,48]. For the guest the two levels have similar amounts of entropy, depending on the size of the ligand and at the 1 M concentration being used here. The numbers of water molecules in each of the four categories makes clear that the guest is almost entirely desolvated upon binding and that the host loses comparatively fewer water molecules to accommodate the guest, supporting the finding in Figure 6.3 that guest desolvation contributes more than host desolvation for the systems studied here.

6.5 Conclusions

A new energy-entropy method called EE-MCC has been presented to calculate the free energy of binding and applied to a series of aqueous host-guest complexes in the SAMPL8 “Drugs of Abuse” Blind Challenge. EE-MCC accounts for the entropy of all flexible degrees of freedom of the system in a consistent and general manner. The calculated binding Gibbs free energy values are in good agreement with experimental results having average standard error of mean 0.9 kcal mol^{-1}. The main feature of MCC is that it provides the entropy components over all molecules and all degrees of freedom in the system at a hierarchy of length scales. There is a large loss of positional and orientational entropy that is fairly similar for all guests, with the orientational entropy loss larger for larger guests. There is a smaller loss of conformational entropy, depending on the flexibility of the guest. There are also smaller decreases in vibrational entropy of the host, guest and contacting water. These losses are compensated by a large gain in orientational entropy of water released to bulk, with the larger contribution coming from water that was hydrating the guest.

6.6 References

[1] Gilson, M.K., Given, J.A., Bush, B.L. and McCammon, J.A., The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys. J.* **1997,** *72*, 1047-1069.

[2] Luo, H. and Sharp, K., On the calculation of absolute macromolecular binding free energies. *Proc. Natl. Acad. Sci.* **2002,** 99, 10399–10404.

[3] Mobley, D.L. and Gilson, M.K., Predicting binding free energies: frontiers and benchmarks. *Annu. Rev. Biophys.* **2017**, *46*, 531–558.

[4] Gaieb, Z., Liu, S., Gathiaka, S., Chiu, M., Yang, H., Shao, C., Feher, V.A., Walters, W.P., Kuhn, B., Rudolph, M.G. et al., D3R Grand Challenge 2: blind prediction of protein-ligand poses, affinity rankings, and relative binding free energies. *J. Comput. Aided Mol. Des.* **2018**, *32*, 1–20.

[5] Pantsar, T. and Poso, A., Binding affinity via docking: Fact and fiction. *Molecules* **2018,** *23*, 1899.

[6] Böhm, H.J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **1994**, *8*, 243–256.

[7] Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, *11*, 425–445.

[8] Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J. and Koes, D.R., Protein–ligand scoring with convolutional neural networks. *J. Chem. Inf. Model* **2017**, *57*, 942–957.

[9] Skalic, M., Martínez-Rosell, G., Jiménez, J. and De Fabritiis, G., Play molecule bind scope: large scale CNN-based virtual screening on the web. *Bioinformatics* **2019**, *35*, 1237–1238.

[10] Adeshina, Y.O., Deeds, E.J. and Karanicolas, J., Machine learning classification can reduce false positives in structure-based virtual screening. *Proc. Natl. Acad. Sci. USA.* **2020** *117*, 18477–18488.

[11] Gervasio, F.L., Laio, A. and Parrinello, M., Flexible docking in solution using metadynamics. *J. Am. Chem. Soc.* **2005,** *127*, 2600-2607.

[12] Woo, H.-J. and Roux, B., Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA.* **2005,** 102, 6825-6830.

[13] Doudou, S., Burton, N.A. and Henchman, R.H., Standard free energy of binding from a one-dimensional potential of mean force. *J Chem. Theory Compu.* **2009,** *5*, 909-918.

[14] Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M.K., Greenwood, J. et al. Accurate and reliable prediction of relative ligand binding potency in pro- spective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

[15] Cournia, Z., and Allen, B. and Sherman, W., Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model.* **2017**, 57, 2911-2937.

[16] Straatsma, T.P. and McCammon, J.A., Computational alchemy. *Annu. Rev. Phys. Chem.* **1992,** *43*, 407-435.

[17] Bhati, A.P., Wan, S., Wright, D.W. and Coveney, P.V., Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theory Compu.* **2017,** *13*, 210-222.

[18] Honig, B. and Nicholls, A., Classical electrostatics in biology and chemistry. *Science* **1995**, *268*, 1144–1149.

[19] Kollman, P.A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D.A. and Cheatham, T.E., Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000,** *33*, 889-897.

[20] Wang, C., Greene, D.A., Xiao, L., Qi, R. and Luo, R., Recent developments and applications of the MMPBSA method. *Fron. Mol. Biosci.* **2018,** *4*.

[21] Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J.Z.H. and Hou, T., End-Point binding free energy calculation with MM/PBSA and MM/GBSA: Strategies and applications in drug design. *Chem. Rev.* **2019,** *119*, 9478-9508.

[22] Massova, I. and Kollman, P.A., Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug. Discov. Des.* **2000,** *18*, 113–135.

[23] Tidor, B. and Karplus, M., The contribution of vibrational entropy to molecular association: The dimerization of insulin. *J. Mol. Bio.* **1994**, *238*, 405–414.

[24] Kongsted, J. and Ryde, U., An improved method to predict the entropy term with the MM/PBSA approach. *J. Comput. Aided Mol. Des.* **2008**, *23*, 63.

[25] Swanson, J.M.J., Henchman, R.H. and McCammon, J.A., Revisiting free energy calculations: A theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* **2004,** 86, 67-74.

[26] Luo, H. and Sharp, K., On the calculation of absolute macromolecular binding free energies. *Pro. Natl. Acad. Sci.* **2002**, *99*, 10399–10404.

[27] Chang, C. -E., Chen, W. and Gilson, M. K., Evaluating the accuracy of the quasiharmonic approximation. *J. Chem. Theory Comput.* **2005**, *1*, 1017–1028.

[28] Chang, C.-e.A., Chen, W. and Gilson, M.K., Ligand configurational entropy and protein binding. Proc. Natl. Acad. Sci. **2007,** *104*, 1534-1539.

[29] Chang, C. -E. and Gilson, M. K., Free energy, entropy, and induced fit in host−guest recognition: Calculations with the second-generation mining minima algorithm. *J. Am. Chem. Soc.* **2004**, *126*, 13156–13164.

[30] Diehl, C., Genheden, S., Modig, K., Ryde, U. and Akke, M., Conformational entropy changes upon lactose binding to the carbohydrate recognition domain of galectin-3. *J. Biomol. NMR* **2009**, *45*,157–169.

[31] Fenley, A.T., Killian, B.J., Hnizdo, V., Fedorowicz, A., Sharp, D.S. and Gilson, M. K., Correlation as a determinant of configurational entropy in supramolecular and protein systems. *J. Phys. Chem. B* **2014,** *118*, 6447-6455.

[32] King, B.M., Silver, N.W. and Tidor, B., Efficient calculation of molecular configurational entropies using an information theoretic approximation. *J. Phys. Chem. B* **2012,** *116*, 2891-904.

[33] Suárez, D. and Díaz, N., Ligand strain and entropic effects on the binding of macrocyclic and linear inhibitors: Molecular modeling of penicillopepsin complexes. *J. Chem. Inf. Model.* **2017**, *57*, 2045–2055.

[34] Suárez, D. and Díaz, N., Affinity calculations of cyclodextrin host–guest complexes: Assessment of strengths and weaknesses of end-point free energy methods. *J. Chem. Inf. Model.* **2019**, *59*, 421–440.

[35] Irudayam, S. J. and Henchman, R. H., Entropic cost of protein−ligand binding and its dependence on the entropy in solution. *J. Phys. Chem. B* **2009**, *113*, 5871–5884.

[36] Li, Z. and Lazaridis, T., Thermodynamic contributions of the ordered water molecule in HIV-1 protease. *J. Am. Chem. Soc.* **2003**, *125*, 6636–6637.

[37] Abel, R., Young, T., Farid, R., Berne, B. J. and Friesner, R. A., Role of the active-site solvent in the thermodynamics of factor Xa ligand binding. *J. Am. Chem. Soc.* **2008**, *130*, 2817–2831.

[38] Nguyen, C. N., Young, T. K. and Gilson, M. K., Grid inhomogeneous solvation theory: Hydration structure and thermodynamics of the miniature receptor cucurbit[7]uril. *J. Chem. Phys.* **2012**, *137*, 044101–044101.

[39] Genheden, S., Luchko, T., Gusarov, S., Kovalenko, A. and Ryde, U., An MM/3D-RISM approach for ligand binding affinities. *J. Phys. Chem. B* **2010**, *114*, 8505–8516.

[40] Wong, S., Amaro, R. E. and McCammon, J. A., MM-PBSA captures key role of intercalating water molecules at a protein−protein interface. *J. Chem. Theory Comput.* **2009** *5*, 422–429.

[41] Raman, E. P. and MacKerell, A. D., Spatial analysis and quantification of the thermodynamic driving forces in protein-ligand binding: binding site variability. *J. Am. Chem. Soc.* **2015**, *137*, 2608–2621.

[42] Verteramo, M. L., Stenström, O., Ignjatović, M. M., Caldararu, O., Olsson, M. A., Manzoni, F., Leffler, H., Oksanen, E., Logan, D. T., Nilsson, U. J., Ryde, U. and Akke, M., Interplay between conformational entropy and solvation entropy in protein–ligand binding. *J. Am. Chem. Soc,* **2019**, *141,* 2012–2026.

[43] Higham, J., Chou, S.-Y., Gräter, F. and Henchman, R.H., Entropy of flexible liquids from hierarchical force–torque covariance and coordination. *Mol. Phys.* **2018,** 116, 1965-1976.

[44] Ali, H.S., Higham, J. and Henchman, R.H., Entropy of simulated liquids using multiscale cell correlation. *Entropy* **2019**, *21*, 750.

[45] Ali, H.S., Higham, J., de Visser, S.P. and Henchman, R.H., Comparison of free-energy methods to calculate the barriers for the nucleophilic substitution of alkyl halides by hydroxide. *J. Phys. Chem. B* **2020**, *124*, 6835–6842.

[46] Henchman, R.H., Partition function for a simple liquid using cell theory parametrized by computer simulation. *J. Chem. Phys.* **2003,** *119*, 400-406.

[47] Henchman, R.H., Free energy of liquid water from a computer simulation via cell theory. *J. Chem. Phys.* **2007,** *126*, 064504.

[48] Irudayam, S.J., Plumb, R.D. and Henchman, R.H., Entropic trends in aqueous solutions of the common functional groups. *Faraday Discuss.* **2010,** *145*, 467-485.

[49] Gerogiokas, G., Calabro, G., Henchman, R.H., Southey, M.W.Y., Law, R.J. and Michel, J., Prediction of small molecule hydration thermodynamics with grid cell theory. *J. Chem. Theory Comput.* **2014,** *10*, 35-48.

[50] Hensen, U., Gräter, F. and Henchman, R.H., Macromolecular entropy can be accurately computed from force. *J. Chem. Theory Comput.* **2014,** *10*, 4777-4781.

[51] Chakravorty, A., Higham, J. and Henchman, R.H., Entropy of proteins using multiscale cell correlation. *J. Chem. Inf. Model.* **2020,** *60*, 5540-5551.

[52] Muddana, H.S., Daniel, V.C., Bielawski, C.W., Urbach, A.R., Isaacs, L., Geballe, M.T. and Gilson, M.K., Blind prediction of host–guest binding affinities: A new SAMPL3 challenge. *J. Comput. Aided Mol. Des.* 2012, 26, 475–487.

[53] Muddana, H.S., Fenley, A.T., Mobley, D.L. and Gilson, M.K., The SAMPL4 host–guest blind prediction challenge: An overview. *J. Comput. Aided Mol. Des.* **2014**, *28*, 305–317.

[54] Yin, J., Henriksen, N.M., Slochower, D.R., Shirts, M.R., Chiu, M.W., Mobley, D.L. and Gilson, M.K., Overview of the SAMPL5 host–guest challenge: Are we doing better? *J. Comput. Aided Mol. Des.* **2017**, *31*, 1–19.

[55] Rizzi, A., Murkli, S., McNeill, J.N., Yao, W., Sullivan, M., Gilson, M.K., Chiu, M.W., Isaacs, L., Gibb, B.C., Mobley, D.L. and Chodera, J.D., Overview of the SAMPL6 host-guest binding affinity prediction challenge. *J. Comput. Aided Mol. Des*. **2018**, *32*, 937-963.

[56] Higham, J. and Henchman, R.H., Locally adaptive method to define coordination shell. *J. Chem. Phys.* **2016**, *145*, 084108.

[57] Trott, O. and Olson, A.J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Compu. Chem.* **2010**, *31*, 455–461.

[58] Case, D.A., Ben-Shalom, I.Y., Brozell, S.R., Cerutti, D.S., Cheatham, T.E., Cruzeiro, V.W.D., Darden, T.A., Duke, R.E., Ghoreishi, D., Giambasu, G., et al. **2019**, AMBER 2019, University of California, San Francisco.

[59] Träg, J. and Zahn, D., Improved GAFF2 parameters for fluorinated alkanes and mixed hydro- and fluorocarbons. *J. Mol. Model.* **2019**, *25*, 39.

[60] Wang, J.M., Wang, W., Kollman, P.A. and Case, D.A., Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph Model.* **2006**, *25*, 247–260.

[61] Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[62] Joung, I.S. and Cheatham, T.E., Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

[63] Abraham, M.J., van der Spoel, D., Lindahl, E. and Hess, B., GROMACS development team, GROMACS User Manual version 2018.4, www.gromacs.org, **2018**.

[64] Bussi, G., Donadio, D. and Parrinello, M., Canonical sampling through velocity rescaling. *J. Chem. Phys*. **2007**, 126, 014101.

[65] Parrinello, M. and Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

[66] Steven, M., Jared, K., Brockett, T. A., Michael, S., Volker, B., Matthew, R. R., et al. In vitro and in vivo sequestration of phencyclidine by Me4Cucurbit[8]uril. ChemRxiv. Preprint. **2020**. https://doi.org/10.26434/chemrxiv.12994004.v1

Chapter 7. Conclusions and Future Work

7.1 Conclusions

The present work presents a new, efficient and general method to calculate the free energy values of a wide range of systems, spanning liquids, solutions, chemical reactions and host-guest systems. It also presents detailed studies to investigate the stability and reaction kinetics for a number of chemical processes that are catalyzed by various heme and non-heme iron enzymes.

In relation to the new free energy method, we developed the theory Multiscale Cell Correlation (MCC) to calculate the entropy from the force and torque covariance matrices and probabilities distributions of molecular coordinations and conformations in a molecular dynamics simulation. MCC theory was applied to calculate the entropy of 58 important industrial liquids which were modelled with the General AMBER Force Field (GAFF) and Optimized Potentials for Liquids Simulations (OPLS) force fields. The calculated entropy values when compared with experiment had values with unsigned errors of 8.7 J K^{-1} mol^{-1} and 9.8 J K^{-1} mol^{-1} for the GAFF and OPLS force fields, making GAFF slightly better than OPLS for these liquids, and showing thatMCC is also better than the 2-Phase Thermodynamics (2PT) method which is currently the only other method available to calculate the entropy of such liquids.

Next, we combined the MCC theory with the density functional theory (DFT) method in the quantum mechanics / molecular mechanics (QM/MM) formalism to develop a new energy-entropy (EE) method (EE-MCC). The EE-MCC method is applied to a series of six S_N2 chemical reactions where a halogen atom is replaced by a hydroxyl group in aqueous solution. The EE-MCC is used to calculate the Gibbs free energy barriers for six reactions and compared with experimental values. We also calculated the Gibbs free energy barriers for these six reactions in implicit solvent using the EE-NMA method with B3LYP/6-31+G* and M06/6-31+G* level of DFT where entropy is calculated with Normal Mode Analysis (NMA) and in explicit solvent using the Potential Mean Force (PMF). We have used two Hamiltonians, self-consistent charge density functional tight binding (SCC-DFTB) method and DFT method with

B3LYP/6-31+G* in explicit solvent model with respective PMF values while experimental free energy barriers are derived from rate constants. The EE-MCC free energy barriers calculated using the SCC-DFTB method are in good agreement with experimental and PMF values.

Most recently, we also applied the EE-MCC method to calculate the binding Gibbs free energies of host-guest systems directly from MD simulations. The binding free energy values calculated by EE-MCC method are in good agreement with experimental values and a standard error of mean is 0.9 kcal mol^{-1}. It also gives us detailed insight into other thermodynamic quantities such as entropy and enthalpy values which are directly calculated from MD simulations. The entropy and enthalpy calculated by the EE-MCC method are also in reasonable agreement with the experimental values, having a mean average error 2.0 kcal mol^{-1} and 1.8 kcal mol$^{–1}$.

This work studied a range of P450 isozymes by using either a large model quantum mechanics (QM) cluster technique or full QM/MM method or both. GcoA is an isozyme of cytochrome P450 and used to convert lignin units into useful products. We studied the syringol activation by an iron(IV)-oxo heme cation radical oxidant using QM-cluster model on a large active site cluster model of GcoA enzyme. We studied the detailed reaction free energy profile which leads to the selective o-demethylation and acetal formation. We also provided information about selective product formation by doing in-silico mutations. OxyB is another isozyme of P450 which is used to catalyze aromatic cross-linking of two glycopeptide units for the biosynthesis of vancomycin. Similarly, OleT$_{JE}$ and TxtE are isozymes of cytochrome P450 and the former is used to convert fatty acids into a range of hydroxylation, desaturation and decarboxylation products while the later one is used for the aromatic nitration of L-tryptophan (L-Trp).

We also studied the dihydroxylation of L-arginine (L-Arg) catalyzed by OrfP enzyme. We have used large active site cluster model consist of 278 atoms and run the density functional theory method to explore the full potential energy profile. We show that the dihydroxylation of L-Arg occurs via a two-step hydroxylation reaction. We also studied the different product distributions of L-Arg and L-homoarginine (L-hArg) which are catalyzed by another non-heme iron enzyme known as viomycin (VioC). Interestingly, selective hydroxylation of L-arginine at the C^3-position for antibiotics biosynthesis while experimental studies showed that using the substrate analogue, namely L-homo-arginine, a mixture of products was obtained originating from C^3-hydroxylation, C^4-hydroxylation and C^3–C^4 desaturation. To understand how the addition of one CH_2 group to a substrate can lead to such a dramatic change in selectivity and activity, we did a computational study using QM cluster models. We set up a large active-site cluster model of 245 atoms that includes the oxidant with its first- and second-coordination sphere influence as well as the substrate-binding pocket. The model was validated against experimental work on related enzymes and previous computational studies. Thereafter, possible pathways leading to products and byproducts were investigated for a model containing L-Arg and one for L-homo-Arg as substrate. The calculated free energies of activation predicted product distributions that matched experimental observation and gave a low-energy C^3-hydroxylation pathway for L-Arg, while for L-homo-Arg several barriers were found to be close in energy, leading to a mixture of different products.

7.2 Future Work

We will extend our EE-MCC method to calculate the ligand-protein binding Gibbs free energies which will be helpful in computational biophysics and structure-based drug design. Furthermore, we are working on some other nonheme iron dioxygenases enzymes. One such enzyme, namely HygX performs an oxidative ring-closure reaction to form the ortho-ether linkage. Another one is Taurine/α-ketoglutarate dioxygenase (TauD) which is an important enzyme that takes part in the cysteine catabolism process in the human body and hydroxylates taurine. The activation of substrate by TauD is highly stereo- and regioselective and only takes place on the pro-*R* C^1-position. To understand the regio- and stereoselectivity of TauD and HygX enzymes we performed a detailed density functional theory study on large active site cluster models with 244 and 302 atoms respectively that include the first- and second-coordination sphere of the nonheme iron center and substrate binding pockets. For further investigation of kinetics of the enzymatic reaction mechanisms, we are also using full QM/MM simulations to understand the activities of these enzymes. Both these projects are still in progress and I am running some additional QM/MM simulations.

Chapter 8. List of Publications

[1] **Ali H. S**., Higham J. and Henchman R. H., (2020): Entropy of simulated liquids using multiscale cell correlation (MCC). *Entropy*, 21, 750.

[2] **Ali H. S.**, Higham J., de Visser S. P. and Henchman R. H., (2020): Comparison of free-energy methods to calculate the barriers for the nucleophilic substitution of alkyl halides by hydroxide. *The Journal of Physical Chemistry B*, 124, 6835-6842.

[3] **Ali H. S.,** Henchman R. H. and de Visser S. P., (2020): Cross-linking of aromatic phenolate groups by cytochrome P450 enzymes: A model for the biosynthesis of vancomycin by OxyB. *Organic and Biomolecular Chemistry*, 18, 4610-4618.

[4] **Ali H. S.,** Henchman R. H. and de Visser S. P., (2020): Lignin biodegradation by a cytochrome P450 enzyme: A computational study into syringol activation by GcoA. *Chemistry – A European Journal*, 26, 13093-13102.

[5] Louka S., Barry S. M., Heyes D. J., Mubarak M. Q. E., **Ali H. S.,** Alkhalaf L. M., Munro A. W., Scrutton N. S., Challis G. L. and de Visser S. P., (2020): The catalytic mechanism of aromatic nitration by cytochrome P450 TxtE: Involvement of a ferric-peroxynitrite intermediate. *The Journal of American Chemical Society*, 142, 15764-15779.

[6] Chowdhury A.S., **Ali H. S**., Faponle A. S. and de Visser S. P., (2020): How external perturbations affect the chemoselectivity of the reaction and product distributions in cytochrome P450 OleTJE. *Physical Chemistry Chemical Physics*, 22, 27178-27190.

[7] **Ali H. S.,** Henchman R. H. and de Visser S. P., (2021): What determines the selectivity of arginine dihydroxylation by the nonheme iron enzyme OrfP? Accepted in: *Chemistry – A European Journal,* 27, 1795-1809.

[8] **Ali H. S.,** Henchman R. H., Warwicker J. and de Visser S. P., (2021): Negative catalysis by a nonheme iron dioxygenase; an example of selective hydroxylation and desaturation by VioC. *Journal of Physical Chemistry A*, 125, 1720-1737.

[9] de Visser S. P., Lin Y. T., **Ali H. S.**, Bagha U. K., Mukherjee G. and Sastri C. V., (2021): Negative catalysis by metalloenzymes: Examples from heme and nonheme iron oxygenases. *Coordination Chemistry Review*, 439, 213914.

[10] Han S. B, **Ali H. S.** and de Visser S. P., (2021): Glutarate hydroxylation by the carbon starvation-induced protein D. A computational study into the stereo- and regioselectivity of the reaction. *Inorganic Chemistry*, 60, 4800-4815.

[11] Lin Y. T, **Ali H. S.** and de Visser S. P., (2021): Negative catalysis by a nonheme iron dioxygenases: The selectivity of the TmpA biosynthesis enzyme. *Chemistry – A European Journal*, (accepted/in press).

[12] **Ali H. S.,** Henchman R. H. and de Visser S. P., (2021): Mechanism of oxidative ring-closure as part of the orthosmycin biosynthesis step by a nonheme iron dioxygenase. *ChemCatChem* (accepted/in press).

[13] **Ali H. S.,** Ghafoor S. and de Visser S. P., (2021): Density functional theory study into the reaction mechanism of isonitrile biosynthesis by the nonheme iron enzyme ScoE. *Topics in Catalysis*, (accepted/in press).

[14] **Ali H. S.,** Chakravorty A., Kalayan J., de Visser S. P. and Henchman R. H., (2021); Binding free energies for the SAMPL8 cucurbit[8]uril host-guest challenge using multiscale cell correlation. *Journal of Computer-Aided Molecular Design*, (accepted/in press)