Measuring the Effects of Reaction Coordinate and Electronic Treatments in the QM/MM Reaction Dynamics of *Trypanosoma cruzi trans*-Sialidase

THESIS SUBMITTED TO THE

UNIVERSITY OF CAPE TOWN

IN FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

IAN LLOYD ROGERS

SUPERVISOR: PROFESSOR KEVIN J. NAIDOO

SCIENTIFIC COMPUTING RESEARCH UNIT

THE DEPARTMENT OF CHEMISTRY

JULY 2016

Abstract

The free energy of activation, as defined in transition state theory, is central to calculating reaction rates, distinguishing between mechanistic paths and elucidating the catalytic process. Computational free energies are accessible through the reaction space that is comprised of the conformational and electronic degrees of freedom orthogonal to the reaction coordinate. The overarching aim of this thesis was to address theoretical and methodological challenges facing current methods for calculating reaction free energies in glycoenzyme systems. Tractable calculations balance chemical accuracy and sampling efficiency that necessitates simplification of these complex reaction spaces through quantum mechanics/molecular mechanics partitioning and use of a semi-empirical electronic method to sample an approximated reaction coordinate. Here I directly and indirectly interrogate both the appropriate levels of sampling as well as the accuracy of the semi-empirical method required for reliable analysis of glycoenzyme reaction pathways.

Free Energies from Adaptive Reaction Coordinates Forces, a method that builds the potential of mean force from multiple iterations of reactive trajectories, was used to construct reaction surfaces and volumes for the glycosylation and deglycosylation reactions comprising the *T. cruzi trans*-sialidase catalytic itinerary. This enzyme was chosen for the wealth of experimental data available for it built from its significance as a potential drug target against Chagas disease. Of equal importance is the identification of an elimination reaction competing with the primary transferase activity. The identification of this side reaction, that is observable only in the absence of the *trans*-sialidase or sialic acid acceptor, presented the opportunity to study the means by which enzymes selectivity bias in favor of a single reaction path. I therefore set out to explore the molecular details of how *T. cruzi trans*-sialidase asserts a precision and selectivity synonymous with enzyme catalysis.

The chemical nature of the transition sate, formally defined as a dividing hypersurface separating the reactant and product regions of phase space, was characterized for the deglycosylation reaction. More than 40 transition state configurations were isolated from reactive trajectories, and the sialic acid substrate conformations were analyzed as well as the substrate interactions with the nucleophile and catalytic acid/base. A successful barrier crossing requires that the substrate pass through a family of E_5, $^4H_5$ and $^6H_5$ puckered conformations, all of which interact slightly differently with the enzyme.

This work brings new evidence to the prevailing premise that there are several pathways from reactant to product passing through the saddle and successful product formation is not restricted to the minimum energy path. Increasing the reaction space with use of a multi-dimensional (3-D) reaction coordinate allowed simultaneous monitoring of the hitherto unexplored competition between a minor

elimination reaction and the dominant displacement reaction present in both steps of the catalytic cycle. The dominant displacement reactions display lower barriers in the free energy profiles, greater sampling of favorable reactant stereoelectronic alignments and a greater number of possible transition paths leading to successful crossing reaction trajectories. The effects on the electronic degrees of freedom in reaction space were then investigated by running density functional theory reactive trajectories on the semi-empirical free energy. In order to carry out these simulations Free Energies from Adaptive Reaction Coordinates Forces was ported as a Fortran 90 library that interfaces with the NWChem molecular dynamics package. The resulting B3LYP/6-31G/CHARMM crossing trajectory provides a molecular orbital description of the glycosylation reaction. Direct investigation of the underlying potential energy functions for B3LYP/6-31G(d), B3LYP/6-31G and SCC-DFTB/MIO point to the minimal basis set as the primary limitation in using self-consistent charge density functional tight binding as the quantum mechanical model for modeling of enzymatic reactions transforming sialic acid substrates.

Declaration

I declare that the work in this thesis: 'Measuring the Effects of Reaction Coordinate and Electronic Treatments in the QM/MM Reaction Dynamics of *Trypanosoma cruzi trans*-Sialidase' is based on research performed at the Scientific Computing Research Unit, Department of Chemistry, University of Cape Town, South Africa. No part of this thesis has been submitted elsewhere for any other degree or qualification. The work in this thesis is my own, unless otherwise stated in the text.

Ian Rogers

July 2016

Acknowledgements

Firstly, I would like to express my sincere gratitude to Professor Kevin Naidoo, who provided valuable guidance for this dissertation and for helping me shape much of the material presented here for publication.

I thank Dr Martin Field for hosting me at Institut de Biologie Structurale (IBS).

I would like to thank the SCRU group in general for their collegiality, the coffee club for light-hearted conversations and thought-provoking debate, and I am indebted to Dr Chris Barnett, Dr Gerhard Venter and Dr Karl Wilkinson for their role in reviewing this thesis.

I am also grateful to the NRF and UCT for the funding provided to me, as well as to the Centre for High Performance Computing for the availability of computational resources.

Lastly, to my loving family and partner for their patience and unwavering support.

Abbreviations

CAZy: Carbohydrate-Active enZymes Database

CHARMM: Chemistry at Harvard Macromolecular Mechanics

CF3MU-SA: 4-trifluoromethylumbelliferyl sialic acid

CV: collective variable

DFT: density functional theory

SA: *N*-acetylneuraminic acid, sialic acid

DANA: 2-deoxy-2,3-didehydro-*N*-acetylneuraminic acid

fs: femtosecond

FEARCF: Free Energies from Adaptive Reaction Coordinate Forces

FEV: free energy volume

FES: free energy surface

Gal: galactose

GHO: generalized hybrid orbital

HF: Hartree Fock

KIE: kinetic isotope effect

LCAO: linear combination of atomic orbitals

MD: molecular dynamics

MM: molecular mechanics

NAC: near attack conformation

NAO: natural atomic orbital

NBO: natural bond orbital

PES: potential energy surface

PMF: potential of mean force

PNP-SA: *p*-nitrophenyl sialic acid

ps: picosecond

NR: Newton-Raphson

QM: quantum mechanics

RMSD: root-mean-square deviation

SCC-DFTB: self-consistent charge density functional tight binding

SCF: self-consistent field

TS: transition state

TST: transition state theory

VTST: variational transition state theory

WHAM: weighted histogram analysis method

Table of Contents

1 Reaction Dynamics Techniques for Calculating Enzymatic Free Energies of Activation 1

1.1 Theories of Enzyme Catalysis 2

1.2 Understanding Enzyme Catalysis through Transition State Theory 5

1.3 Enzyme Reaction Paths and Free Energies of Activation 9

1.4 Challenges of Exploring Enzyme Catalysis using Free Energy Methods 16

1.5 *Trypanosoma cruzi trans*-Sialidase 20

1.6 References 28

2 Aims and Objectives 32

3 Treating Enzyme Reaction Spaces Using DFT/MM and SCC-DFT/MM Methods 34

3.1 Treating Exchange and Correlation with DFT 34

3.2 Formulation and Approximations of SCC-DFTB 45

3.3 QM/MM Potential Energy 50

3.4 References 53

4 Simulating Enzyme-Catalyzed Reactions with FEARCF 56

4.1 Flat Histogram Methods 56

4.2 FEARCF 60

4.3 Parallelized FEARCF Reaction Dynamics Trajectories 64

4.4 References 67

5 Profiling Transition State Configurations on the *Trypanosoma cruzi trans*-Sialidase Free Energy Reaction Surface 69

5.1 Computational Methods 71

5.2 Results and Discussion 76

5.3 Concluding Remarks 81

5.4 References 82

6 Multi-dimensional Reaction Dynamics Reveal How the TcTS Enzyme Suppresses Competing Side Reactions and their Side Products 85

6.1 Computational Methods 87

6.2 Results and Discussion 90
6.3 Concluding Remarks 99
6.4 References 100
7 Evaluation of the Electronic Treatment of the TcTS QM Region 102
7.1 Theory and Computation 105
7.2 Computational Details 108
7.3 Results and Discussion 109
7.4 Concluding Remarks 116
7.5 References 117
8 Conclusion 119
Appendix A: Simulation Methods 123
Explicit Solvation Models 123
Treating Long-range Electrostatic Interactions 125
Integrating the Equations of Motion 127
Appendix B: Guthrie-Jencks Mechanistic Nomenclature 129
Appendix C: Analysis Methods and Conditions 131
TS Volume Analysis 131
QM/MM TS Optimization 131
Natural Bond Orbital Analysis 132
Appendix D: Results 133
RMSD Analyses for TcTS Equilibration MD Trajectories 133
SA C-4 OH Hydrogen Bond 134
References 135

1 Reaction Dynamics Techniques for Calculating Enzymatic Free Energies of Activation

Reactive trajectories are rare events initiated from molecular species in equilibrium. The kinetic energy of the reacting molecules under thermal motion is transferred through molecular collisions into structural deformation. As a result, the internal energy increases due to structural strains and distortions, as well as loss of translational, rotational and vibrational degrees of freedom. Provided kinetic energy is injected into the appropriate vibrational modes, the highest energy point along the reactive trajectory is reached. The molecular structure at this point is known as the activated complex and its passage to product is related to the vibrational motion along reaction progress (transition coordinate).[1] As such, the activated complex has a lifetime not longer than that of the vibration (typically 10^{-13} to 10^{-14} s).[2] These rare events can proceed along many different trajectories that depend on starting conformations, initial momenta and individual vibrational states of the reacting species.[3,4] For example, different angles of alignment between the nucleophile and departing group in a nucleophilic attack would lead to different energies for the activated complex. The result is an ensemble of reaction paths through phase space, and the rate of reaction is observed as a weighted average for all the possible paths.[1,5]

Evolutionary pressure to optimize particular chemical pathways for survival has led to the expression of enzymes that catalyze reactions with remarkable speed, specificity and selectivity.[6] In a general sense, enzymes perform this role by establishing an equilibrium with the substrate and influencing the reaction to follow a more favorable pathway to the product state that undergoes equilibrium dissociation. The alternate pathway offered by the enzyme active site can be a different chemical mechanism (for example, using a general base instead of water as a base, or involvement of covalent catalysis), and will eminently provide an optimized environment to reduce the energy barrier associated with the original reaction.[7] Precisely how the latter is achieved is the matter of some debate and sits at the frontier of computational biology. Elucidating enzyme activity holds exciting potential for rational design of transition state analog-based inhibitors,[2] rational approaches to enzyme engineering,[8] and catalyzing synthetic reactions.[9,10] However, the gestalt of enzyme activity arises from a complexity that has made it difficult to delineate a complete understanding of the factors that govern the observed rate enhancements and selectivities.[11]

This thesis addresses the use of computational reaction dynamics methods to investigate the catalytic action of a carbohydrate active enzyme, *Trypanosoma cruzi trans*-sialidase. The purpose of this chapter

is to introduce the broad topic of calculating generalized free energy of activation profiles for enzyme-catalyzed reactions. Chapter 2 will lay out the aims and objectives of the thesis and the methods integral to achieving them will be described in detail in Chapters 3 and 4. In the following sections, the current understanding of enzyme catalysis is outlined (Section 1.1), whereupon the phenomenological understanding of enzyme catalysis within the framework of transition state theory (TST) is detailed (Section 1.2). Following the discussion of this foundational theory, the methods used to access the enzyme reaction paths and transition states are briefly overviewed (Section 1.3) along with the current challenges facing the computation of enzymatic reaction free energies (Section 1.4). It is within this context that *T. cruzi trans*-sialidase is introduced in Section 1.5 – both as an enzyme system that has been well characterized by experimental work due to its importance to *T. cruzi* (the parasite responsible for Chagas disease), and as a model system for exploring enzyme selectivity.

1.1 Theories of Enzyme Catalysis

1.1.1 Rate Enhancement in Enzyme-Catalyzed Reactions

Differential binding to stabilize the activated complex is a principle that underlies the rate enhancement achieved in many enzymatic reactions (Figure 1.1A). It is understood that in stabilizing the activated complex the reaction energy barrier is reduced and this is associated with a faster observed turnover. Preferential stabilization of the activated complex is a complicated phenomenon, since the reactant is necessarily sequestered from solution. Therefore, the reactant must be bound, but resembles the activated complex to some degree.[1,12] Additional binding interactions, with favorable enthalpic or entropic contributions, must act on the activated complex to a larger extent than the substrate. These stabilizing interactions are formed through geometric and electrostatic complementarity within the enzyme active site. The electrostatic stabilization of the fleetingly existent activated complex has been highlighted as the predominant means of rate enhancement.[13] Accordingly, the active site microenvironment significantly alters the reactivity of key catalytic groups. Here, bulk solvents that tend to screen electrostatic interactions are excluded while charges and hydrogen bond donor/acceptor groups are arranged proximal to the bound substrate. An electrostatic field is thus created that stabilizes charges developing along a reactive trajectory.

The idea of substrate activation[14] (Figure 1.1B) goes 'hand-in-hand' with preferential stabilization of the activated complex. On binding to the active site that is designed to be complementary in structure and electronic characteristics to the activated complex, the substrate may distort in order to optimize its interactions with the enzyme. The introduced steric and/or electronic strain causes the substrate to adopt a high-energy conformation with increased reactivity compared to the substrate alone. If this strain acts in the direction of the activated complex, and if the strain is relieved upon reaching the

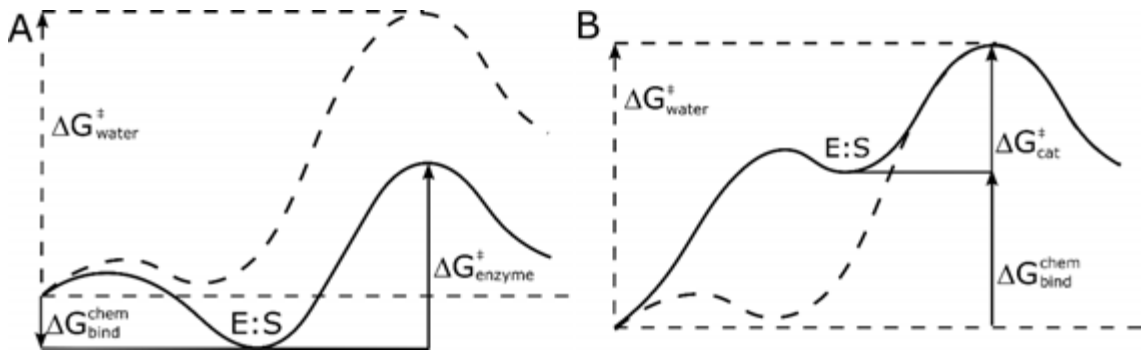


Figure 1.1 Illustrations explaining the free energy changes associated with (A) transition state stabilization and (B) reactant state destabilization. In the latter case, the energy cost of distorting the reacting substrate is 'paid' in the binding step. Figure adapted from Warshel et al.[15]

high-energy structure, the substrate activation can be viewed as a ground-state mechanism of catalysis.[1] The primary distinction between differential binding and reactant destabilization is that the enthalpic cost of destabilizing the reactant is 'paid' in the binding step.[13] Provided that the enzyme displays a Michaelis constant value (K_m, an inverse measure of the substrate's affinity for the enzyme notated) near natural concentration of the substrates, the rate-determining step will rather be the chemical step.

The spatial-temporal approach is a closely related view of enzyme catalysis.[16,17] This theory is based on the premise that the observed reaction rate is proportional to the time that the reactants reside within a critical distance and orientation. That is to say, the longer the reactants spend in the correct geometry, the greater the probability that the vibration that defines the transition coordinate is excited to a high enough level to achieve passage through the activated complex. An upper bound of $T\Delta S^{\ddagger}$ = 4.6 kcal/mol (rate enhancement of 2×10^3) has been extrapolated for this phenomenon at 25 °C by comparing a series of reactions that have directly comparable inter- and intramolecular routes.[1] This effect has been defined using the concept of effective molarity: the concentration needed to give a pseudo first-order rate constant that is identical to the first-order rate constant for the intramolecular reaction.[1] Again, it is useful to distinguish between the kinetic and thermodynamic steps of enzyme catalysis and note that the cost of unfavorable entropy changes associated with sequestering the freely moving reactant is a binding phenomenon.

Near attack conformation (NAC, Figure 1.2)[18-20] is a unifying theory that attempts to fully cover the methods of enzyme catalysis. This approach uses geometric criteria to define an optimally aligned reactant state. This structure then represents the turnstile through which the equilibrated ground state must pass to achieve the activated complex. Hindrances to forming the NAC structure include the entropic considerations of bringing the reactants together accompanied by the steric strain due to the proximity of the bond-forming atoms. Therefore, according to NAC theory, the rate constant can

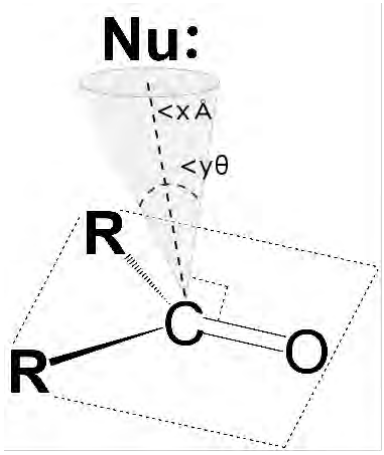


Figure 1.2 Illustration of the concept of a NAC structure for a nucleophilic step, adapted from Bruice and Lightstone.[18] The structure is defined by a distance between the nucleophile and electrophile, and variation of angle of attack from 90°. Enzymatic rate acceleration is achieved by increasing the concentration of reactant states optimally aligned for nucleophilic attack.

be correlated to (i) the mole fraction of the reactant present as NACs, (ii) the difference in solvation between the reactant complex in the NAC and solvation of the NAC without a catalyst, and (iii) the electrostatic binding forces that can stabilize the activated complex.

Finally, it has been pointed out that the changes to the protein conformation are key to the understanding of the drivers that evolve the reaction.[21] The roles of protein dynamics and conformational selection are being actively investigated to explain enzyme catalysis. There is a lot of debate, both substantive and semantic,[22] that surrounds the role of dynamic motions in enzymes. While one camp emphasizes electrostatic stabilization as the sole primary catalytic effect,[13,23-26] the other posits that these protein motions enhance transit from reactant to product.[3,11,27-31] Specifically, it is proposed that fast (femtosecond), protein-promoting motions dynamically modulate the shape of the potential energy profile and drive certain types of reactions (for example, proton transfer reactions dominated by tunneling). On the other hand, long-time scale dynamics (up to milliseconds) have been associated with sampling of the protein conformational landscape to achieve sub-states that optimize enzyme-substrate interactions.[6,31]

1.1.2 Selectivity in Enzyme-Catalyzed Reactions

Reaction stereo- and regioselectivity dovetails with the principles of rate acceleration in enzyme catalysis. The reactivity-selectivity principle suggests that the stabilization of activated complexes and reactive intermediates leads to mitigation of side products and improved selectivity.[1] In the extreme case, enzyme participation progresses to covalent catalysis. This is seen in retaining hydrolases and some retaining glycosyltransferases,[32] where a double-displacement reaction ensures the desired stereochemical product. The enzyme can also create an optimized spatial and chemical environment to lower the activation energy of one reaction relative to competing reactions, and so enhance the

kinetic favorability of the primary product. Finally, spatial-temporal tuning ensures optimal alignment of reacting atoms that leads to strict regioselectivity.[1]

1.2 Understanding Enzyme Catalysis through Transition State Theory

Investigating which of the phenomena outlined above drive enzyme catalysis is a difficult challenge for experimental techniques. The observed catalytic action arises from complex kinetics that encompass an ensemble of possible reactive trajectories. The short time and length scales, in the presence of high noise to signal ratios, make it impossible for current and near-future experimental techniques to extract molecular-level details during the progress of a chemical reaction. Various attempts to follow the evolution of derivative reactions using kinetic isotope effect (KIE) experiments of modified substrates are being used to extract clues about reaction mechanisms.[33] However, these experimental techniques performed in complex solvents and molecular environments are unable to provide clear mechanistic pathways between natural substrates and activated complexes. The difficulty of interpreting experimental mutant studies has also been highlighted.[6] On the other hand, computational tools provide atomic and femtosecond resolution that can isolate activated complexes,[34-36] and TST uses the quasi-thermodynamic concept of the free energy of activation, $\Delta G^{\ddagger}$, to connect microscopic computational simulations to experimental rates of reaction.[37] $\Delta G^{\ddagger}$ is the free energy difference between the reactants and the transition state (TS), where the TS is defined as an infinitesimally thin hypersurface that separates the reactant regions of phase space from product regions.

Canonical TST[34] considers the 6*N*-dimensional space of a system for which all points are populated according to a Boltzmann distribution. Reactive motion takes a phase point from the equilibrated reactant through a sparsely populated (6*N* - 1)-dimensional dividing hypersurface to the unpopulated product region. The missing degree of freedom is the vibrational mode that transits points in phase space from TS to product and is known as the reaction coordinate. If it is assumed that there is no recrossing of the TS dividing surface, then the one-way flux (chemiflux) through the dividing surface is the same as the rigorous case where both reactants and product regions of phase space are at equilibrium.[5] However, the net flux will be zero because the system is assumed to be at equilibrium.[5] Finally, the free energy of activation is calculated as the classical barrier height between the TS and the equilibrium energy of reactants. Within the context of a classical system, enzymatic rate acceleration can be explained by reduction of $\Delta G^{\ddagger}$ such that there is a greater number of configurations at the TS compared to the uncatalyzed system at the same temperature. This is surmised to ultimately result in a faster reaction rate for the chemical step.

The equilibrium constant for the quasi-equilibrium between the reactants and the TS is used to quantitatively relate $\Delta G^{\ddagger}$ to the reaction rate. Consider a bimolecular reaction that proceeds through an activated complex $[AB]^{\ddagger}$:

$$[A] + [B] \rightleftharpoons [AB]^{\ddagger} \rightleftharpoons [C]$$

Equation 1.1

The chemiflux out of the reactant, equal to the rate constant when all concentrations are unity,[5] is calculated as the derivative with respect to time:

$$\frac{d[C]}{dt} = k[A][B] = k^{\ddagger}[AB]^{\ddagger}$$

Equation 1.2

Now, assuming a quasi-equilibrium between the reactants and TS, the concentration of the TS can be found in terms of the concentrations $[A]$ and $[B]$:

$$[AB]^{\ddagger} = K^{\ddagger}[A][B]$$

Equation 1.3

where the equilibrium constant, $K^{\ddagger}$, can, in analogy to a true equilibrium, be expressed as:

$$K^{\ddagger} = \left(\frac{k_B T}{h\upsilon}\right) e^{\frac{-\Delta G^{\ddagger}}{RT}}$$

Equation 1.4

In Equation 1.4, k_B, h, R and T are, respectively, Boltzmann's and Planck's constants, the gas constant and the temperature. Substituting Equation 1.3 and Equation 1.4 into Equation 1.2 and assuming concentrations of unity gives:

$$k = k^{\ddagger}\left(\frac{k_B T}{h\upsilon}\right) e^{\frac{-\Delta G^{\ddagger}}{RT}}$$

Equation 1.5

Finally, if it is assumed that $k^{\ddagger}$ is related to the vibrational frequency, υ, for the mode along reaction progress, and a transmission coefficient, κ, is used to correct for activated complexes that do not proceed to product due to misaligned atoms or interference of rotational states, we obtain the Eyring equation:[36]

$$k = \kappa\left(\frac{k_B T}{h}\right) e^{\frac{-\Delta G^{\ddagger}}{RT}}$$

Equation 1.6

1.2.1 Validity of TST

TST makes two fundamental assumptions. In the first, the TS forms a quasi-equilibrium with the reactant state (Equation 1.3). The free energy of activation is a generalization of the concept of free energy of reaction (ΔG), which is rigorously defined for a true reactant ⇌ product equilibrium, where each of the reactant and product states locally reflects a Boltzmann distribution of energies. Analysis of such a system describes the equilibrium constant $K = e^{-\Delta G/RT}$, and ΔG can be calculated as a one-to-one mapping. The same equality is not strictly true for $\Delta G^{\ddagger}$ because the fleeting existence of the activated complexes does not allow for redistribution of energy among the rotations and vibrations to reflect a Boltzmann distribution of states. It can be rationalized that the assumed local Boltzmann distribution is an approximation of the actual energy distribution for the ensemble of activated complexes arising from different reactive trajectories.[1] This distribution should be most populated at the lowest energy activated complex, with transit points becoming less probable as a function of the distance from this minimum. Indeed, the validity of TST is supported by the correlation observed between calculated barriers and experimental rate constants for some enzymes.[37]

However, there are reactions for which the assumption that all points along the reaction path are in thermal equilibrium is not appropriate. For example, when a high-energy intermediate with low barriers to alternate products exists, the relative differences in barrier $\Delta G^{\ddagger}$ values may not explain selectivity.[38] If bond cleavage or formation at the intermediate occurs with a rate comparable to the timescale for redistribution of its vibrational motions, the entrance channel may possess a better 'dynamic match' to a specific exit channel. In this case, molecular motions will preferentially direct the system to one of the products, where both might be equally favorable if the intermediate had sufficient lifetime to allow all internal energy to randomize. Similarly, a single TS may lead to several isomeric products via bifurcation of the reaction pathway. In that case the selectivity is also controlled by the distribution of energy in its vibrational modes. It has been postulated that the electrostatic environment of the active site steers the reaction pathway towards the natural product.[39,40]

The second fundamental assumption of TST is that there is no recrossing of the dividing surface. This is only true if the motion in the reaction coordinate is globally separable. If this condition is met, the reaction coordinate is uncoupled to other degrees of freedom. In other words, motion along the reaction coordinate at the TS is independent of the values of the other coordinates.[5] Although the reaction-coordinate-separability assumption is often a defensible one, it does not always hold, and the accuracy of TST is limited by the extent of recrossing.[37] Advances in TST have focused on the transmission coefficient that corrects for the breakdown of this assumption.[5] In variational transition state theory (VTST), the transition state hypersurface is variationally optimized to minimize the one-

way flux coefficient. This effectively optimizes the reaction coordinate to minimize the effects of recrossing.[41] Practically, the transmission coefficient can also include semi-classical approximations to account for tunneling and other non-equilibrium contributions in the motion along the reaction coordinate (VTST/MT) and so derive a quasi-classical rate constant from the classical version written in Equation 1.6. VTST and VTST/MT have been extended to enzyme-catalyzed reactions through ensemble averaged TST, where the recrossing factor needs to be calculated for each configuration in the TS ensemble generated during the reaction free energy calculation.[42,43]

1.2.2 The Reaction Coordinate, Progress Variable and Reaction Space

The concept of the TS as a dividing surface orthogonally intersecting the reaction coordinate has been introduced. The TS is practically located by monitoring the energy changes along a progress variable, that is the signed distance along the reaction path. The reaction coordinate and the progress variable are locally the same if the TS orthogonally intersects the reaction path, and the progress variable is used interchangeably with the reaction coordinate in the literature without confusion.[5] That convention is adopted here and the reaction coordinate is notated as ξ, unless otherwise specified.

Usually, the reaction coordinate is deconvoluted into a *n*-dimensional order parameter $\xi(\xi_1, \xi_2, \cdots)$ that should include all important contributions to the mode transiting points in phase space from TS to product. Energy changes along the reaction coordinate are monitored by sampling the reaction phase space comprising the electronic and conformational degrees of freedom orthogonal to the order parameter. The treatment of the reaction phase space (or simply reaction space) is dependent on the Hamiltonian employed, the sampling scheme used and the temperature simulated. For the *n*-dimensional ξ, when the reaction space is sampled at 0 K through geometry optimizations, we retrieve a reaction potential energy surface (PES) if *n*=2 and a potential energy volume if *n* = 3. Likewise, by including thermal fluctuations using enhanced sampling techniques discussed later in Section 1.3.3, the reaction free energy surface (FES) and reaction free energy volume (FEV) can be calculated. The free energy calculated as a function of ξ is also generally known as the potential of mean force (PMF).

The definition of ξ can be very general, such as the energy gap between the reactant and the product electronic potential energy surfaces,[44] but an intuitive and convenient choice is a geometric coordinate composed from a number of variables drawn from an internal coordinate system, such as bonds or angles.[5] For example, the length of the breaking bond can be subtracted from the length of the forming bond ($\xi = \xi_{forming} - \xi_{breaking}$), so that an increasing value of ξ represents reaction progress. Use of the energy gap between reactant and product valence bond states has been compared to the use of a geometrical reaction coordinate in dihydrofolate reductase.[45] It was concluded that use of either reaction coordinate in computing reaction free energies is expected to give similar results.

1.3 Enzyme Reaction Paths and Free Energies of Activation

The TS is quantitatively related to a reaction rate within the framework of TST through the quasi-thermodynamic concept of the free energy of activation, $\Delta G^{\ddagger}$. In this way models can be validated against experimental rate constants, k_{cat}, and can be used to distinguish between mechanisms.[37] Once the enzyme system has been prepared from an experimental starting point, various approaches can be pursued to access the free energy of activation. The techniques adopted will depend on the focus of the study: whether only analysis of stationary points is desired or a complete reaction path is sought.

1.3.1 Preparing the Enzyme System for QM/MM Reaction Dynamics

The enzymatic reaction space is often simplified using a quantum mechanics/molecular mechanics (QM/MM) partitioning that is discussed in more detail in Chapter 3. Essentially, the electron distribution changes within the catalytic site are modeled using quantum mechanics (QM) and the effect of the polarizing protein environment is included using a classical molecular mechanics (MM) force field. A good experimental starting point for QM/MM simulations is a well-resolved crystal structure. The enzyme co-crystallized with the substrate (Michaelis complex) can be obtained by mutation of catalytic residues/substrates. If the Michaelis complex crystal structure is not available, the reactant state can be derived using structural superposition of binary complexes or molecular docking. The starting structure is prepared by reverting mutations and/or substrate modifications, adding missing atoms/residues, and protonating titratable side chains. The system is then solvated before the Hamiltonian is defined and used to minimize any steric clashes.

It is important to examine and thermally equilibrate the protein crystal structure prior to production simulations for two reasons. Firstly, errors in structural parameters are introduced by crystal effects, as well as by ignoring correlated motion in the refinement process. Typically, a single, average structure is obtained by taking time and ensemble averages of the many heterogeneous molecules in the crystal.[32] These errors may be significant when studying reactions that rely on accurate modeling of atom positions. Secondly, the crystal structure needs to be in a realistic microscopic state. Equilibration is achieved using a molecular dynamics (MD) or Monte Carlo scheme to sample the potential energy surface. In this research, classical MD trajectories were simulated by solving the differential equations embodied in Newton's second law (see Appendix A for more detail):

$$\frac{d^2 x_A(t)}{dt^2} = \frac{\boldsymbol{F}_A\{x(t)\}}{m_A}$$

Equation 1.7

In this way, MD methods generate a collection of microscopic replications of the system (an ensemble) that belong to a particular thermodynamic state and are connected in time. The ensemble is characterized by the parameters of the system held constant during the dynamics. Conditions of a constant number of particles, pressure and volume (NPT; Isothermal-isobaric ensemble), or a constant number of particles, volume and temperature (NVT; Canonical ensemble), correlate most closely to experimental conditions.

Structural and energetic properties of the system are monitored during initial dynamics of the system, and representative snapshots are analyzed to test for equilibration. Typically, the protein root-mean-square deviation (RMSD) should converge. A representative microscopic state is used as the starting point for QM/MM simulations to determine the enzyme reaction mechanism. If long time-scale transitions are present in the equilibrium ensemble, either the catalytically competent structure should be chosen or else slowly interconverting sets of reactant states should be treated as separate reactants in a multi-species mechanism.

1.3.2 Calculating the Minimum Energy Reaction Path

The reaction path on the potential energy is estimated as a minimum energy path (MEP): the lowest energy path for a rearrangement of a group of atoms from reactant to product passing through the TS structure. The TS structure is the first-order saddle point on the potential energy. It represents the minimum of the TS dividing surface that is located at the intersection of the TS and the reaction coordinate. When a 2-D reaction coordinate is employed to calculate the PES, the available entropy from the orthogonal degrees of freedom (primarily influenced by bond strengths that determine vibration) defines the reaction energy landscape (widths of the ‘valleys’ and ‘saddles’ by analogy to geographical features).

The simplest approach to searching the reaction path is to scan a discretized reaction coordinate optimizing the orthogonal degrees of freedom at each point. Reaction coordinate driving requires only the gradient and works well if the reaction can be described with a 1-D or 2-D reaction coordinate. However, when the potential energy landscape becomes complex, the scheme may miss the saddle point configuration when relaxing the remaining degrees of freedom (susceptible to drag hysteresis effects[46]). Chain-of-states methods, such as the nudged elastic band[47] algorithm and conjugate peak refinement, provide an alternative method of estimating the reaction path as a series of points interpolating the reactant and products. Chain-of-states methods hold the advantage of being able to treat multiple degrees of freedom at 0 K since there is no need for a pre-defined reaction coordinate.

TS structures obtained using only the potential gradient can be refined using local optimization methods that explicitly calculate the Hessian. Then, with the TS structure in hand, the reaction path

(known as the intrinsic reaction coordinate) may be traced by steepest descent in mass-weighted coordinates to the reactant and product. The most popular local optimization routines include rational function optimization[48,49] and the quasi-Newton-Raphson (quasi-NR) methods. The Hessian eigenvector with the lowest imaginary frequency is selected as the reaction coordinate, and the system's energy is maximized along this mode while the energy in all other directions is minimized. An approximate NR method implemented in NWChem[50] is used in this thesis to relax conformations extracted from semi-empirical reaction dynamics (self-consistent charge density functional tight binding [SCC-DFTB]) onto the density functional theory (DFT) potential energy.

Local methods such as quasi-NR require a good estimate of the TS in order to converge, since the initial uphill step is only appropriate if the local potential function, V, has a negative curvature. The NR scheme expands V in a Taylor series, truncated after the second-order term. For step $i+1$:

$$V(\boldsymbol{R}_i + \Delta\boldsymbol{R}_i) = V(\boldsymbol{R}_i) + \boldsymbol{g}_i^T(\Delta\boldsymbol{R}_i) + \frac{1}{2}\Delta\boldsymbol{R}_i^T\boldsymbol{H}_i\Delta\boldsymbol{R}_i$$

Equation 1.8

where $\Delta\boldsymbol{R}_i = \boldsymbol{R}_{i+1} - \Delta\boldsymbol{R}_i$ is the displacement vector to the new geometry, $\boldsymbol{g}$ (or gradient) is the first derivative of V with respect to nuclear coordinates and $\boldsymbol{H}$ is the second derivative and is known as the Hessian. Assuming a quadratic V, the saddle point, for which $\boldsymbol{g}_i = 0$, can be found in a single step as

$$\Delta\boldsymbol{R}_i = -\boldsymbol{H}^{-1}\boldsymbol{g}$$

Equation 1.9

In the coordinate system $(\boldsymbol{R}')$ where $\boldsymbol{H}$ is diagonal,

$$\Delta\boldsymbol{R}'_k = -\frac{\boldsymbol{f}_k}{\varepsilon_k}$$

Equation 1.10

In Equation 1.10, $\boldsymbol{f}_k$ is the projection of the gradient along each Hessian eigenvector with eigenvalue ε_k. If only one of the eigenvalues is negative, the optimization step detailed above will be along the gradient towards the saddle point. A suitable shift parameter, λ, is chosen in eigenvector following schemes so that that the step component in the reaction mode is guaranteed to be along the gradient:

$$\Delta\boldsymbol{R}'_k = -\frac{\boldsymbol{f}_k}{\varepsilon_k - \lambda}$$

Equation 1.11

Since the local potential function is not realistically quadratic, optimization is an iterative process. At each step the Hessian can be updated in an approximate fashion (quasi-NR), using the gradient and

displacement vectors from the previous steps. This is sufficient while the coordinates have not changed considerably, but a new, non-approximate Hessian should be generated if the optimization does not converge quickly.

Thus, the stationary points for a single protein structure can be found using one of the path finding/ optimization schemes introduced above, and the entropic contributions can be included[51-54] within the framework of the rigid-rotor, harmonic oscillator approximation.[55] The free energy differences can then be related to experimental values using TST. While this strategy has the advantage that it can be used together with high-level *ab initio* methods, it also has several limitations. Firstly, this treatment provides an incomplete characterization of the reaction since the rigid-rotor, harmonic oscillator approximation can only be applied to the stationary points along the reaction coordinate. Secondly, a single path on the potential energy neglects the effects of multiple protein conformations along the enthalpic pathway and across the barrier separating reactants and products. A protein progressing from a local minimum may not relax as the substrate advances toward product formation, which could result in an overestimated activation barrier. In addition, thermal fluctuations in the protein conformation significantly affect the FES through polarization of the QM region or accompanied changes in protein–ligand interaction mode.[56] Indeed, barriers and TS structures for different starting conformations can differ significantly.[37]

1.3.3 Calculating the Free Energy Reaction Path

In order to include the thermal fluctuation of the protein environment, multiple reaction trajectories need to be considered. There is a TS dividing surface intersecting the energy saddle point for each of the many possible reactive trajectories in condensed phase. Since it is not possible to find all the saddle points, statistical methods are used to find the 'generalized' TS located at the maximum of the PMF.[5]

1.3.3.1 Calculating the Generalized Free Energy of Activation Profile

The system's global free energy (the Helmholtz free energy for a closed thermodynamics system at constant temperature) is given by the well-known expression:

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

Equation 1.12

where Q_{NVT} is the partition function for the canonical ensemble. In analogy to the above expression, if a unidimensional reaction coordinate is considered, the free energy at each point on the PMF, $W(\xi_0)$, can be calculated from the integral over combinations of the nuclear coordinates, $\boldsymbol{R}$, that give the reference value of the reaction coordinate, ξ_0:[35,57]

$$W(\xi_0) = C - k_B T \ln\left(\int \delta(\xi(\boldsymbol{R}) - \xi_0) e^{-\beta V(\boldsymbol{R})} d\boldsymbol{R}\right)$$

Equation 1.13

where $\beta = 1/k_B T$ and terms in the partition function that are independent of the reaction coordinate have been subsumed into C. It is practically useful to obtain the expression for the ensemble average of the probability distribution function:[57]

$$\langle \rho\ (\xi_0) \rangle = \frac{\int \delta(\xi - \xi_0) e^{-\beta V} d\boldsymbol{R}}{\int e^{-\beta V} d\boldsymbol{R}}$$

Equation 1.14

It is clearly not possible to calculate the phase-space integrals present in Equation 1.14. However, as the sampling of an ergodic system increases, the ensemble average tends to the time average:[58]

$$\langle \rho\ (\xi_0) \rangle = \lim_{t \to \infty} \frac{1}{t} \int P[\xi_0(t)] dt$$

Equation 1.15

In Equation 1.15, every point in phase space is visited and P counts the frequency of occurrence of ξ_0 in a given interval that is of infinitesimal width. The reaction coordinate probability density can then be used to calculate the free energy at each point of the PMF:

$$W(\xi_0) = C' - k_B T \ln \mathrm{P}(\xi_0)$$

Equation 1.16

The problem is that simulations drawing from the Boltzmann distribution of energy states will preferentially sample low energy configurations and a reliable probability density will not be obtained. High-energy regions may contribute significantly to the free energy due to the exponential term of Equation 1.14. Therefore, a range of sampling techniques has been developed to enhance sampling of the unfavorable regions along the reaction coordinate. Methods that have been most commonly applied to enzyme reactions involve simulations that are run on the Born-Oppenheimer surface in the presence of an external biasing force. In this way reaction coordinate probability histograms are constructed with intervals of finite width, whereupon histogram reweighting methods,[59,60] such as the weighted histogram analysis method (WHAM),[61,62] are applied to retrieve an estimate of the unbiased free energy. These methods include restrained dynamics schemes that apply a pre-defined restraining potential to enhance sampling of unfavorable regions along the reaction coordinate, and flat histogram methods that calculate a biasing/driving potential on the fly to achieve equiprobable sampling of the reaction coordinate space.

1.3.3.2 Restrained Dynamics Methods

Umbrella sampling[63-65] is a widely-used restrained dynamics technique wherein a biasing potential is added to the system's Hamiltonian. Under the influence of the biasing potential, the ensemble average $\langle A \rangle$ of a quantity A is determined by:

$$\langle A \rangle = \frac{\langle A/e^{-\beta V^b} \rangle_b}{\langle 1/e^{-\beta V^b} \rangle_b}$$

Equation 1.17

where V^b includes the biasing potential in addition to the system Hamiltonian, and the brackets $\langle \cdots \rangle_b$ specify an ensemble average determined with the biased distribution function, ρ^b:

$$\rho^b = \frac{e^{-\beta V^b}}{\int e^{-\beta V^b} d\boldsymbol{R}}$$

Equation 1.18

If the negative value of an accurate PMF were to be used as the biasing potential, a uniform distribution of configurations along the reaction coordinate would be obtained, and an accurate free energy could be computed. However, because the PMF is not known before the start of the calculation, windowed potentials, w_i, are employed to restrain the system's reaction coordinates along the reaction path. The biasing potential is commonly a harmonic function:

$$w_i(\xi(\boldsymbol{R})) = \frac{K}{2}\left(\xi(\boldsymbol{R}) - \xi_0^i\right)$$

Equation 1.19

for which the total biased potential becomes:

$$V^b = V + w_i(\xi(\boldsymbol{R}))$$

Equation 1.20

The ensemble average of the probability distribution can be derived analogously to that obtained for the Boltzmann distribution function in Equation 1.14:

$$\langle \rho(\xi) \rangle_i^b = \frac{\int \delta\left(\xi - \xi_0^i\right) e^{-\beta V^b} d\boldsymbol{R}}{\int e^{-\beta V^b} d\boldsymbol{R}}$$

Equation 1.21

The bias in the exponential of the numerator is constant over the integral and can be separated out:

$$\langle \rho(\xi) \rangle_i^b = e^{-\beta w_i} \times \frac{\int \delta\left(\xi - \xi_0^i\right) e^{-\beta V} d\boldsymbol{R}}{\int e^{-\beta V^b} d\boldsymbol{R}}$$

Equation 1.22

The unbiased distribution needs to be recovered in order to calculate $W(\xi)$, and can be obtained as follows:

$$\langle \rho\ (\xi) \rangle_i = \langle \rho\ (\xi_0) \rangle_i^b \times e^{\beta w_i} \times \langle e^{-\beta w_i} \rangle$$

Equation 1.23

Computing the PMF this way, the non-Boltzmann simulation would be implemented via a biasing potential that comprises several windows along the reaction coordinate. Using WHAM, the full distribution function can be written as a weighted sum of the unbiased window distribution functions:

$$\langle \rho\ (\xi) \rangle = \sum_i^{windows} p_i\ (\xi) \times \langle \rho\ (\xi) \rangle_i$$

Equation 1.24

where the weights, p_i, are found in an iterative fashion to minimize the statistical error in the estimate function and are related to the last term in Equation 1.23, $\langle e^{-\beta w_i} \rangle$. Provided histograms from each window are suitably overlapping, one obtains an accurate estimate of the free energy. This often requires adapting the force constant of the restraint. Constructing a 2-D PMF using adaptive umbrella sampling typically involves an initial set of simulations. These are performed with the center of each harmonic function equally spaced along a line corresponding to the concerted path (though this path may be mechanistically prohibitive) connecting the reactant and product state. In the next set of calculations, the harmonic functions are restricted to lie along the minimum energy path of the PMF estimated from previous simulations.[66] The PMF is related to the unbiased probability distribution, P:

$$W(\xi_1, \xi_2, \cdots) = -\frac{1}{\beta} \ln P(\xi_1, \xi_2, \cdots)$$

Equation 1.25

1.3.3.3 Flat Histogram Methods

While umbrella sampling-type simulations use a biasing potential that is necessarily close to the quantity that they are attempting to determine, flat histogram methods[67,68] attempt to automatically search for the optimal biasing potential on the fly without prior knowledge of the free energy profile. This work employs a flat histogram method called the Free Energies from Adaptive Reaction Coordinate Forces (FEARCF)[69] that is discussed in detail in Chapter 4. Briefly, FEARCF introduces memory-dependence by including the gradient of the current PMF estimate as a driving force in an

iterative process. The FEARCF routine therefore consists of a number of iterations for which the PMF gradient from the previous iteration is applied to the reaction coordinate as a perturbing force driving the system 'uphill' to unsampled regions of phase space. An equal ratio of sampling of the global minimum compared with that of the TS implies a flat histogram in configuration space has been achieved, and by definition.

1.3.4 Rare Event Simulations

Rare event simulations are another class of reaction dynamics methods that can be used to determine enzyme reaction paths, but without direct access to $\Delta G^{\ddagger}$. Rare event algorithms sample small sketches of MD simulations that enhance the occurrence of reactive events. The transition path sampling scheme[70-72] conducts a Monte Carlo-type sampling of trajectory space. Sampling is initiated with the definition of a successful event, and an estimate of a reactive MD trajectory that may be prepared in an artificial manner such as high-temperature MD or geometric interpolation. Each Monte Carlo step can then be implemented using a shooting move algorithm applied to an intermediate configuration for the preceding reactive event. For example, the velocities for the snapshot in time can be modified such that the conservation of total momentum and energy is obeyed, and then propagated forward and backward in time.[73] The trajectory is monitored and if the path shows the reactive event it will be accepted. In this way transition path sampling iteratively scans the trajectory space for barrier crossings through low energy activated complexes, and convergence is tested by starting from several pathways. Transition path sampling makes no *a priori* judgments about degrees of freedom important in the transition process, and yields realistic reactive trajectories that are not biased by an external potential. However, there are technical challenges associated with implementing time-reversible MD.[73] Furthermore, the protocol to identify the TS ensemble is unclear and computationally expensive. In the transition path sampling study of chorismate mutase,[74] the TS ensemble was considered to be those points that have an equal probability of ending at the reactant or product state when initiated with random velocities. Such a structure needed to be found for each trajectory in the transition path sampling ensemble.

1.4 Challenges of Exploring Enzyme Catalysis using Free Energy Methods

The calculation of the reaction free energy profile and $\Delta G^{\ddagger}$ within the context of TST has been discussed up until now and is summarized in Figure 1.3. The illustration highlights the concept of a deconvoluted progress variable that is notated ξ' to distinguish it from the reaction coordinate that is formally the missing degree of freedom in the TS.

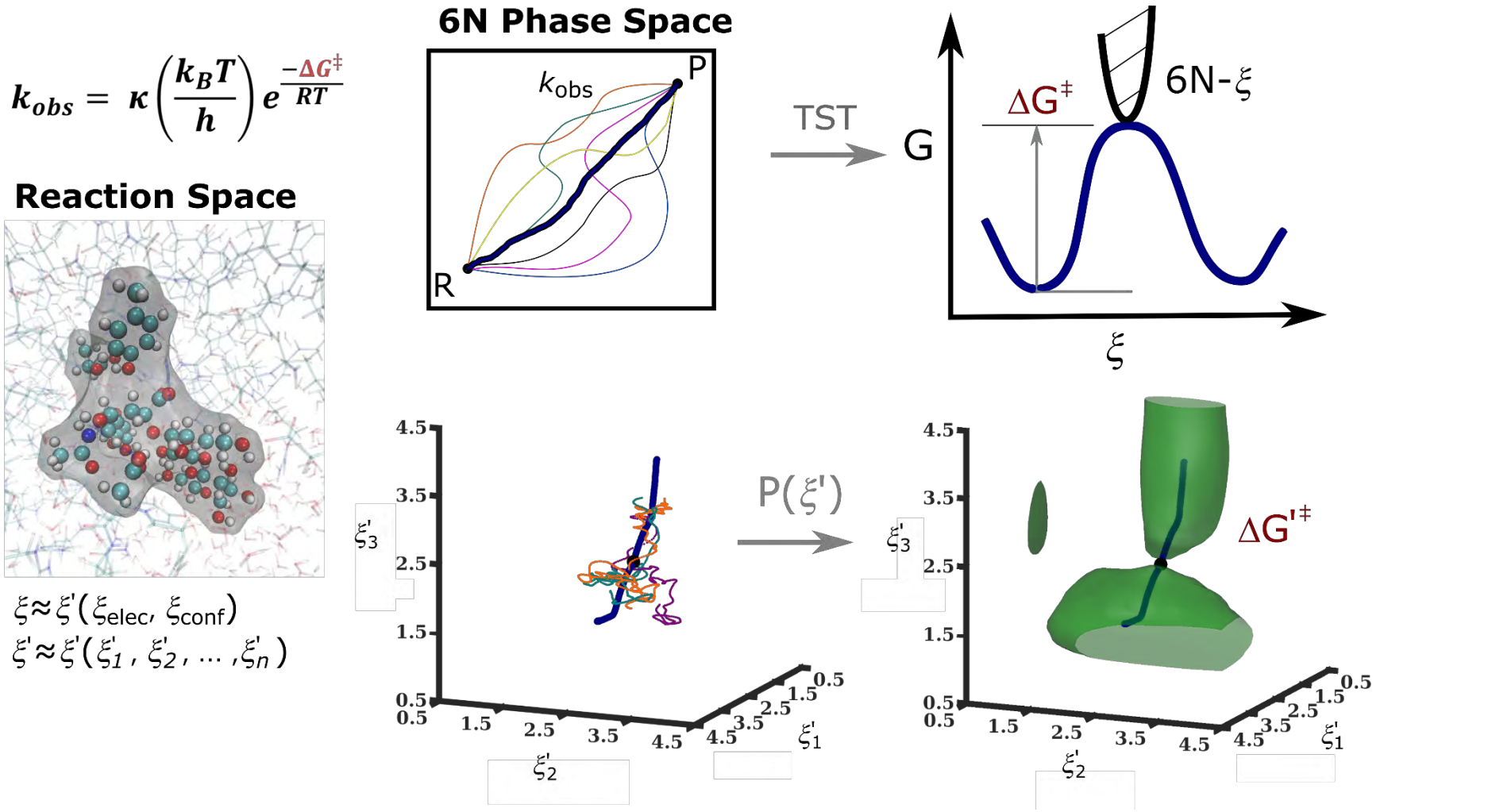


Figure 1.3 The relation of computational free energies to TST and the observed reaction rate is illustrated. The top pane shows that the reaction rate, observed as the result of many possible pathways, is related to the quasi-thermodynamic free energy of activation within the framework of TST. The most probable path is described by the reaction coordinate, ξ, which is the missing degree of freedom in the hypothetical dividing surface. In free energy methods, phase space is approximated by the reaction space defined by the QM/MM potential energy function and deconvolution of the reaction coordinate into a progress variable, ξ'. Averaging along the progress variable yields probability histograms that are mapped to free energy values. The prime notation has been used to clearly distinguish between the reaction and progress coordinates.

The progress variable includes important contributions to the mode transiting points from the TS to the product, and defines the reaction space orthogonal to it. A QM/MM Hamiltonian is used in enhanced sampling of the reaction coordinate space from which the unbiased reaction coordinate probability density $P(\xi')$ is retrieved. $\Delta G'^{\ddagger}$ can then be obtained from the PMF, $W(\xi_1', \xi_2') = -\frac{1}{\beta}\ln P(\xi_1', \xi_2')$, and is quantitatively related to the observed rate constant through the Eyring equation.

The development of free energy methods has been a boon for the study of enzyme reactions. Computational studies interact synergistically with experiments that provide the starting structures for molecular simulations as well as the macroscopic thermodynamic data against which models may be validated. Computation can interpret experimental findings on a fine spatial and temporal resolution to elucidate enzyme reaction mechanisms and can suggest new experiments. However, it is important to note that calculating enzymatic reaction free energies within the context of TST is not a black-box procedure due to theoretical and technical challenges that are now described. Thereafter, *T. cruzi* trans-sialidase is introduced as an apt enzyme system for addressing some of these challenges.

1.4.1 Defining the Reaction Coordinate

Deconvolution of the reaction coordinate requires an understanding of the biological system in order to define an order parameter that captures the reaction kinetics. Typically, enzyme reaction free energies are sampled along a 1-D or 2-D geometric reaction coordinate. The advantages of these low-dimensional free energy computations are rapid sampling, simplified analysis and convenient visualization using standard 1-D or 2-D graphing tools. However, such models are unable to comprehensively describe the complexities of enzyme catalyzed reaction mechanisms. This is because enzymatically catalyzed reactions are invariably dependent on a number of correlated events that need to be treated as components of a higher-dimensional reaction coordinate. The validity of the reaction-coordinate-separability assumption will be influenced by construction of the reaction coordinate if motions correlated to reaction progress are not included.[5] In addition, a poorly defined order parameter will yield an inaccurate representation of the TS orthogonal to it. To compensate, a composite order parameter can be used in 1-D or 2-D reaction coordinate computations.[75] Each bond that is formed or broken during a reaction can be allocated to a specific dimension or can be merged with other bonds as chemically appropriate. Although this reduces the reliance of the free energy path on a specific variable, distances between atoms that are involved in bond breaking and formation are lost through conflation. As a result, the simplified picture lacks physically and chemically important information.[3,69] In the case of competing reactions an alternative approach to conflation is to run multiple simulations, each one modeling a distinct chemical transformation. Alternative reaction

coordinates can be delineated and the relative barrier heights computed. However, this approach will miss important features that arise because of the interaction of the two separate reaction paths.

1.4.2 Characterizing the TS

The TS holds central importance in calculating free energies of activation and elucidating enzyme mechanisms. Characterization of the TS is closely related to the definition of the reaction coordinate since the location of the generalized TS is found at the free energy maximum along the reaction path. The TS population arises from an ensemble of activated complexes from possible reactive trajectories in phase space. Within the context of rare event dynamics, identification of the activated complexes has either been defined as those structures that have the same probability of ending at the reactant region as the product region, or alternatively as the points in configuration space with the highest probability that equilibrium trajectories passing through them are reactive.[74] It is clear that, although $\Delta G^{\ddagger}$ is conveniently calculated from a single point on the PMF, a characterization of the TS requires exploration of configuration space around this point. The local TS landscape can be qualitatively compared to the landscape along the complete free energy path. Importantly, the width of the entry and exit channels to the TS are a function of entropic contributions (primarily the vibrational degrees of freedom). The physical interpretation of the changes in these properties is that the relative widening of the landscape indicates a more favorable entropy change and so allows a greater rate of passage of reactive complexes.[1]

1.4.3 Accuracy-Efficiency Trade-off

The determinations of reliable reaction coordinate probability densities are contingent on extensive sampling of reaction space. This leads to the overarching challenge of QM/MM free energy methods, namely to balance computational accuracy and efficiency. A well-defined reaction coordinate and accurate Hamiltonian need to be defined to model the reaction phase space while remaining tractable. For example, a highly accurate, computationally expensive potential energy function that precludes statistically reliable sampling in multi-dimensional reaction coordinate space will return inaccurate results. Likewise, extensive sampling of inaccurate electronic and conformational configurations will also yield unreliable free energies. Therefore, armed with chemical insight into the model system, judicious approximations need to made,[53] and it is imperative to assess the reliability of the information extrapolated from the PMF.[76,77]

Unfortunately, there is no clear protocol for verifying free energy results, and this task is often complicated by lack of experimental data. When considering alternative/competing mechanisms, qualitative agreement of the computational activation energies with experiment is often sufficient to distinguish the most probable pathway.[37] Importantly, agreement with the experimental reaction

barrier alone does not guarantee validation of the model as this may be serendipitously achieved from an incorrect mechanism. It has been shown in the hairpin ribozyme system[76] that semi-empirical methods can give comparable enthalpic results to *ab initio* methods but along an incorrect reaction mechanism. Further kinetic data, such as KIEs, and comparison of the sampled conformations against higher levels of theory should also be considered. Path integral sampling of reactant and TS ensembles can be used to treat the quantum behavior of the nuclei and so 'correct' PMFs resolved using classical Newtonian dynamics. In this way comparative KIE ratios can be generated but, unfortunately, path integral simulations are suited to enzyme systems where only a few degrees of freedom involved in the chemistry need to be quantized.[78] Therefore, computational KIEs are often limited to calculation from vibrational frequencies obtained from normal mode analysis of stationary structures using the Bigeleisen-Mayer equation.[79,80]

1.5 *Trypanosoma cruzi trans*-Sialidase

In this thesis, reaction free energy profiles are calculated for the reactions catalyzed by *T. cruzi trans*-sialidase (TcTS). TcTS has been identified as a drug target against *T. cruzi* infection that leads to trypanosomiasis or Chagas disease. TcTS fulfils a number of important functions in the course of *T. cruzi* infection, one of which is to coat the parasite's membrane with neuraminic acid derivatives (Figure 1.4A) and thus evade detection as a foreign cell by the host's immune response.[81-83] TcTS achieves this by catalyzing the selective transfer of the α(2,3)-linked sialic acids (SAs) from host sialyl-glycoconjugates to the terminal residues of cell surface galactopyranosyl-containing glycoconjugates. The importance of TcTS to *T. cruzi* has prompted extensive research into its activity in which lactose (Figure 1.4B) has often served as the sialic acid donor and/or acceptor.

TcTS is also of interest as a member of the large set of glycosyltransferases (GTs) and glycosylhydrolases (GHs, commonly called glycosidases) that synthesize glycoconjugates through concerted action. GTs catalyze monosaccharide transfer to either initiate formation of glycoconjugates or to extend existing glycans.[84] The donor sugar is commonly activated by a phosphoester bond to a mono- or dinucleotide moiety. GHs, on the other hand cleave glycosidic linkages to form intermediates that are then acted on by GTs. GTs and GHs have evolved to display high selectivity for their substrates and glycosylation typically proceeds as a sequential process. In this way oligosaccharide chains, or glycans, are attached to polypeptides, lipids, small organic molecules or DNA. The different glycosidic bond types, as well as the numerous positions on monosaccharides available for linking, impart glycans with a large structural variability and increasing complexity as further subunits are added. Such conjugates frequently occur attached to the membrane of eukaryotic and bacterial cells, or as part of the extracellular matrix. Their location and diversity impart cell surface glycans with the means to fulfill

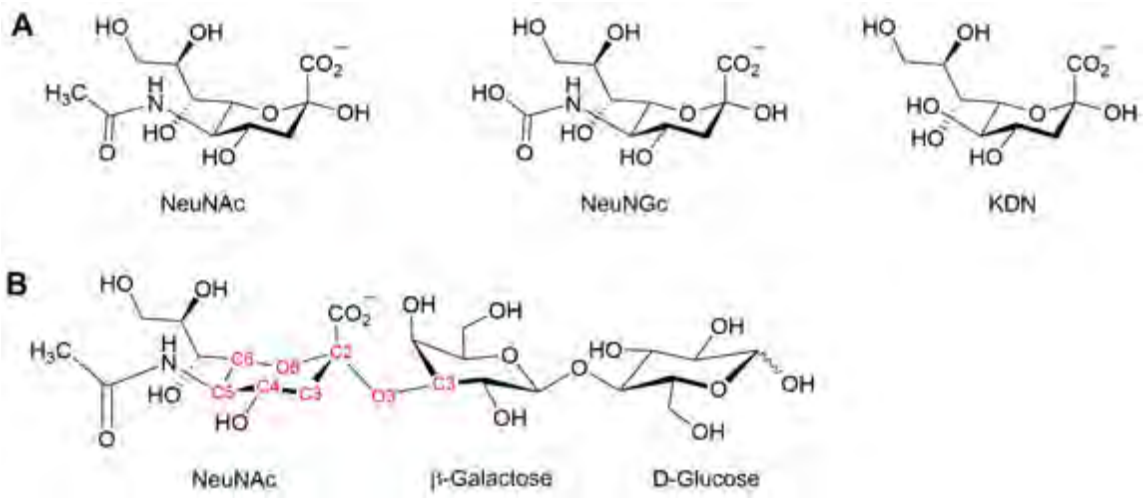


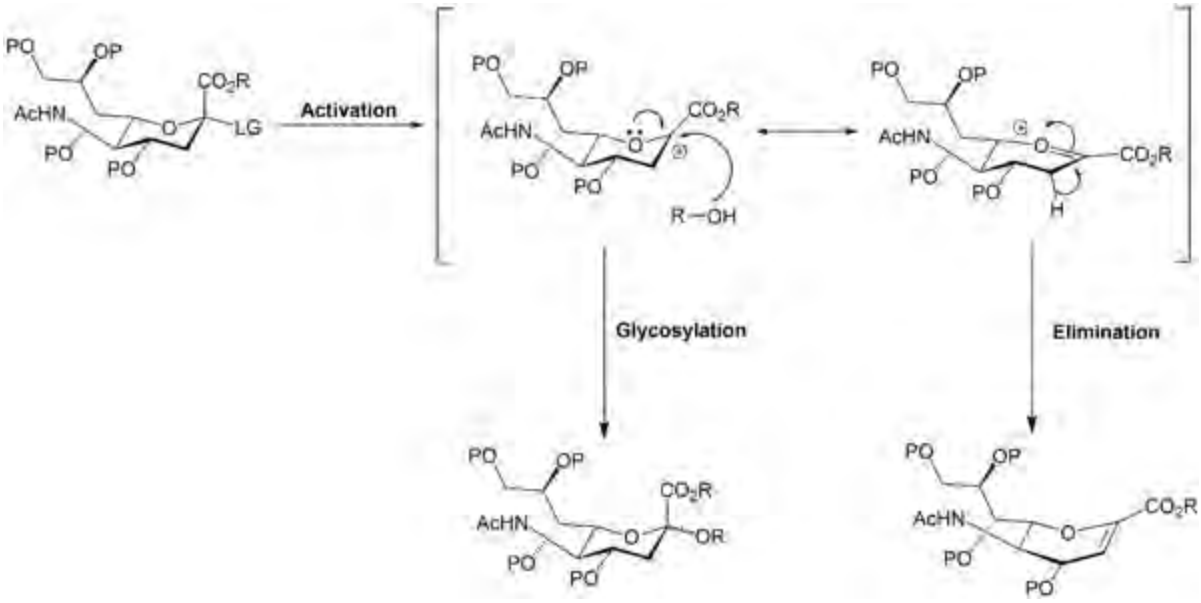
Figure 1.4 (A) Some common sialic acids (SAs) that are a diverse family of N- or O-substituted derivatives of neuraminic acid.[94] The anomeric carboxylate group, as displayed, is normally deprotonated at physiological pH. The resulting net negative charge strongly influences SA physicochemical properties. (B) The structure of α(2,3)-sialyllactose.

numerous roles including signaling, recognition and adhesion. In fact, the complexity achieved in multi-cellular organisms from a relatively a small number of genes has been attributed to the cells glycome.[84] Factors that result in under or over glycosylation will disrupt the normal functioning of cells, and it is important to elucidate the structure-function relationship and biochemical pathways of cell surface glycans.[84]

Sialyltransferases (STs) and sialylhydrolases (SHs, commonly called sialidases) are a subset of the carbohydrate-active enzymes that metabolize sialic acids. Mechanistic action is either retaining or inverting according to the retention or inversion of the stereochemistry of reaction centers of the substrates.[85] Retaining exo-sialidases are grouped into the Carbohydrate-Active enZymes Database (CAZy) families GH33 (bacterial and eukaryotic enzymes), GH34 (sialidases from different Influenza virus strains), and GH83 (other viral sialidases), while the fourth family, GH58, contains bacteriophage endo-sialidases.[85] Bacterial and eukaryotic STs are inverting GTs thought to operate through a concerted displacement reaction (A_ND_N-like)[†] that proceeds via a single oxocarbenium-ion-like activated complex. Bacterial STs are grouped into four families GT42, GT52, GT80 and GT38, whereas all eukaryotic STs are grouped into GT29. STs show remarkable selectivity considering the polyfunctional and complicated molecular architecture of sialic acids. These characteristics have presented considerable challenges to synthetic chemists.[86] Constructing sialic acid-containing

[†] In this thesis the Guthrie-Jencks mechanistic nomenclature, recommended by IUPAC, has been adopted in preference to the Ingold system (S_N2, E1, etc.). The Guthrie-Jencks indicates pre-association steps and proton transfers, which are particularly important in carbohydrate chemistry. Please refer to Appendix B for a description of the nomenclature.

saccharides in solution proceeds via an oxocarbenium ion intermediate (D_N+A_N; Scheme 1.1). However, the anomeric carboxylic acid destabilizes the formation of the oxocarbenium ion through an inductive electron withdrawing effect. The bulky ion also contributes to steric constraints at the C-2 center that results in the formation of 2,3-didehydro side-products from a competing $D_N+A_{xh}D_H$ elimination reaction. Selectivity is further hindered by the lack of anchimeric assistance from a neighboring C-3 substituent, as well as nucleophilic attack of water (hydrolysis).



Scheme 1.1 Elimination side reaction reducing selectivity in synthetic reactions of protected sialic acid donors adapted from Ress and Linhardt.[86]

1.5.1 TcTS Structure and Activity

The TcTS structure is built from a catalytic *N*-terminal domain linked, by a long α-helix, to the globular C-domain comprising two antiparallel β-sheets arranged in a β-sandwich-like structure (Figure 1.5A).[87] The *N*-terminal domain is folded into a six-bladed β propeller, and contains the sialic acid-binding residues. Within the catalytic site, the anomeric carboxyl group of the sialoside binds to an arginine triad composed of Arg_{35}, Arg_{245} and Arg_{314}, while the C-5 acetamido group hydrogen bonds to Asp_{96}. The deep catalytic pocket is lined with hydrophobic residues such as Met_{95}, Phe_{115}, Trp_{120} and Val_{176}.

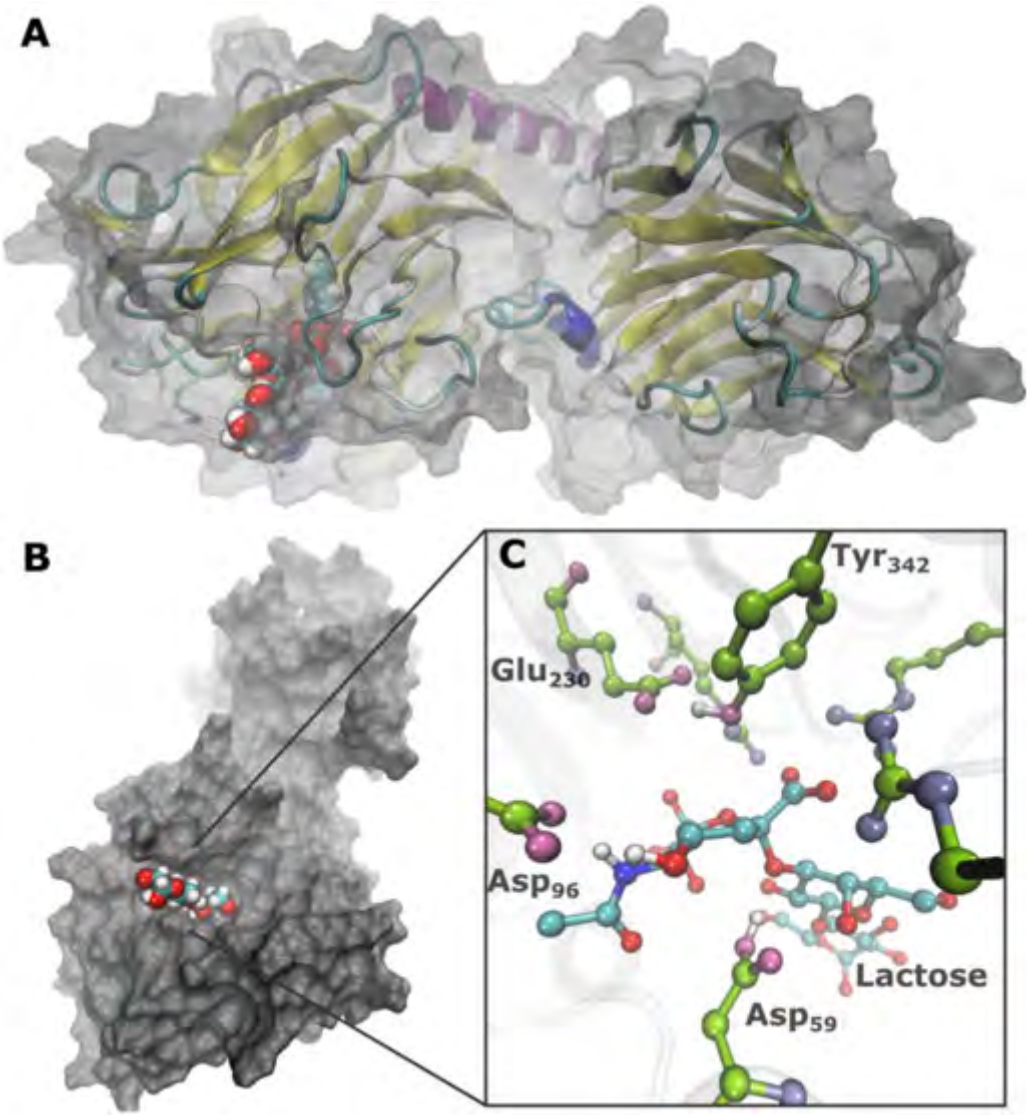


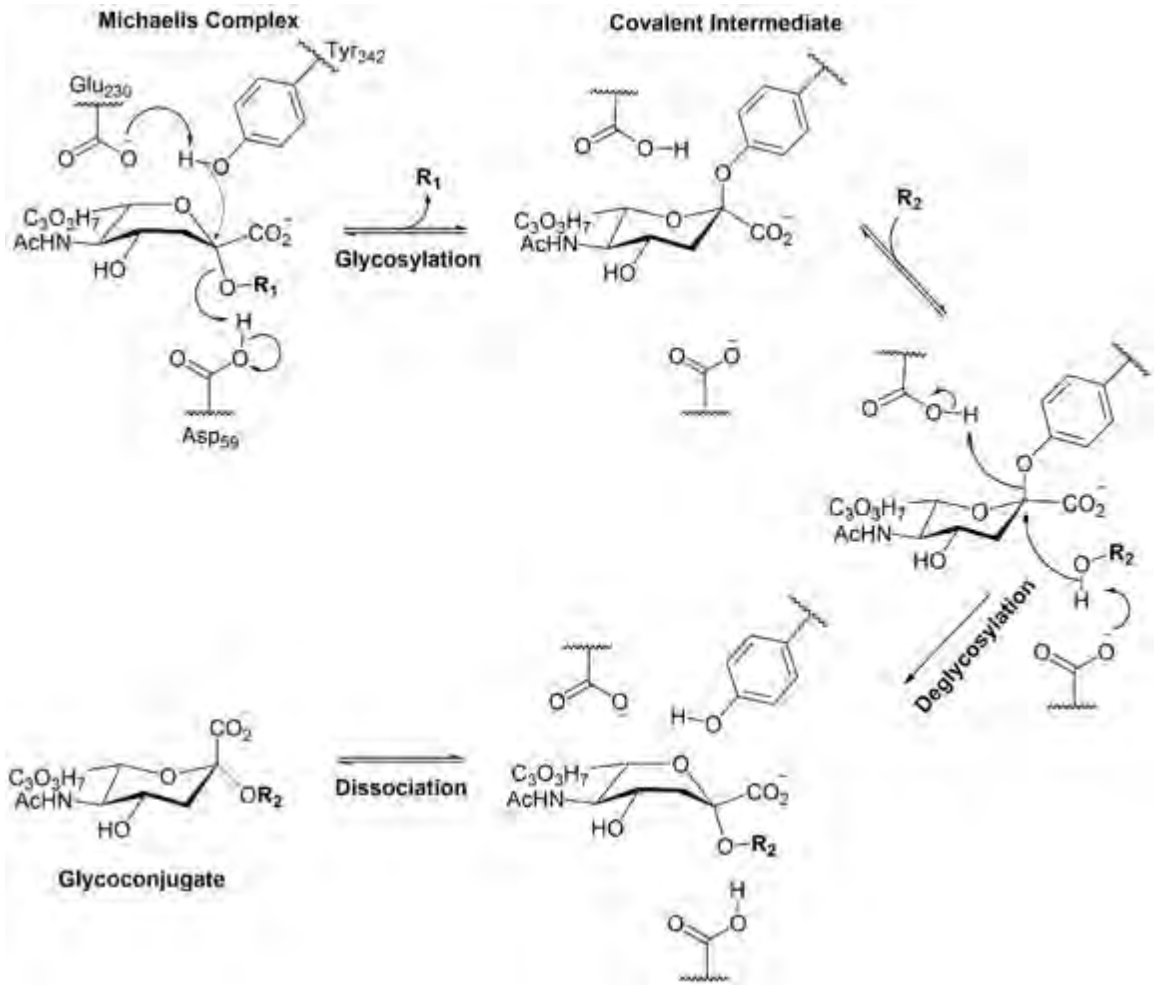
Figure 1.5 The structure of TcTS is summarized in (A) and (B) which display the protein secondary structure and sialyllactose binding mode. The Michaelis complex is described in atomic detail in (C) for which only hydrogens involved in important hydrogen bonding interactions are displayed for clarity. The amino acid residues are shown in an alternative color scheme and include the catalytic acid/base residue Asp_{59}, the nucleophile pair Tyr_{342}/Glu_{230}, the arginine triad that stabilizes the SA carboxylate, and Asp_{96} that assists in distorting the SA ring pucker into a $B_{2,5}$ conformation through hydrogen bonding interactions.

Although TcTS is classified as an exo-sialidase in the GH33 family, the enzyme preferentially transfers sialic acid units to β-galactosyl-containing molecules. Activity shows the selective formation of α(2,3)-linkages and retention of configuration at the anomeric position. Uncommonly to GTs, the sialic acid unit is not chemically activated by a phosphoester bond. TcTS achieves net sialyltransferase activity through a double-displacement reaction. Strong nucleophilic participation was inferred from primary ^{13}C and β-dideuterium KIE ratios,[88] and the collapse into a covalent intermediate was confirmed by the X-Ray crystal structure of catalytically incompetent $TcTS_{Ala59}$ trapped with Tyr_{342} covalently bonded to 3-fluorosialoside.[89]

There has been some debate on whether the double-displacement proceeds along a classical Ping-Pong mechanism in which the sialic acid donor dissociates before the acceptor is bound, or whether a ternary complex is formed with both the sialic acid donor and acceptor bound at the active site. MD simulations of the TcTS with α(2,3)-sialyllactose at the active site[90] have supported a previous proposal[91] of a distinct aglycone binding site formed via sialic acid-induced conformational changes. Similar evidence was not seen in simulations for the $TcTS_{D247A}$:sialyllactose complex, and NMR-derived dissociation constants showed that the complex did not bind lactose. Based on these results, a hybrid Ping-Pong sequential mechanism was proposed where a covalent intermediate is formed after both the sialic acid donor and acceptor molecules are bound. However, the current opinion leans in favor of a classic Ping-Pong mechanism (Scheme 1.2) in which Asp_{59} serves as the catalytic acid/base. Such a mechanism is consistent with the lack of a distinct second binding site for the acceptor molecule in crystal structures,[89] and has been supported by kinetic and chemical rescue[92] experiments.[†] In the latter study, assay measurements monitored the transfer of activated sialic acid from *p*-nitrophenyl sialic acid (PNP-SA) and 4-trifluoromethylumbelliferyl sialic acid (CF3MU-SA) to lactose. The second step, deglycosylation of sialylated TcTS, was identified as the rate-limiting step, and rate constants of 7.4 and 6.7 s^{-1} were calculated for the respective sialic acid donors.

According to the Ping-Pong mechanism, the reaction is initiated by binding of the sialic acid donor substrate. The size of the catalytic cleft is defined by the intramolecular distance between Trp_{312} and Tyr_{119} at the protein surface. These residues serve as gatekeepers and the loops that carry them shovel sialic acid oligosaccharides into the binding site.[93] On binding, the mobile loops are locked through the induced interaction mode of Arg_{314} with the SA C-2 carboxylate. In this conformation, Trp_{312} and Tyr_{119} stabilize the galactoside moiety of the substrate in the enzyme pocket through CH – π interactions.

[†] The addition of azide restored the activity of a mutant enzyme that had been rendered inactivate by mutation of Asp_{59} to Ala. This result confirmed the role of Asp_{59} as the acid/base catalyst.



Scheme 1.2 Putative mechanism for TcTS catalytic activity.[86] In the first step, nucleophilic attack by the nucleophile pair on the host's SA donor (R_1) forms the covalent intermediate. Deglycosylation of TcTS-SA by the parasite's galactosyl acceptor (R_2) forms the sialylated glycoconjugate.

The sialic acid ring is distorted into the $B_{2,5}$ pucker through interactions between its C-5 acetamido moiety and Asp_{96} (Figure 1.5). As a result, the leaving group is orientated in a pseudo-axial position, thereby optimizing proton donation from Asp_{59}. The anomeric carbon is also placed in an optimal position for in-line attack by the Tyr_{342}/Glu_{230} nucleophile pair with minimal encumbrance from 1,3-diaxial repulsions.[89] QM/MM analysis[94] has shown that the proton transfer from Tyr_{342} to Glu_{230} is unfavorable in the apoenzyme, but becomes favorable on substrate binding due to altered interaction modes for Arg_{35}, Arg_{314} and Arg_{53}. In this way, substrate binding enables Tyr_{342} to be deprotonated and act as a nucleophile. The use of a neutral tyrosine as a nucleophile reduces charge repulsion with the negatively charged SA C-2 carboxylate group hindering access to the anomeric center. Formation of a Tyr_{342} – SA covalent bond and proton donation from Asp_{59} to the substrate aglycone yields the covalent intermediate in which the sialic acid ring adopts a $^2C_5$ conformation. The sialic acid donor subsequently leaves the TcTS binding pocket to make space for the incoming β-galactose (Gal) acceptor. In the

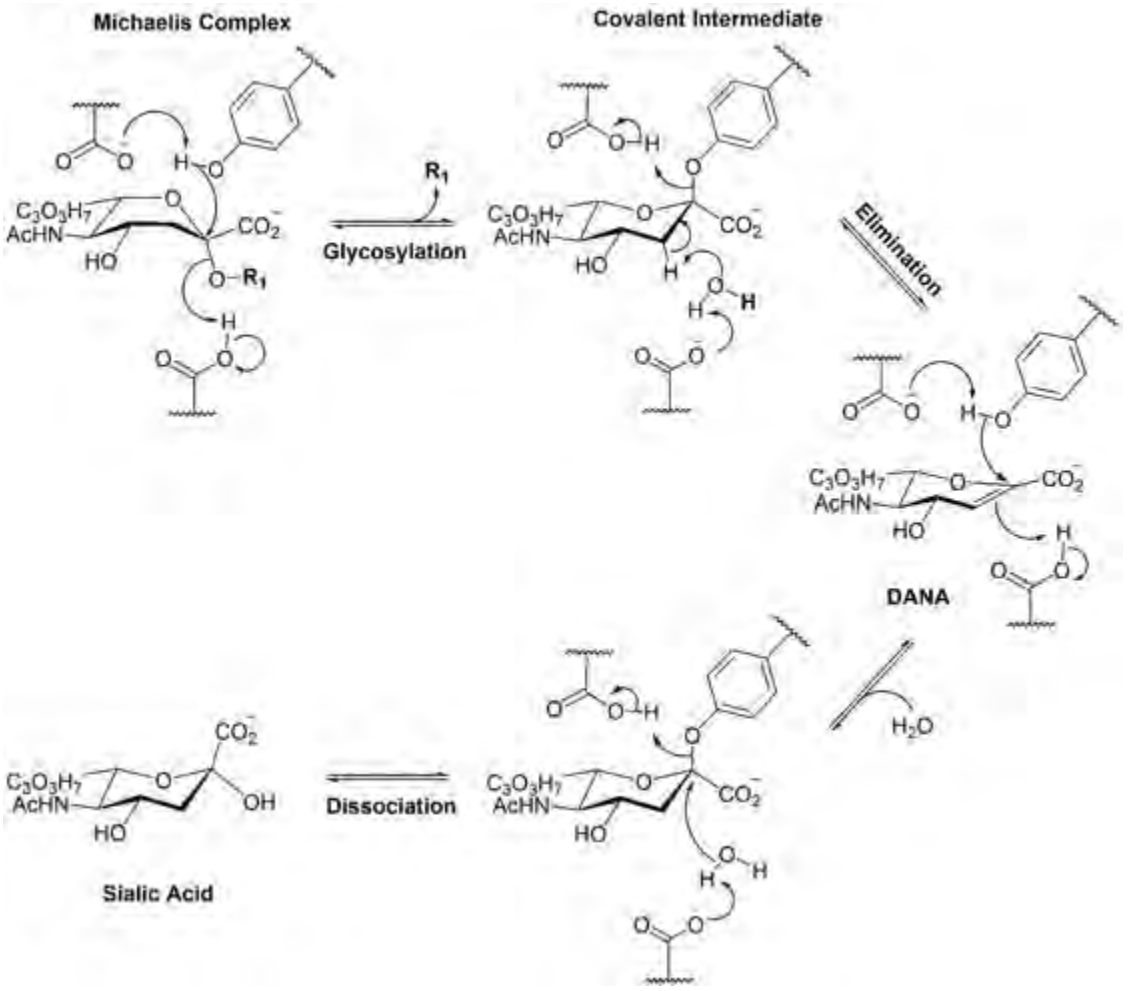
second step of the catalytic itinerary, sialic acid is transferred in a deglycosylation step through the attack of the β-galactose C-3 hydroxyl group that is deprotonated by the Asp_{59} acid/base catalyst.

1.5.2 Side Reactions to TcTS Transferase Activity

The complete two-step transferase reaction is susceptible to two side reactions. In the absence of galactoside acceptor, TcTS catalyzes sialoside hydrolysis of the covalent intermediate by a water molecule at the active site.[95] MD simulations show that TcTS mitigates the hydrolysis side reaction by excluding water from the active site[96] and adopting a catalytically competent Tyr_{342} – Glu_{230} conformation only in the presence of the acceptor molecule.[97] QM/MM free energy calculations further reveal that selectivity is explained by the ~7 kcal/mol difference in energy barrier for the chemical steps.

Interestingly, NMR experiments have also revealed the biosynthesis of 2-deoxy-2,3-didehydro-*N*-acetylneuraminic acid (DANA) when an excess of TcTS is incubated with sialyllactose (Scheme 1.3).[95] Of significance here is the phenomenon that DANA is a known (albeit weak) inhibitor of TcTS, and as such has previously been co-crystalized in the TcTS active site.[98] This competition of elimination side reactions, seen in synthetic experiments, has been observed as a minor reaction in other sialidases such as those expressed by *Influenza B* virus,[99] *Vibrio cholera*[100] as well as *Streptococcus pneumonia*.[101] A contrary example is that of *Streptococcus pneumonia* NanC where the elimination reaction is the primary reaction. The opposite display in selectivity by this sialidase was recently attributed to the exclusion of water from the active site and proximity of catalytic Asp_{315} to SA C-3 in crystal structures of the NanC:DANA complex.[102] It was argued that the catalytic aspartate abstracts the C-3 proton directly. On the other hand, the production of DANA in TcTS has not been explored, and the computational study of the competing elimination reaction forms one of the major aims of this thesis.

The general mechanism for sialic acid elimination in sialidases, as outlined by Jongkee and Withers,[103] shows SA H-3 abstraction from the covalent intermediate[103] or oxocarbenium[99,100] species mediated by an active site water molecule. For TcTS under physiological conditions, a bound galactose acceptor forms hydrogen bond interactions between O-3 and Asp_{59}. In this scenario, sialic acid will be susceptible to abstraction mediated by β-galactose O-3 rather than an active site water molecule. The detection of DANA in only trace amounts for TcTS incubated with sialyllactose under hydrolyzing conditions, while the SA C-3 proton is readily eliminated in synthetic experiments, points to the ability of TcTS to mitigate the unfavorable side reaction. Selective transformation of substrates to a single product, alongside efficiency, is one of the hallmarks of enzyme catalysis, and characterization of the competing reactions at an atomistic level allows a unique opportunity to examine how enzymes mitigate competing reactions.



Scheme 1.3 General mechanism for the elimination side reaction observed for a number of sialidases proposed by Jongkee and Withers.[103] Proton abstraction of the covalent intermediate by the catalytic aspartate is mediated by a water molecule.

1.5.3 TcTS as a Model System for Reaction Free Energy Simulations

In this thesis the challenges facing free energy methods discussed in Section 1.4 are addressed by using the FEARCF flat histogram method to construct reaction surfaces and volumes for the glycosylation and deglycosylation reactions comprising the TcTS catalytic itinerary. TcTS was chosen for the wealth of experimental data available for it, built from its significance as a potential drug target against Chagas disease. Despite the availability of this data, the complexity of the multistep transfer of sialic acid has denied complete characterization by experiment. Of importance is the identification of an elimination reaction competing with the primary transferase activity. The identification of this side reaction, that is observable only in the absence of the *trans*-sialidase or sialic acid acceptor, presented the opportunity to study the means by which enzymes selectivity bias in favor of a single reaction path. I therefore set out to explore the molecular details of how *T. cruzi trans*-sialidase asserts a precision and selectivity synonymous with enzyme catalysis. The aims and objectives of this thesis are laid out in detail in the next chapter.

1.6 References

(1) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books: Sausalito, California, 2005.

(2) Schramm, V. L. *Annu. Rev. Biochem.* **2011**, *80*, 703.

(3) Masgrau, L.; Truhlar, D. G. *Acc. Chem. Res.* **2015**, *48*, 431.

(4) Zheng, J.; Truhlar, D. G. *Faraday Discuss.* **2012**, *157*, 59.

(5) Truhlar, D. G. *Arch. Biochem. Biophys.* **2015**, *582*, 10.

(6) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186.

(7) Warshel, A. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 425.

(8) Steiner, K.; Schwab, H. *Comput. Struct. Biotechnol. J.* **2012**, *2*, 1.

(9) Scott, L. T. *Nat. Chem.* **2014**, *6*, 177.

(10) Breslow, R.; Dong, S. D. *Chem. Rev.* **1998**, *98*, 1997.

(11) Hanoian, P.; Liu, C. T.; Hammes-Schiffer, S.; Benkovic, S. *Acc. Chem. Res.* **2015**, *48*, 482.

(12) Pauling, L. *C&EN* **1946**, *24*, 1375.

(13) Kamerlin, S. C. L.; Warshel, A. *Proteins* **2010**, *78*, 1339.

(14) Menger, F. M. *Biochemistry* **1992**, *31*, 5368.

(15) Strajbl, M.; Shurki, A.; Kato, M.; Warshel, A. *J. Am. Chem. Soc.* **2003**, *125*, 10228.

(16) Menger, F. M. *Acc. Chem. Res.* **1985**, *18*, 128.

(17) Menger, F. M. *Acc. Chem. Res.* **1993**, *26*, 206.

(18) Bruice, T. C.; Lightstone, F. C. *Acc. Chem. Res.* **1999**, *32*, 127.

(19) Bruice, T. C.; Benkovic, S. J. *Biochemistry* **2000**, *39*, 6267.

(20) Zhang, X.; Bruice, T. C. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18356.

(21) García-Meseguer, R.; Martí, S.; Ruiz-Pernía, J. J.; Moliner, V.; Tuñón, I. *Nat. Chem.* **2013**, *5*, 566.

(22) Kohen, A. *Acc. Chem. Res.* **2015**, *48*, 466.

(23) Burschowsky, D.; van Eerde, A.; Okvist, M.; Kienhöfer, A.; Kast, P. et al. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, 17516.

(24) Schopf, P.; Mills, M. J. L.; Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 4328.

(25) Giraldo, J.; Roche, D.; Rovira, X.; Serra, J. *FEBS Lett.* **2006**, *580*, 2170.

(26) Pisliakov, A. V.; Cao, J.; Kamerlin, S. C. L.; Warshel, A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 17359.

(27) Eisenmesser, E. Z.; Millet, O.; Labeikovsky, W.; Korzhnev, D. M.; Wolf-Watz, M. et al. *Nature* **2005**, *438*, 117.

(28) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679.

(29) Schwartz, S. D.; Schramm, V. L. *Nat. Chem. Biol.* **2009**, *5*, 551.

(30) Veglia, G. *Nat. Chem. Biol.* **2013**, *9*, 410.

(31) Klinman, J. P. *Acc. Chem. Res.* **2015**, *48*, 449.

(32) Ardevol, A.; Rovira, C. *J. Am. Chem. Soc.* **2015**, *137*, 7528.

(33) Schramm, V. L. *Ann. Rev. Biochem.* **1998**, *67*, 693.

(34) Wigner, E. *Z. Physik. Chem. B.* **1932**, *19*, 203.

(35) Laidler, K. J.; King, M. C. *J. Phys. Chem.* **1983**, *87*, 2657.

(36) Eyring, H. *Chem. Rev.* **1935**, *17*, 65.

(37) van der Kamp, M. W.; Mulholland, A. J. *Biochemistry* **2013**, *52*, 2708.

(38) Carpenter, B. K. *Angew. Chem. Int. Ed.* **1998**, *37*, 3340.

(39) Kemsley, J. *C&EN* **2014**, *92*, 34.

(40) Hong, Y. J.; Tantillo, D. J. *Nat. Chem.* **2014**, *6*, 104.

(41) Truhlar, D. G.; Garrett, B. C. *Acc. Chem. Res.* **1980**, *235*, 440.

(42) Pu, J.; Gao, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3140.

(43) Alhambra, C.; Corchado, J.; Sánchez, M. L.; Garcia-Viloca, M.; Gao, J. et al. *J. Phys. Chem. B* **2001**, *105*, 11326.

(44) Warshel, A. *Acc. Chem. Res.* **2002**, *35*, 385.

(45) Agarwal, P. K.; Billeter, S. R.; Hammes-Schiffer, S. *J. Phys. Chem. B* **2002**, *106*, 3283.

(46) Jensen, F. *Introduction to Computational Chemistry*; 2nd ed.; John Wiley & Sons, Ltd, 2007.

(47) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; World Scientific: 1998, p 385.

(48) Simons, J.; Jorgensen, P.; Taylor, H.; Ozment, J. *J. Phys. Chem.* **1983**, *87*, 2745.

(49) Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. *J. Phys. Chem.* **1985**, *89*, 52

(50) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P. et al. *Comput. Phys. Commun* **2010**, *181*, 1477.

(51) Glowacki, D. R.; Harvey, J. N.; Mulholland, A. J. *Nat. Chem.* **2012**, *4*, 169.

(52) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J. et al. *Angew. Chem. Int. Ed.* **2006**, *45*, 6856.

(53) Lodola, A.; Mulholland, A. J. In *Methods in Molecular Biology*; Humana Press: Totowa, NJ, 2013; Vol. 924.

(54) Senn, H. M.; Thiel, W. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198.

(55) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, California, 2000.

(56) Hammes, G. G.; Benkovic, S. J.; Hammes-Schiffer, S. *Biochemistry* **2011**, *50*, 10422.

(57) Field, M. J. *A Practical Introduction to the Simulation of Molecular Systems*; Second ed.; Cambridge University Press: New York, 2007.

(58) Kästner, J. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 932.

(59) Ensing, B.; De Vivo, M.; Liu, Z. W.; Moore, P.; Klein, M. L. *Acc. Chem. Res.* **2006**, *39*, 73.

(60) Barnett, C. B.; Wilkinson, K. A.; Naidoo, K. J. *J. Am. Chem. Soc.* **2011**, *133*, 19474.

(61) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.

(62) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339.

(63) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578.

(64) Torrie, G. M.; Valleau, J. P. *J. Chem. Phys.* **1977**, *66*, 1402.

(65) Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.

(66) Rajamani, R.; Naidoo, K. J.; Gao, J. *J. Comp. Chem.* **2003**, *24*, 1775.

(67) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.

(68) Wang, F. G.; Landau, D. P. *Phys. Rev. E* **2001**, *64*.

(69) Naidoo, K. J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9026.

(70) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108* 1964.

(71) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877.

(72) Geissler, P. L.; Dellago, C.; Chandler, D. *J. Phys. Chem. B* **1999**, *103*, 3706.

(73) Zahn, D. *Mol. Simul.* **2012**, *38*, 211.

(74) Crehuet, R.; Field, M. J. *J. Phys. Chem. B* **2007**, *111*, 5708.

(75) Pierdominici-Sottile, G.; Horenstein, N. A.; Roitberg, A. E. *Biochemistry* **2011**, *50*, 10150.

(76) Jir, S.; Kamp, M. W. V. D.; Mulholland, A. J.; Bana, P.; Otyepka, M. *J. Chem. Theory Comput.* **2014**, *10*, 1608.

(77) Plata, R. E.; Singleton, D. *J. Am. Chem. Soc.* **2015**, *137*, 3811.

(78) Vardi-Kilshtain, A.; Nitoker, N.; Major, D. T. *Archives of Biochemistry and Biophysics* **2015**, *582*, 18.

(79) Bigeleisen, J.; Mayer, M. G. *J. Chem. Phys.* **1947**, *15*, 261.

(80) Wolfsberg, M. *Acc. Chem. Res.* **1972**, *5*, 225.

(81) Schenkman, S.; Jiang, M. S.; Hart, G. W.; Nussenzweig, V. *Cell* **1991**, *65*, 1117−1125.

(82) Paris, G.; Cremona, M. L.; Amaya, M. F.; Buschiazzo, A.; Giambiagi, S. et al. *Glycobiology* **2001**, *11*, 305.

(83) Freire-de-Lima, L.; Oliveira, I. A.; Neves, J. L.; Penha, L. L.; Alisson-Silva, F. et al. *Front. Immunol.* **2012**, *3*, 356.

(84) Varki, A.; Cummings, R. D.; Esko, J. D.; Freeze, H. H.; Stanley, P. et al. *Essentials of Glycobiology*; Cold Spring Harbor Laboratory Press, 2009.

(85) Buschiazzo, A.; Alzari, P. M. *Curr. Opin. Chem. Biol.* **2008**, *12*, 565.

(86) Ress, D.; Linhardt, R. *Curr. Org. Synth.* **2004**, *1*, 31.

(87) Oliveira, I. A.; Freire-de-lima, L.; Penha, L. L.; Dias, W. B.; Todeschini, A. R. *Proteins and Proteomics of Leishmania and Trypanosoma*; Springer Netherlands: Dordrecht, 2014; Vol. 74.

(88) Yang, J.; Schenkman, S.; Horenstein, B. *Biochemistry* **2000**, *39*, 5902.

(89) Amaya, M. F.; Watts, A. G.; Damager, I.; Wehenkel, A.; Nguyen, T. et al. *Structure* **2004**, *12*, 775.

(90) Oliveira, I. A.; Gonçalves, A. S.; Neves, J. L.; von Itzstein, M.; Todeschini, A. R. *J. Biol. Chem.* **2014**, *289*, 423.

(91) Todeschini, A. R.; Dias, W. B.; Girard, M. F.; Wieruszeski, J.; Mendonça-Previato, L. et al. *J. Biol. Chem.* **2004**, *279*, 5323.

(92) Damager, I.; Buchini, S.; Amaya, M. F.; Buschiazzo, A.; Alzari, P. et al. *Biochemistry* **2008**, *47*, 3507.

(93) Mitchell, F. L.; Miles, S. M.; Neres, J.; Bichenkova, E. V.; Bryce, R. a. *Biophys. J.* **2010**, *98*, L38.

(94) Pierdominici-Sottile, G.; Roitberg, A. E. *Biochemistry* **2011**, *50*, 836.

(95) Todeschini, A. R.; Mendonça-Previato, L.; Previato, J. O.; Varki, A.; van Halbeek, H. *Glycobiology* **2000**, *10*, 213.

(96) Demir, O.; Roitberg, A. E. *Biochemistry* **2009**, *48*, 3398.

(97) Bueren-Calabuig, J. A.; Pierdominici-Sottile, G.; Roitberg, A. E. *J. Phys. Chem. B* **2014**, *118*, 5807.

(98) Buschiazzo, A.; Amaya, M. F.; Cremona, M. L.; Frasch, A. C.; Alzari, P. M. *Molecular cell* **2002**, *10*, 757.

(99) Burmeister, W. P.; Henrissat, B.; Bosso, C.; Cusack, S.; Ruigrok, R. W. H. *Structure* **1993**, *1*, 19.

(100) Moustafa, I.; Connaris, H.; Taylor, M.; Zaitsev, V.; Wilson, J. C. et al. *J. Biol. Chem.* **2004**, *279*, 40819.

(101) Xu, G.; Kiefel, M. J.; Wilson, J. C.; Andrew, P. W.; Oggioni, M. R. et al. *J. Am. Chem. Soc.* **2011**, *133*, 1718.

(102) Owen, C. D.; Lukacik, P.; Potter, J.; Sleator, O.; Taylor, G. L. et al. *J. Biol. Chem.* **2015**, *290*, 27736.

(103) Jongkees, S. A. K.; Withers, S. G. *Acc. Chem. Res.* **2014**, *47*, 226.

2 Aims and Objectives

The free energy of activation, as defined by TST, is central to calculating reaction rates, distinguishing between reaction mechanisms, and elucidating vehicles of catalysis. The overarching aim of this thesis is to address some of the challenges facing current methods for calculating QM/MM reaction free energies in glycoenzyme systems. TcTS serves as a pertinent system due to the role of *T. cruzi* as an agent of Chagas disease, the extensive experimental data available for it, and the intriguing observation of side reactions to the primary *trans*-sialidase activity. Previous QM/MM MD simulations have focused on the role of substrate binding in modulating the active site, while a conflated 2-D reaction coordinate has been used with restrained dynamics to calculate profiles for the transition between Michalis complex and covalent intermediate and the competing hydrolysis of the covalent intermediate.

It is common for computational investigations of enzyme reactions to consider only $\Delta G^{\ddagger}$ associated with the opaque concept of a hypothetical dividing hypersurface to explain rate enhancement. The first objective of this thesis is to gain a more complete understanding of the second TcTS-catalyzed reaction, the deglycosylation step. In Chapter 5, deglycosylation by a lactose acceptor is modeled explicitly as the forward reaction, treating the translational and rotational freedom of the β-galactose O-3 nucleophile. The free energy for the complete reaction path is calculated using FEARCF, a method that avails both the activation free energy as well as sketches of rare event dynamics. An ensemble of activated complexes can be gained through running multiple reactive FEARCF trajectories, and allows profiling of the well-configured TS structures that successfully cross to product. In doing so, the chemical nature of the conformations local to the generalized TS may be characterized.

The second objective is to increase the TcTS reaction space to explore the elimination side reaction to DANA under physiological conditions. The use of a multi-dimensional (3-D) reaction coordinate allows exploration of DANA production in Chapter 6 using lactose as the sialic acid donor or acceptor in the glycosylation and deglycosylation steps. The choice of reaction coordinate is an important consideration when treating two correlated pathways, and a multi-dimensional decomposition is sought to account for features that arise because of interaction between the competing displacement and elimination paths. The resultant free energy paths can be analyzed to extract pertinent conformations and gain an understanding of the intermolecular interactions that explain how TcTS mitigates the elimination reaction. Such an analysis directly addresses the concept of enzyme reaction selectivity as observed in TcTS.

The final objective of this thesis is to verify the reliability of information extracted from free energy volumes resolved using SCC-DFTB. It is desirable to verify both the free energy results and the underlying structures. In lieu of structural data for the TS, the latter can be evaluated by comparing SCC-DFTB/MM results against QM/MM geometry optimizations for the large QM region (~100 atoms) at a higher level of theory. In addition, a high-level reactive trajectory from Michaelis complex to covalent intermediate is prepared in Chapter 7 by differentiating FEARCF driving forces from the SCC-DFTB/MM PMF. A successful DFT/MM crossing is contingent on comparable gradients for the SCC-DFTB and DFT energy functions, and would grant access to a molecular orbital description along the reaction path.

3 Treating Enzyme Reaction Spaces Using DFT/MM and SCC-DFT/MM Methods

The electronic degrees of freedom in the reaction space can be modeled with state-of-the-art QM methods for small catalytic-site models. However, such models are limited to calculations on the potential energy surface, may lack some functional groups that contribute electronically to the reaction, and neglect the polarizing protein environment. In order to calculate reliable free energies for TcTS-catalyzed reactions (~100 atom catalytic site), the QM level of theory should accurately model the electron distribution whilst still offering the speed required for statistically reliable sampling along a multi-dimensional reaction coordinate.

DFT explicitly includes local correlation while being faster than post-Hartree-Fock wavefunction methods. However, DFT is still too slow to construct accurate probability histograms along most enzyme reaction coordinates. SCC-DFTB, an approximate DFT method, is therefore used in the current research since it is fast (comparable to semi-empirical wavefunction methods) and has been benchmarked for the TcTS glycosylation reaction.[1] In this thesis, SCC-DFTB reaction paths are validated by comparing structures optimized onto the potential energy surface with those obtained using the modern M06-2x functional. Deviations from *ab initio* behavior also determine the extent of overlap with DFT that is necessary for running successful DFT/MM reactive trajectories on the SCC-DFTB PMF. An understanding of the approximations introduced in the derivation of SCC-DFTB is important to evaluating its performance. DFT and SCC-DFTB will be discussed in this context, whereupon further simplification of reaction space using the QM/MM scheme will be introduced.

3.1 Treating Exchange and Correlation with DFT

The Born-Oppenheimer approximation, which postulates that electrons can instantaneously respond to any changes in the relative positions of the nuclei, allows the electronic and nuclear energy to be treated independently. By keeping nuclei fixed, the electronic energy can be solved as a function of the molecular wavefunction (using molecular orbital or valence bond methods) or electron density (DFT). Hartree-Fock (HF) and density functional theory share a similar self-consistent field (SCF) approach that arises from use of the variational principle and application of a linear combination of atomic orbitals (LCAO) basis set.[2]

3.1.1 Hartree-Fock

Molecular orbital methods theorize that electrons do not belong to particular bonds but are spread throughout the entire molecule. The electronic state is described mathematically using a molecular wavefunction, Ψ, which is dependent on one spin and three spatial coordinates for every electron. Ψ, possessing units of probability density to the one-half power, is only interpretable when probed by an appropriate quantum mechanical operator to retrieve physically meaningful results. Below, the Hamiltonian, $\hat{\mathrm{H}}$, is the electronic energy operator that acts on the *n*-electron wavefunction, and thus contains the appropriate terms for the kinetic and potential energy of the system:

$$\hat{\mathrm{H}}\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2 \dots \boldsymbol{r}_n) = E\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2 \dots \boldsymbol{r}_n)$$

Equation 3.1

for which the energy is formulated:

$$E = \int \Psi(\boldsymbol{r}_1, \boldsymbol{r}_2 \dots \boldsymbol{r}_n)^* \hat{\mathrm{H}}\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2 \dots \boldsymbol{r}_n) d\tau$$

Equation 3.2

In Equation 3.2 $\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_n$ are the position vectors of the *n* electrons. When Ψ is expressed as a single Slater determinant of one-electron molecular orbitals, ψ_i, and the explicit Hamiltonian is used, the following expression for the energy is derived:

$$E = 2\sum_{i}^{n/2} H_{ii} + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(2J_{ij} - K_{ij}\right)$$

$$H_{ii} = \langle \psi_i^*(1) \left| -\frac{1}{2}\nabla_{\mathrm{i}}^2 - \sum \frac{Z_A}{r_{Ai}} \right| \psi_i(1) \rangle$$

$$J_{ij} = \langle \psi_i^*(1)\psi_i(1) \left| \frac{1}{r_{ij}} \right| \psi_j^*(2)\psi_j(2) \rangle$$

$$K_{ij} = \langle \psi_i^*(1)\psi_j^*(2) \left| \frac{1}{r_{ij}} \right| \psi_i(2)\psi_j(1) \rangle$$

Equation 3.3

where $r_{Ai} = |\boldsymbol{R}_A - \boldsymbol{r}_i|$ and $r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j|$. In the above equations J is the Coulomb integral that sums the repulsive interaction between infinitesimal volumes of charge density, and the Exchange integral K, arising from the anti-symmetry of Ψ, accounts for the phenomenon that electrons of the same spin avoid each other more so than expected from only the Coulombic repulsion. The energy in Equation 3.3 is solved using a variational approach in which the energy of the system is bounded by the true

energy. Accordingly, E is minimized with respect to the set of one-electron molecular orbitals that yields a set of $n \times n$ HF equations:

$$\hat{F}\psi_i(1) = -\frac{1}{2}\sum_{j=1}^{n} l_{ij}\psi_j(1)$$

Equation 3.4

where l_{ij} are the Lagrange multipliers and $\hat{F}$ is the Fock operator:

$$\hat{F} = \hat{H}^{core}(1) + \sum_{j=1}^{n}\left(2\hat{J}_j(1) - \hat{K}_j(1)\right)$$

In the transformed set of molecular orbitals where the matrix of Lagrange multipliers is diagonal, the generalized eigenvalue problem is obtained in which the diagonal elements of $\boldsymbol{L}$ are the molecular orbital energies:

$$\hat{F}\boldsymbol{\psi} = \varepsilon\boldsymbol{\psi}$$

Equation 3.5

Expressing each molecular orbitals as a LCAO, $\psi_i = \sum_{\mu}^{m} c_{\mu i}\phi_{\mu}$, leads to the practically useful $m \times m$ set of Roothan-Hall equations:

$$\boldsymbol{FC} = \boldsymbol{SC\varepsilon}$$

Equation 3.6

where $\boldsymbol{F}$ is the Fock matrix, $\boldsymbol{C}$ is the matrix of atomic orbital coefficients and $\boldsymbol{S}$ is the overlap matrix. In order to calculate the eigenvalues of ψ_i, the Fock matrix is diagonalized after orthogonalization by the overlap matrix:

$$\mathbf{F}' = \mathbf{C}'\boldsymbol{\varepsilon}\boldsymbol{C}'^{-1}$$

Equation 3.7

Since the LCAO coefficients are present in the Fock matrix itself, they must be optimized from an initial trial wavefunction through iterative SCF calculations. This allows the energetically best wavefunction to be found within the given basis set.

Within the HF formalism, electron-electron interactions are described by the Coulombic repulsion between each electron and an averaged electrostatic field, subject to a quantum mechanical correction by the exchange energy (Fermi correlation; Equation 3.3). Thus, the effects of the short- and long-range contributions to Coulomb correlation are ignored: the so-called short-range 'dynamic'

correlation is neglected because the one-electron Fock operators do not treat electrons as discrete point particles, and the long-range 'static' correlation is neglected because Ψ is formulated as a single determinant.[3] The probability of locating an electron at a particular time is, therefore, independent of the position of the other electrons. Indeed, the correlation energy is defined as the difference between the energy from the HF limit and the exact solution of the Schrödinger equation. Neglecting correlation allows electrons to come too close to each other, resulting in absolute energies that are too high.[4] This is a severe limitation for modeling reactions in which correlation energy changes significantly when bonds break. Electron-electron correlation can be explicitly treated through consideration of a multi-determinant Ψ using post-HF formalisms, such as Møller–Plesset perturbation and Configuration Interaction methods. In this way, wavefunction theory can be systematically improved by going to bigger basis sets and improved levels of theory (Figure 3.1).[4] Unfortunately, the computational expense of post-HF methods precludes their routine application to enzyme active sites.

3.1.2 Density Functional Theory

DFT has become very popular for modeling larger systems because this level of theory includes electron correlation but is not as demanding as post-HF methods.[5] DFT encodes the *n*-many-body problem of correlated electrons into a single-bodied problem by formulating the molecular energy in terms of the experimentally measurable electron probability density function, $\rho(x, y, z)$.

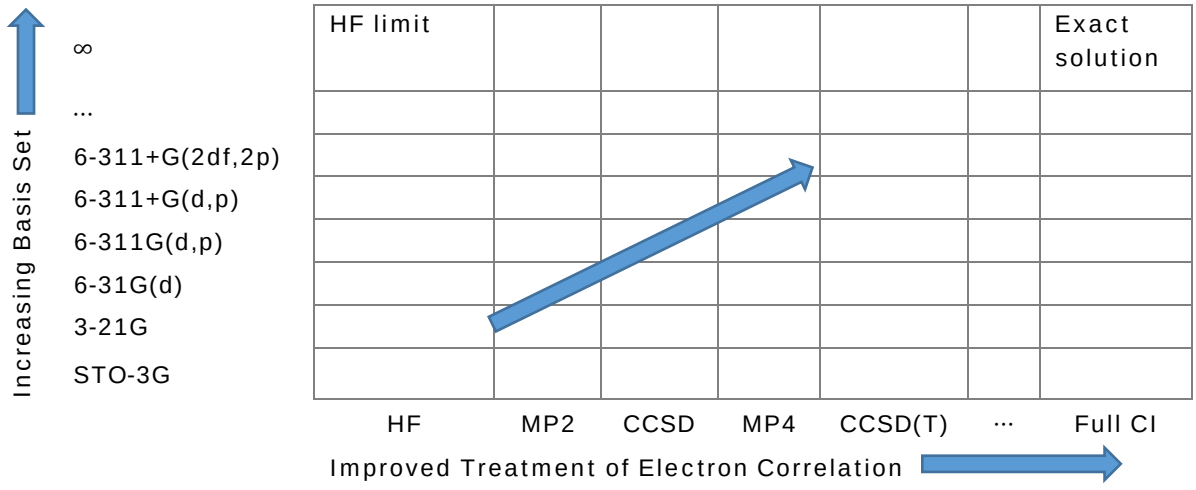


Figure 3.1 Illustration of the systematic convergence to the exact solution of the Schrödinger equation by improved treatment of electron-electron correlation and increasing the basis set size. Figure adapted from Jensen.[3]

3.1.2.1 Kohn-Sham Formalism

The theoretical framework for modern-day DFT stems from two critical theorems rigorously proved by Hohenberg and Kohn.[6,7] In the first, a system of *n*-interacting electrons is set in an external potential of N nuclei, v_{ext}, that is associated with a unique ground-state molecular wavefunction, Ψ_0. It can then be shown that there exists a one-to-one mapping from the ground-state density of the interacting electrons to v_{ext}. Therefore, the total electron density uniquely and exactly determines the ground-state properties of the system, and a universal energy functional can be defined:

$$E_0 = F[\rho_0] + V_{Ne}[\rho_0]$$

Equation 3.8

where F is the universal functional, and $V_{Ne}[\rho_0]$ accounts for the interaction energy between the molecular electron density and the nuclei. It was shown in the second theorem that, for a non-degenerate ground state, the electron density obeys the variational principle:

$$F[\rho'] + \int V_{Ne}\rho' d\boldsymbol{r} \geq F[\rho_0] + \int V_{Ne}\rho_0 \, d\boldsymbol{r}$$

Equation 3.9

That is, the input of the exact ground-state electron density into the universal functional yields the global minimum of this energy functional, E_0. The energy of the system can therefore be determined by variational optimization of the energy with respect to the electron density.

The Hohenberg-Kohn theorems lay the foundation for the Kohn-Sham (KS) equations[7] that provide the practical machinery to calculate the electronic energy. To circumvent the unknown form of $F[\rho]$, a fictitious reference system of non-interacting electrons is considered for which the ground-state density is necessarily the same as the real system; ρ then gives the correct unique position and atomic numbers of the nuclei. The energy is decomposed into the classical energies of the non-interacting electrons and the non-classical interactions which result in deviations of the reference system from the actual system of interacting electrons. The kinetic and potential correlation energies are subsumed into an unknown exchange-correlation functional, E_{xc}:

$$E[\rho] = \langle T[\rho]_{ref} + \Delta T[\rho] + V_{Ne}[\rho] + V_{ee}[\rho]_{ref} + \Delta V_{ee}[\rho] \rangle$$

$$E[\rho] = \langle T[\rho]_{ref} + V_{Ne}[\rho] + V_{ee}[\rho]_{ref} + E_{xc}[\rho] \rangle$$

$$E = T[\rho]_{ref} - \int \sum_A \frac{Z_A}{|\boldsymbol{r} - \boldsymbol{R}_A|}\rho(\boldsymbol{r})\, d\boldsymbol{r} + \frac{1}{2}\iint \frac{\rho(\boldsymbol{r})\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r} d\boldsymbol{r}' + E_{xc}[\rho]$$

Equation 3.10

where $T[\rho]_{ref}$ and $V_{ee}[\rho]_{ref}$ are the classical kinetic and potential energy functionals, and the deviations of these energies from the actual values are given by the delta terms. The potential energy, comprising nuclei-electron attraction and electron-electron repulsion terms, are described by classical Coulomb interactions. It is the classical form of the electron-electron repulsion that gives rise to the well-known self-interaction error in DFT.[8] On the other hand, the kinetic energy functional is unknown. Therefore, a ground state multi-electron wavefunction is defined in order to access the kinetic energy. $T[\rho]_{ref}$ can then be calculated as the expectation value for the sum of the one-electron kinetic energy operators over Ψ_{ref}:

$$\langle T[\rho]\rangle_{ref} = \left\langle \Psi_{ref} \middle| \sum_i -\frac{1}{2}\nabla_i^2 \middle| \Psi_{ref} \right\rangle$$

Equation 3.11

Since the reference system comprises non-interacting electrons, Ψ_{ref} can be written exactly as a single Slater determinant of occupied spin molecular orbitals ψ_i for which the electron density is defined:

$$\rho = 2\sum_i \left|\psi_i^2\right|$$

Equation 3.12

In practice, it is common to separate E_{xc} into a pure exchange (E_x) and correlation (E_c) components:

$$E_{xc}[\rho] = \int \rho E_x[\rho] d\boldsymbol{r} + \int \rho E_c[\rho] d\boldsymbol{r}$$

Equation 3.13

The differing treatment of exchange and correlation for α- and β-spin electrons is incorporated by expanding the total density as the sum of separate spin densities: $\rho = \rho_\alpha + \rho_\beta$ for which spin polarization, ς, can be defined:

$$\varsigma = \frac{\rho_\alpha - \rho_\beta}{\rho_\alpha - \rho_\beta}$$

Equation 3.14

The total KS energy can finally be written in terms of the KS spatial orbitals as:

$$E[\rho] = \sum_i^n n_i \left\langle \psi_i \middle| -\frac{1}{2}\nabla^2 - \sum_A^N \frac{Z_A}{|\boldsymbol{r} - \boldsymbol{R}_A|} + \frac{1}{2}\int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}' \middle| \psi_i \right\rangle + E_{xc}[\rho] + \frac{1}{2}\sum_A^N \sum_{A\neq B}^N \frac{Z_A Z_B}{r_{AB}}$$

Equation 3.15

where $r_{AB} = |\boldsymbol{R}_A - \boldsymbol{R}_B|$ and Z_A is the nuclear charge of atom A. The advantage of this formalism is that while $\hat{T}$ and $\hat{V}_{ee}$ are exact for the reference system, E_{xc} is only a relatively small part of the total.[5] Variational minimization by differentiation of the energy with respect to the KS orbitals (subject to the constraint that these remain orthonormal) yields:

$$\hat{h}^{KS}\psi^{KS} = -\frac{1}{2}\boldsymbol{L}\psi^{KS}$$

$$\hat{h}^{KS} = \left[-\frac{1}{2}\nabla^2 - \sum_{A}^{N}\frac{Z_A}{|\boldsymbol{r} - \boldsymbol{R}_A|} + \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|}d\boldsymbol{r}' + \frac{\delta E_{xc}[\rho]}{\delta\rho}\right]$$

Equation 3.16

In analogy to HF SCF calculations, by expanding the KS molecular orbitals in a LCAO basis set, the above set of equations can be formulated as a generalized eigenvalue problem $\boldsymbol{FC} = \boldsymbol{SC\varepsilon}$ with the KS Fock matrix elements:

$$F_{\mu\nu} = \langle\phi_\mu|\hat{h}^{KS}|\phi_\upsilon\rangle$$

Equation 3.17

3.1.2.2 Exchange and Correlation in DFT

The DFT formulation is exact, but the explicit functional form of E_{xc} is not known. Developing good exchange-correlation functionals is the primary concern in improving the accuracy of DFT.[4] Furthermore, absence of a theoretical prescription to systematically improve the accuracy of E_{xc} has led to the development of a range of functionals containing parameters. As a result, E_{xc} functionals are complex and the evaluation of the integrals is carried out numerically. The approximate nature of E_{XC} functionals means that DFT is not strictly variational, and the presence of parameters makes DFT philosophically semi-empirical in nature.[2] Parameters are optimized by employing one or both of the predominant philosophies: requiring the functional to fulfil the properties derived for the exact functional (the existence of which has been proven), or taking an empirical approach by fitting the parameters to experimental data. The overall performances of the developed functionals are ultimately benchmarked against experiment.[9]

Although there is no systematic pathway to improving E_{xc}, DFT functionals can be arranged on a conceptual Jacob's ladder[10] to chemical accuracy (Figure 3.2). Climbing the ladder addresses the local definition of the E_{xc} functional that leads to neglect of important physical interactions such as London dispersion forces. Thus, the DFT ladder climbs from 'Hartree World' including $\nabla\rho$ and $\nabla^2\rho$, non-local HF exchange, and finally promotion of electrons into virtual orbitals. The different descriptions of exchange and correlation in DFT and wave mechanics lead to two further considerations. Part of ε_{xc}

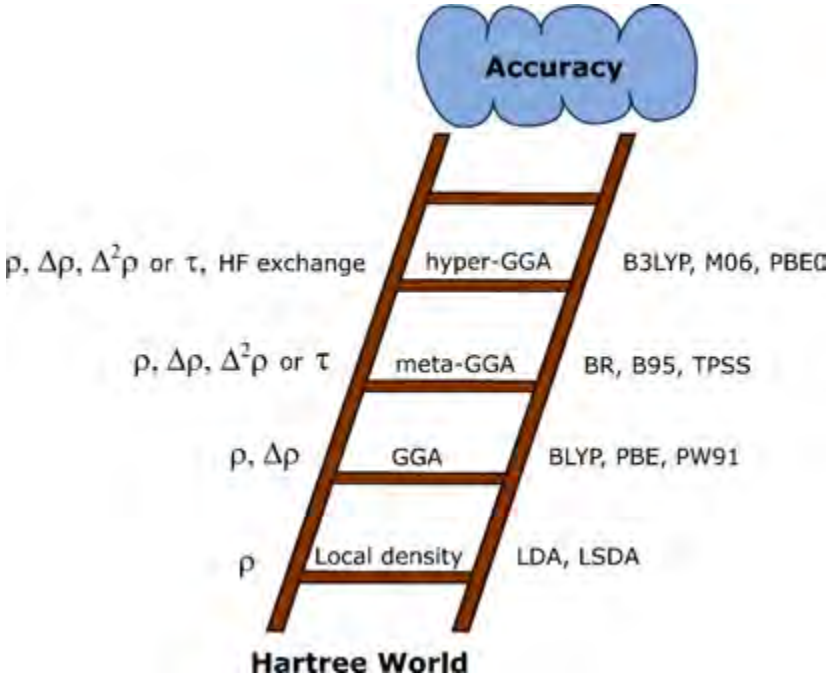


Figure 3.2 Illustration of climbing the ladder to chemical accuracy where each successive rung is occupied by exchange-correlation functionals with reduced locality. The increasing sophistication of the semi-local enhancements are associated with a modest increase in computation times but resource requirements climb much more steeply after that. Figure adapted from Perdew et al.[11]

can be viewed as a correction for self-interaction energy but, unlike wavefunction methods, cancellation is not guaranteed. Secondly, cancellation of the delocalized HF exchange hole by the long-range correlation in wavefunction approaches should be inherent in the functional. For this reason the two functional parts in Equation 3.13 should be developed in an integrated manner.[3]

The first rung, local spin-density approximation, assumes that the local density at each point in the molecule can be treated as a uniform electron gas. E_X can then be formulated as a function of the separate charge densities:

$$E_X^{LSDA}[\rho] = -2^{1/3} C_x \int \left(\rho_\alpha^{4/3} + \rho_\beta^{4/3}\right) d\boldsymbol{r}$$

Equation 3.18

The correlation energy is interpolated from the known energies of a uniform electron gas at $\varsigma = 0$ (paramagnetic system) and $\varsigma = \pm 1$ (ferromagnetic system). Intermediate values are calculated using suitable analytic formulas constructed to fit energies determined by quantum MC methods. The functional form of the *Vosko-Wilk-Nusair* (VWN) implementation[12] is given below:

$$\varepsilon_C^{VWN}(r_s,\varsigma) = \varepsilon_c(r_s,0) + \varepsilon_a(r_s)\left[\frac{f_2(\varsigma)}{f_2''(0)}\right](1-\varsigma^4) + [\varepsilon_c(r_s,1) - \varepsilon_c(r_s,0)]f_2(\varsigma)\varsigma^4$$

$$f_2(\varsigma) = \frac{f_1(\varsigma) - 2}{2^{1/3} - 1}$$

Equation 3.19

where r_s is radius of the effective volume containing one electron:

$$\frac{4}{3}\pi r_s^3 = \rho^{-1}$$

Equation 3.20

Moving up the ladder, rungs two and three of Jacob's Ladder (Figure 3.2) adopt the form:

$$E_{xc}[\rho] = \int \varepsilon_{xc}^{LSDA}(\rho) f$$

Equation 3.21

In Equation 3.21, f is a function of $\nabla\rho$ and/or $\nabla^2\rho$. The kinetic energy density, τ, that displays the same behavior as the Laplacian can be used for computational considerations:[11]

$$\tau(\boldsymbol{r}) = \frac{1}{2}\sum_{i=1}^{occupied} \left|\nabla\psi_i^{KS}(\boldsymbol{r})\right|^2$$

Equation 3.22

The fourth rung in Figure 3.2 is occupied by non-local hyper-GGA methods. These methods formulate E_{XC} as a weighted sum of the DFT exchange-correlation energy and HF exchange energy (an approach justified by the adiabatic connection formula):

$$E_{XC}^{hyb} = \frac{X}{100}E_X^{HF} + \left(1 - \frac{X}{100}\right)E_X^{DFT} + E_C^{DFT}$$

Equation 3.23

Although hybrid functionals led to a major improvement in accuracy, progress since these initial developments has been slower.[2,3,5] Additional parameterization has not resulted in significantly better overall performance, and B3LYP[13] still remains a valid and efficient functional:

$$E_{XC}^{B3LYP} = (1-a_0)E_X^{LSDA} + a_0E_X^{HF} + a_X\Delta E_X^{B88} + a_CE_C^{LYP} + (1-a_C)E_C^{VWN}$$

Equation 3.24

In Equation 3.24, parameters a_0, a_X and a_C were determined by fitting to experimental data that resulted in 72% contribution from HF exchange. Documented flaws in predictions of non-covalent

bonding interactions and reaction barrier heights[8] have been partially overcome by more recent hybrid meta-GGAs, such as M06-2x.[14] This functional is the top-performing functional for main group thermochemistry, kinetics and non-covalent interactions within the M06 suite. M06-2x uses the GGA exchange functional PBE and contains 54% HF exchange. The correlation treats opposite- and parallel-spin using a functional form based on the M05 and meta-GGA VSXC correlation functionals. M06-2x has demonstrated a good response under dispersion forces, arising from parametrization against benchmark databases.

3.1.2.3 Basis Set Considerations

The accuracy of DFT calculations relies primarily on the level of theory, discussed above, and the basis set used. The basis set is fundamentally the set of mathematical functions used to construct molecular orbitals through a linear combination. It is chemically intuitive to choose functions that resemble atomic orbitals and center them on the atomic centers. Slater-type orbitals (STOs) are an attractive choice because they closely resemble hydrogenic atomic orbitals.[2] However, when using an STO basis set, the need to numerically solve the general four-index integral proves a major limitation for *ab initio* methods. On the other hand, the four-index integrals for Gaussian-type orbitals (GTOs) can be solved analytically. The functional form of GTOs is given below:

$$\varphi(x,y,z;\delta,i,j,k) = \left(\frac{2\delta}{\pi}\right)^{3/4}\left[\frac{(8\delta)^{i+j+k}\,i!\,j!\,k!}{(2i)!\,(2j)!\,(2k)!}\right]^{1/2} x^i y^j z^k e^{-\delta(x^2+y^2+z^2)}$$

Equation 3.25

where δ defines the GTO width and i, j and k determine the axial symmetry. While these functions are convenient from a computational standpoint, they give an incorrect shape of the radial portion of the orbital. Firstly, GTOs are smooth and differentiable at r = 0 where hydrogenic atomic orbitals have a cusp corresponding to a zero probability of finding electron density at the nucleus. In addition, the electron density in hydrogenic atomic orbitals decay as e^{-r} while the decay of GTOs is exponential in e^{-r^2}. Computationally efficient primitive Gaussians are therefore used in linear combination to give a contracted basis function, ϕ, that reproduces the desired STO radial behavior (and so can be equated to an atomic orbital):

$$\phi(x,y,z;\delta,i,j,k) = \sum_{a=1}^{M} c_a\ \varphi(x,y,z;\delta_a,i,j,k)$$

Equation 3.26

In Equation 3.26, contraction coefficients, c_a, control the shape of φ and ensure normalization. The parameters c_a and δ are found by a fitting procedure, and common δs within contraction shells reduce the number of distinct integrals that must be calculated.

The smallest basis set, known as the minimal basis set, contain as many contracted basis functions as required to accommodate the electrons of neutral atoms. Improvements to the minimal basis set aim to impart greater flexibility during the SCF process to find a more accurate electron distribution.[2] The first way to do this is to de-contract the basis functions. Although this can be done for all orbitals, it is more common to split only the valence orbitals that are strongly affected by chemical bonding. These split-valence functions represent core orbitals through a single contracted basis functions, while valence orbitals are split into contracted functions (split-valence double-ζ, triple-ζ, etc.). Coefficients of the inner- and outer-shell basis functions can be varied independently allowing a finer-tuning of the electron distribution and a lower energy. The Pople basis sets[15] are popular split-valence basis sets that have the notation X-YZWG etc., where X denotes the number of primitives comprising each core atomic orbital, while Y,Z… indicate the contraction of each of the inner and outer shells. For example, 6-31G is a split-valence double-ζ basis set where '6' refers to the six primitive gaussians used to describe each core atomic orbital, '3' refers to the three primitives used to construct the inner shell of valence orbitals and '1' denotes an outer shell consisting of a single GTO.

Polarization functions can also be added to increase basis set flexibility. These are functions which have an angular momentum higher than the valence orbitals, and so their contribution allows the electron distribution to be displaced along a particular direction (polarized). An illustration of this effect is shown for water in Figure 3.3. So-called balanced basis sets consider the correspondence between de-contracting the valence basis functions and adding polarization functions.[2] Balanced double-ζ basis sets should include *d* functions on heavy atoms and *p* functions on H, while triple-ζ basis sets should

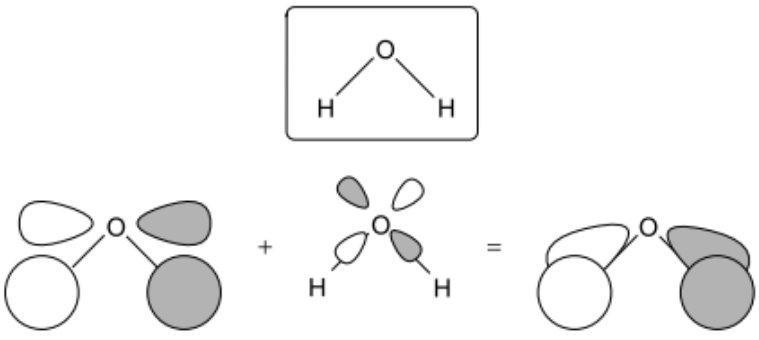


Figure 3.3 Illustration of a more accurate electron distribution obtained for water from the antisymmetric combination of H 1s orbitals and the oxygen p_x orbital. Bonding interactions are enhanced by mixing a small amount of O $d_{x,z}$ character into the molecular orbital. Figure reproduced from Jensen.[2]

include one set of *f* and two sets of *d* functions on heavy atoms, and one set of *d* and two sets of *p* functions on H.

Finally, standard basis sets can be augmented with diffuse functions that have small δ exponents and so fall off very slowly with distance. Weighting c_δs during the SCF cycle can generate electron density at relatively large distances from the nucleus and so simulate the loosely held electrons on heteroatom lone pairs, in anions, and in electronically excited molecules.[4] Using diffuse functions does not guarantee better performance in DFT calculations due to possible introduction of linear dependencies and poor convergence of the SCF equations for larger molecules.[16] In the Pople basis sets, diffuse functions are denoted by '+'.

DFT calculations hold the advantage that basis-set saturation is reached more easily than in wavefunction calculations.[17] In wavefunction methods, calculations come closer to the basis set limit as more functions are added. However, improved performance comes at a computational cost associated with the increase in four-center integrals. DFT, on the other hand, often approaches limiting results with smaller basis sets than for wavefunction calculations, in which case increasing basis-set size does not always improve accuracy. Since the choice of basis set is considered to be relatively minor,[16] smaller basis sets can be used to run calculations that explicitly treat exchange and correlation effects on enzyme systems that are too large for post-HF methods.[4]

3.2 Formulation and Approximations of SCC-DFTB

SCC-DFTB[18] is an approximate method derived from DFT. The SCC-DFTB energy expression is obtained by deconstructing the electron density into a reference density that is perturbed by some fluctuation, $\rho(\boldsymbol{r}) = \rho_0(\boldsymbol{r}) + \delta\rho(\boldsymbol{r})$, and then inserting $\rho(\boldsymbol{r})$ into the KS total energy functional (Equation 3.15):

$$E[\rho_0 + \delta\rho] = \sum_i^n n_i \langle\psi_i \left| -\frac{1}{2}\nabla^2 - \sum_A^N \frac{Z_A}{|\boldsymbol{r} - \boldsymbol{R}_A|} + \int \frac{\rho_0'}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}' + \frac{\delta E_{xc}[\rho_0]}{\delta\rho_0} \right| \psi_i\rangle$$

$$-\frac{1}{2}\iint \frac{\rho_0'(\rho_0 + \delta\rho)}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}d\boldsymbol{r}' - \int \frac{\delta E_{xc}[\rho_0]}{\delta\rho_0}(\rho_0 + \delta\rho)d\boldsymbol{r}$$

$$+\frac{1}{2}\iint \frac{\delta\rho'(\rho_0 + \delta\rho)}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}d\boldsymbol{r}' + E_{XC}[\rho_0 + \delta\rho] + \frac{1}{2}\sum_A^N \sum_{A\neq B}^N \frac{Z_A Z_B}{r_{AB}}$$

Equation 3.27

where $\rho_0' = \rho_0(\boldsymbol{r}')$ and $\delta\rho' = \delta\rho(\boldsymbol{r}')$ are defined as shorthand notations. The second term in Equation 3.27 corrects the double counting in the Coulomb term; the third term corrects the new exchange-correlation contribution; and the fourth term results from splitting the Coulomb energy into

one part related to ρ_0 and another related to $\delta\rho$.[19] The exchange-correlation functional, $E_{XC}[\rho_0 + \delta\rho]$, is then expanded in a Taylor series and truncated after the second-order term:

$$E_{XC}[\rho_0 + \delta\rho] = E_{XC}[\rho_0] + \int \left.\frac{\delta E_{XC}}{\delta\rho}\right|_{\rho_0} \delta\rho d\boldsymbol{r} + \frac{1}{2}\iint \left.\frac{\delta^2 E_{XC}}{\delta\rho\delta\rho'}\right|_{\rho_0,\rho_0'} \delta\rho\delta\rho'\, d\boldsymbol{r}d\boldsymbol{r}' + \frac{1}{6}\iiint \left.\frac{\delta^3 E_{XC}}{\delta\rho\delta\rho'\delta\rho''}\right|_{\rho_0,\rho_0',\rho_0''} \delta\rho\delta\rho'\delta\rho'' d\boldsymbol{r}d\boldsymbol{r}'d\boldsymbol{r}'' \cdots$$

Equation 3.28

Substituting Equation 3.28 into Equation 3.27 and then canceling terms yields the exact (truncated) expression:

$$E[\rho] = \sum_i^n n_i \langle\psi_i|\hat{H}[\rho_0]|\psi_i\rangle + \frac{1}{2}\sum_A^N \sum_{A\neq B}^N \frac{Z_A Z_B}{r_{AB}} - \frac{1}{2}\iint \frac{\rho_0'\rho_0}{|\boldsymbol{r}-\boldsymbol{r}'|} d\boldsymbol{r}d\boldsymbol{r}' - \int \frac{\delta E_{xc}[\rho_0]}{\delta\rho_0}\rho_0 d\boldsymbol{r} + E_{XC}[\rho_0] + \frac{1}{2}\iint \left[\frac{\delta\rho\delta\rho'}{|\boldsymbol{r}-\boldsymbol{r}'|} + \left.\frac{\delta^2 E_{XC}}{\delta\rho\delta\rho'}\right|_{\rho_0}\right] d\boldsymbol{r}d\boldsymbol{r}'$$

Equation 3.29

By grouping the double counting and nuclear repulsion terms into a single energy contribution denoted E^{rep}, Equation 3.29 can be written as:

$$E^{Total} = E^{bnd} + E^{rep} + E^{2nd}$$

Equation 3.30

To accelerate the evaluation of Equation 3.30, the KS molecular orbitals are expanded in an optimized minimal basis set of atomic orbitals, and each of the terms is implemented in an approximate fashion.

In a fundamental assumption, the molecular density is constructed as the superposition of neutral atomic densities, $\rho_0 = \sum_A^N \rho_A$, which forms the basis for calculating E^{bnd} and E^{rep} according to a two-center formulation. Thus, by excluding the kinetic and potential electronic contributions involving three and four centers, the Hamiltonian can be constructed:

$$H_{\mu\upsilon} = \begin{cases} \varepsilon_\mu^{free\ atom} & if\ \mu = \upsilon \\ \left\langle \phi_\mu \middle| -\frac{1}{2}\nabla^2 + \hat{V}_{eff}[\rho_A + \rho_B] \middle| \phi_\upsilon \right\rangle & if\ \mathrm{A} \neq \mathrm{B} \\ 0 & if\ A = B, if\ \mu \neq \upsilon \end{cases}$$

Equation 3.31

The diagonal elements of the Hamiltonian matrix are the DFT energies of the atomic orbital in the free atom, the intra-atomic off-diagonal elements are zero due to the orthonormal basis set, and the interatomic off-diagonal terms are calculated using an effective potential dependent only on the atomic densities of the bonded atoms. Practically, integral evaluation is completely avoided by pre-calculating the Hamiltonian and overlap elements from DFT calculations for every pair of chemical elements in a reference set of molecules. The results are then either fitted to analytic functions or tabulated for discrete interatomic distances for future recall. Such an approach is justified since the matrix elements are a function only of ρ_A^0 which is invariant to the specific chemical environment.

Extending the two-centered approach, E^{rep} is likewise simplified as a pairwise potential dependent only on the reference density:

$$E^{rep} = \frac{1}{2}\sum_{AB} V_{AB}^{rep}$$

Equation 3.32

where V_{ab}^{rep} is again pre-calculated as the difference between E^{bnd} and an appropriate reference DFT result.

Density functional tight binding (DFTB)[20,21] calculates the total energy at this point, including only E^{bnd} and E^{rep} terms. In doing so, a single, non-self-consistent diagonalization of the Hamiltonian matrix is performed to solve the secular equation (in effect modifying only the shape of the KS orbital). As a result, this procedure is only appropriate for systems where the superposition of neutral atomic densities is accurate, and there is no density fluctuation arising from electronegativity differences.[22,23] Charge redistribution is implemented in SCC-DFTB by minimizing the E^{2nd} term that contains charge density and so necessitates solving coefficients for the LCAO self-consistently. In the spirit of the two-center formulations employed for E^{bnd} and E^{rep}, the charge density fluctuation is represented by atomic components, $\delta\rho = \sum_A^N \delta\rho_A$. Charge fluctuations on the atoms are represented in a monopole approximation and the integral is approximated as a sum of pairwise interactions between atomic partial charges ΔQ_A:

$$E^{2nd} = \frac{1}{2}\sum_{AB}^{N} \Delta Q_A \Delta Q_B \gamma_{AB}$$

$$\gamma_{AB}^{h} = \frac{1}{r_{AB}} - S(r_{AB}, U_A, U_B) \times h(r_{AB}, U_A, U_B)$$

Equation 3.33

where Mulliken charges, $\Delta Q_A = Q_A - Q_A^0$, are used for the monopole contributions. The screening factor, γ_{AB}, serves to modulate the charge-charge interaction and interpolate the correct limiting behavior of the Coulomb integral in E^{2nd} (Equation 3.29). The analytical function contains the overlap integral, *S*, that acts over a product of two normalized Slater-type spherical charge densities that take the form:

$$n_A(r) = \frac{\tau_A^3}{8\pi} e^{-\tau_A|\boldsymbol{r}-\boldsymbol{R}_A|}$$

Equation 3.34

At large interatomic distances γ_{AB} reduces to $1/r_{AB}$ and E^{2nd} reduces to the Coulombic interactions between partial charges. For on-site self-repulsion, γ_{AA}^h reduces to the Hubbard parameter, U_A, which is the second derivative of the total energy of a single atom with respect to the occupation number of the highest occupied atomic orbital.[18,22] This is achieved by relating the exponents, τ_A, of the Slater-type spherical charge densities in the integrand of *S* to the Hubbard parameter:

$$\tau_A = \frac{16}{5} U_A$$

Equation 3.35

Thus, the inverse of U_A models the covalent radius of atoms. The inverse relation of chemical hardness and atomic size only holds for elements within a row of the periodic table. A different γ_{AB} should therefore be applied for different rows.[18,24] The superscript h in Equation 3.33 denotes the use of a more repulsive γ_{AB} for bonds to hydrogen.

3.2.1 Limitations of SCC-DFTB

In general, SCC-DFTB reproduces accurate geometries close to that of full DFT-GGA, predicts consistent reaction energies,[25] and, more specifically, model peptide structures and conformation energies well.[26] On the other hand, while SCC-DFTB treats local correlation effects through descent from DFT, it also inherits some deficiencies, including the tendency to overbind covalent bonds and neglecting long-range dispersion interactions.[22,24] The R^{-6} behavior of two separate neutral non-overlapping fragments can be empirically corrected for at large distances using a damping function:[27]

$$E = E - \sum_{A,B} f(r_{AB}) C_6^{A,B} (r_{AB})^{-6}$$

Equation 3.36

where $C_6^{A,B}$ are molecular polarisabilities derived from experimentally determined atomic polarisabilities. When this correction is applied, stacked complexes like DNA base pairs[27] are

realistically modeled. Some further limitations arise from the approximations used to accelerate calculation of the respective energy terms in Equation 3.29. The minimal basis set used in the Hamiltonian matrix and monopole approximation using Mulliken charge evaluation affect the treatment of electrostatic properties.[24] Specifically Mulliken population analysis leads to underestimation of dipole moments.[23]

As a result of these limitations, proton affinities of negatively charged molecules are overestimated, while those for neutral molecules are underestimated. In addition, even when including the γ_{AB}^{h} function, SCC-DFTB underestimates proton transfer barriers (similar to PBE).[24] Finally, given that accurately modeling hydrogen bonding requires treatment of a full range of contributions including electrostatics, charge-transfer effects and dispersion interactions, it is unsurprising that SCC-DFTB is observed to underestimate the strength of hydrogen bonds.[28,29]

3.2.2 Parameterization and Implementations of SCC-DFTB

The available flavors of DFTB are defined by truncation of E_{XC} in Equation 3.28 and the parameter set employed. DFTB calculates the total energy as a function of E^{bnd} and E^{rep} only, DFTB2 (SCC-DFTB) adds E^{2nd} and, more recently, DFTB3 includes the third order term of Equation 3.28 in an approximate fashion as:

$$E^{3rd} = \frac{1}{6}\sum_{ABC} \Delta Q_A \Delta Q_B \frac{d\gamma_{AB}}{dQ_C}\bigg|_{Q_0^C}$$

Equation 3.37

Parameter sets include parameters that are associated with electrostatic or repulsive properties. The first class of parameters includes the wavefunction or density compression radii. These compression radii are used to modify DFT calculations that yield an optimized LCAO basis set and neutral atomic densities for condensed-phase systems.[19] While these parameters allow only a fine-tuning in performance, the repulsive parameters that determine bond energies, bond distances and stretch vibrational frequencies are crucial for accuracy and are fitted to more accurate DFT calculations.[23]

The popular MIO parameter set for organic molecules was developed with DFTB2 by fitting repulsive potentials to reference B3LYP/6-31G(d) calculations.[18] The same parameter set was initially used for the DFTB3 after adding DFTB3-specific parameters. This is a reasonable approach because the extension only influences systems with significant atomic net charges, and was found to improve the description of hydrogen bonded and charged molecules.[28] The 3OB parameters set, intended for use in organic and biological applications, was later parameterized specifically for DFTB3 starting from MIO parameter values.[28] Both the repulsive potentials and electronic parameters were optimized against

bond lengths, vibrational stretching frequencies and atomization energies. 3OB reportedly improves bond distances and energies of noncovalent interactions, and reduces overbinding apparent in DFTB3/MIO. However, test calculations on the TcTS active site with DFTB2/MIO, DFTB3/MIO and DFTB3/3OB showed that DFTB3/3OB in particular gave spuriously low energies for dissociation of the sialic acid glycosidic bond. DFTB2/MIO gave more consistent results and has previously shown favorable structural and energetic comparison with MP2 and B3LYP.[1]

3.3 QM/MM Potential Energy

QM/MM partitioning is commonly used in enzyme free energy calculations to simplify the reaction space. In this scheme, the electronic degrees of freedom are limited to the reaction region while the conformational degrees of freedom for the polarizing protein environment are treated using an MM force field. Interaction between the QM reaction region and MM environment (QM-MM interactions) are typically described by nonbonded electrostatic and van der Waals interactions, as well as any bonded terms crossing the boundary.[9,30-32] The electrostatic interactions can be treated in two ways, each associated with a different total energy expression for the combined QM/MM energy.[9] In mechanical embedding, the QM system is represented by invariant point charges, and QM-MM electrostatics are modeled by pairwise Coulomb interactions. In such an approach the total QM/MM Hamiltonian is subtractive (for example in mechanically embedded ONIOM[33]):

$$E_{QM/MM} = E_{QM}(QM) + E_{MM}(QM + MM) - E_{MM}(QM)$$

Equation 3.38

This scheme can be extended and applied to combinations of high and low levels of QM treatment (QM/QM). Mechanical embedding neglects the influence of the polar environment along the reaction path. This is a severe approximation since a fundamental function of the protein is to electrostatically stabilize the activated complex. An additive QM/MM scheme can be used to account for this limitation:

$$E_{QM/MM} = E_{MM} + E_{QM} + E_{QM-MM}$$

Equation 3.39

In this case, E_{QM} and E_{QM-MM}^{elec} are computed together in a self-consistent manner such that the polarizing effect of MM point charges is included at the QM level via one-electron integrals (so-called electrostatic embedding):

$$E_{QM/MM} = \langle \Psi | \hat{H}_{QM} + \hat{H}_{QM-MM}^{elec} | \Psi \rangle + E_{QM-MM}^{VDW} + E_{MM}$$

Equation 3.40

where

$$\hat{H}_{QM-MM}^{elec} = \sum_{A \in MM} \sum_{B \in QM} \frac{Q_A Z_B}{r_{AB}} - \sum_{A \in MM} \sum_{i=1}^{n} \frac{Q_A}{|\boldsymbol{r}_i - \boldsymbol{R}_A|}$$

Equation 3.41

QM/MM SCC-DFTB in CHARMM is implemented using an additive QM/MM scheme. When the QM-MM electrostatic Hamiltonian is treated analytically, computationally intensive two- and three-center integrals of the form

$$\left\langle \phi_\mu \left| \frac{Q_A}{|\boldsymbol{r} - \boldsymbol{R}_A|} \right| \phi_\nu \right\rangle$$

Equation 3.42

would have to be computed. In the spirit of the approximate approach taken by SCC-DFTB, the QM-MM electrostatic interactions are calculated as the Coulombic interaction between the Mulliken charge of the QM atoms and the MM partial charges:[34]

$$\hat{H}_{QM-MM}^{elec} \approx \sum_{A \in MM} \sum_{B \in QM} \frac{\Delta Q_A \Delta Q_B}{r_{AB}}$$

Equation 3.43

QM-MM van der Waals interactions, which are significant at short distances and for interactions with neutral MM groups, are usually calculated by an MM procedure (12-6 Lennard-Jones function). MM van der Waals parameters are therefore chosen for each QM atom, often from analogous MM functional groups.[9] Although these parameters do not take into account changes in chemical nature of the QM system along reaction progress, it has been shown for SCC-DFTB/CHARMM that condensed phase thermodynamic quantities are not overly sensitive to their values and that other factors such as the treatment of long-range electrostatic interactions have a greater impact on the reliability of simulation results.[32,35]

3.3.1 Long range QM-MM Electrostatic Interactions

The treatment of long range MM-MM and QM-MM electrostatic interactions depends on how the bulk environment is modeled. In the case of periodic boundary simulations, the Particle Mesh Ewald scheme[36] can be used after the system has been neutralized with counter ions. However, when using periodic boundary conditions with the minimum image convention, the box should be large enough so that the solute does not interact with itself and that no water molecules interact with the solute twice. When dealing with macromolecules, solvation in a sufficiently large unit cell requires a large number of water molecules to be included in the simulation. A more tractable approach is to explicitly include

only a partial sphere of the system (protein and solvent atoms) around the active site as in stochastic boundary conditions[37] (see Appendix A for more detail). In this method the bulk solvent is modeled by introducing a deformable boundary potential and stochastic contributions to dynamics. Long-range electrostatic interactions are only included up to a user-specified distance, where after, the charge-charge interactions are truncated with either a shifting or a switching potential.[38] This treatment is defensible since monopole interactions diminish as $1/r^2$, but it is important to ensure that spherical cutoff values include significant electrostatic contributions.

3.3.2 MM Force Field

The reaction zone is influenced by the mobile protein environment that is modeled using an MM energy function and associated parameters. The popular CHARMM energy expression[39,40] consists of the sum over the individual inter- and intramolecular interactions and takes the form:

$$V(\boldsymbol{R}) = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} K_\varphi\big(1 + cos(n\varphi - \delta)\big) +$$

$$\sum_{impropers} K_\omega(\omega - \omega_0)^2 + \sum_{residues} U_{CMAP}(\varphi, \psi) + \sum_{Urey-Bradley} K_{UB}(S - S_0)^2 +$$

$$\sum_{\substack{non-bonded \\ pairs}} \left\{ \varepsilon_{AB}^{min} \left[\left(\frac{r_{AB}^{min}}{r_{AB}} \right)^{12} - 2 \left(\frac{R_{AB}^{min}}{r_{AB}} \right)^{6} \right] + \frac{Q_A Q_B}{4\pi\varepsilon_0 \varepsilon r_{AB}} \right\}$$

Equation 3.44

Accordingly, bond stretching and angle bending are treated as classical harmonic terms that are a function of the current distance from the equilibrium value given in the parameter set, while dihedral angles are modeled using a sinusoidal function. The CHARMM bonded potential function contains three additional terms: an out-of-plane bending term that is used to maintain planar geometries and stereochemistry of chiral centers, the Urey-Bradley term that enforces coupling of bond angle and bond stretch to reduce repulsive interactions between the 1, 3 atoms, and the CMAP correction to protein backbone torsional angles[41]. Nonbonded interactions are modeled using effective pairwise potentials. Dispersion and exchange-repulsion forces comprising van der Waals interactions are included as a 12-6 Lennard-Jones function. Finally, a Coulomb potential governs the electrostatic interaction between invariant point charges localized on atoms. The point charges are partial atomic charges that model the unequal distribution of charge due to electronegativity differences in the real system. Thus, electronic polarization is included only in an average, invariant manner. Despite the promise of next-generation protein force fields with explicit polarization, the CHARMM force field remains a popular choice for simulating protein dynamics. Its simplicity allows long timescale

simulations of large proteins,[9] and the terms in the energy function are physically meaningful. CHARMM22 has been expressly designed for modeling proteins with explicit TIP3P[42] solvation, and has been shown to give reliable results when used with the CMAP correction.[43]

3.3.3 Boundary Atoms

When including protein residues in the QM region, the boundary between the QM and MM regions typically cuts the covalent bond to the protein backbone. The two general techniques used to satisfy the valence shell of the frontier QM atom are to employ a modified atomic orbital basis set for the frontier atom, using, for example, the generalized hybrid orbital (GHO) method, or alternatively adding a fictitious link-atom across the bond linking the QM and MM regions.[44] The fictitious atom is usually a hydrogen, but other atom types have been used, notably halogens. This scheme is simple and widely used, although the implicitly assumed equivalency of C – H and actual C – X bonds, as well as sensitivity to the positioning of the link atom, may introduce inaccuracy.[9] These issues are overcome in the GHO method that uses hybrid orbitals as basis functions on the frontier atom. For example, a frontier sp^3 carbon is expanded into four hybrid orbitals, one of which is included in the SCF cycle, while the other three are treated as auxiliary orbitals. Parameters for the localized orbitals are optimized to reproduce properties of full QM systems, and can be readily transferred. Both the link atom and hybrid orbital approaches can be sensitive to the classical electrostatic field that interacts with the QM region. MM charges are redistributed and can be scaled to avoid spurious results from overestimated charge interactions.[45] Despite these limitations, boundary methods can give reasonable results by choosing a boundary suitably far from chemical changes and not close to highly charged MM groups.[9]

3.4 References

(1) Pierdominici-Sottile, G.; Horenstein, N. A.; Roitberg, A. E. *Biochemistry* **2011**, *50*, 10150.

(2) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*; Second ed.; John Wiley & Sons, Ltd: West Sussex, England, 2004.

(3) Jensen, F. *Introduction to Computational Chemistry*; 2nd ed.; John Wiley & Sons, Ltd, 2007.

(4) Lewars, E. G. In *Computational Chemistry*; Springer Netherlands: Dordrecht, 2011.

(5) Becke, A. D. *J. Chem. Phys.* **2014**, *140*.

(6) Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964**, *136*, B864.

(7) Kohn, W.; Sham, L. J. *Phys. Rev.* **1965**, *140*, A1133.

(8) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. *Science* **2008**, *321*, 792.

(9) Lodola, A.; Mulholland, A. J. In *Methods in Molecular Biology*; Humana Press: Totowa, NJ, 2013; Vol. 924.

(10) Perdew, J. P.; Schmidt, K. *AIP Conf. Proc.* **2001**, *577*, 1.

(11) Perdew, J. P.; Ruzsinszky, A.; Constantin, L. a.; Sun, J.; Csonka, G. b. I. *J. Chem. Theory Comput.* **2009**, *5*, 902.

(12) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.

(13) Devlin, F. J.; Finley, J. W.; Stephens, P. J.; Frisch, M. J. *J. Phys. Chem.* **1995**, *99*, 16883.

(14) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.

(15) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *The Journal of Chemical Physics* **1971**, *54*, 724.

(16) Boese, A. D. *ChemPhysChem* **2015**, *16*, 978.

(17) Bauschlicher, C. W.; Partridge, H. *Chem. Phys. Lett.* **1995**, *240*, 533.

(18) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M. et al. *Phys. Rev. B* **1998**, *58*, 7260.

(19) Oliveira, A. F.; Seifert, G.; Heine, T.; Duarte, H. A. *J. Braz. Chem. Soc.* **2009**, *20*, 1193.

(20) Slater, J. C.; Koster, G. F. *Phys. Rev.* **1954**, *94*, 1498.

(21) Chadi, D. J. *Phys. Rev. Lett.* **1979**, *43*, 43.

(22) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614.

(23) Elstner, M.; Seifert, G. *Phil. Trans. R. Soc. A* **2014**, *372*, 20120483.

(24) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931.

(25) Gaus, M.; Chou, C.-P.; Witek, H.; Elstner, M. *J. Phys. Chem. A* **2009**, *113*, 11866.

(26) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203.

(27) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.

(28) Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013**, *9*, 338.

(29) Domínguez, A.; Niehaus, T. A.; Frauenheim, T. *J. Phys. Chem. A* **2015**, *119*, 3535.

(30) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. *Chem. Soc. Rev.* **2012**, *41*, 3025.

(31) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2006**, *117*, 185.

(32) Senn, H. M.; Thiel, W. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198.

(33) Chung, L. W.; Sameera, W. M. C.; Ramozzi, R.; Page, A. J.; Hatanaka, M. et al. *Chem. Rev.* **2015**, *115*, 5678.

(34) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569.

(35) Riccardi, D.; Li, G.; Cui, Q. *J. Phys. chem. B* **2004**, *108*, 6467.

(36) Ewald, P. P. *Ann. Phys.* **1921**, *369*, 253.

(37) Berkowitz, M.; McCammon, J. A. *Chem. Phys. Lett.* **1982**, *90*, 215.

(38) Steinbach, P. J.; Brooks, B. R. *J. Comp. Chem.* **1994**, *15*, 667.

(39) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. et al. *J. Comput. Chem.* **1983**, *4*, 187.

(40) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J. et al. *J. Comput. Chem.* **2009**, *30*, 1545.

(41) Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400.

(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(43) Mackerell, A. D. *J. Comp. Chem.* **2004**, *25*, 1584.

(44) Pu, J.; Gao, J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 5454.

(45) König, P. H.; Hoffmann, M.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 9082.

4 Simulating Enzyme-Catalyzed Reactions with FEARCF

Computation of reliable enzyme reaction free energies is contingent on the accurate definition and reliable sampling of the reaction space. Importantly, enzyme active sites are surrounded by a thermally fluctuating environment of protein conformational states, and a realistic calculation of chemical reaction rates must take account of this environmental diversity.[1] Chapter 3 discussed the treatment of the reaction space using SCC-DFTB/MM, a Hamiltonian offering the necessary speed to undertake the extensive sampling required. FEARCF[2,3] is now introduced as a method to enhance sampling of the reaction space that is central to addressing the objectives of this thesis.

4.1 Flat Histogram Methods

As introduced in Chapter 1, free energy activation profiles can be accessed within the framework of classical statistical mechanics through the ensemble average of the probability distribution function, $\langle \rho(\xi) \rangle$. Invoking ergodicity, the ensemble average is calculated from the frequency occurrence of configurations meeting the conditions of ξ, $P(\xi)$. However, MD or Monte Carlo simulations at room temperature preferentially sample low energy configurations and a reliable average for $P(\xi)$ will not be obtained for chemical processes traversing large energy barriers (in other words the required sample size becomes too large and all the possible equilibrium distributions of the partition function, Z, will not be accessed[4]). Enhanced sampling methods commonly alter the potential of the system to artificially induce uniform sampling of the altered phase space. In umbrella sampling simulations, the system is restrained about a position of the reaction coordinate such that the restraining function is centered about that position. The function commonly takes the form of a harmonic function. Having knowledge of the restraining function allows a subsequent reweighting of the recorded histograms to recover results corresponding to the original potential.

An alternative strategy seeks equiprobable sampling of states by a macroscopic parameter, such as potential energy, by iteratively searching for the optimal state weights on the fly. Since the sampling does not follow pre-defined paths, these methods have the advantage of not needing an initial estimate of the energy landscape. The chemical intuition is that we do not need a precise knowledge of the reactant and product states, nor the path connecting them. The biasing potential is thus initially unknown, but iteratively improved in generating a flat histogram.

4.1.1 Monte Carlo Implementations

Early work applied this strategy to the Monte Carlo study of a first-order phase transition embodied in the 2-D 10-point Potts model. Under Monte Carlo simulations the ensemble average $\langle A \rangle$ of a quantity A is determined by:

$$\langle A \rangle = \frac{\sum_i A(\boldsymbol{R}_i) P^{-1}(\boldsymbol{R}_i) e^{-\beta V(\boldsymbol{R}_i)}}{\sum_i P^{-1}(\boldsymbol{R}_i) e^{-\beta V(\boldsymbol{R}_i)}}$$

Equation 4.1

where $\boldsymbol{R}_i$ represents a configuration at time i of a given system with a potential function V, and $P(\boldsymbol{R})$ is the probability density sampled. Metropolis sampling aims to compute averages on a system in contact with a heat bath by choosing points with a Boltzmann-weighted probability density:

$$P(\boldsymbol{R}) \propto e^{-\beta V(\boldsymbol{R})}$$

Equation 4.2

The Boltzmann distribution of energies is obtained by imposing the detailed balanced condition:

$$\frac{W(\boldsymbol{R} \to \boldsymbol{R}')}{W(\boldsymbol{R}' \to \boldsymbol{R})} = e^{-\beta[V(\boldsymbol{R}') - V(\boldsymbol{R})]}$$

Equation 4.3

Berg and Nehaus[5] partitioned the total action, S, of the Potts model into $k = 0, \cdots, N$ intervals of internal energy. Then, instead of sampling with the Boltzmann factor corresponding to the canonical ensemble, configurations were sampled with the weight:

$$P(S) = e^{-(\alpha^k + \beta^k)S} \; for \; S^k < S \leq S^{k+1}$$

Equation 4.4

In Equation 4.4, β^k is varied by choosing different temperatures (thus 'multicanonical' ensemble) and α ensures that $P(S)$ is a steady function of the entropy. These parameters were tuned to obtain a flat behaviour of the action density, whereupon the canonical probability density could be recovered by appropriate reweighting in order to calculate the free energy profile. Therefore, the multicanonical approach consists of separate weight-determining and production periods. Lee applied a more general Monte Carlo algorithm, based on sampling the system's entropy, to the Potts model.[4] It was noted that the partition function can be written as a sum over energy states:

$$Z(\beta) = \sum_i e^{-\beta V(\boldsymbol{R}_i)} = \sum_E e^{S(E)-\beta E}$$

Equation 4.5

where $S(E)$ is the entropy for a given energy, E. An arbitrary distribution $P(E) \propto e^{A[E]} = e^{S(E)-J(E)}$ can be obtained by imposing the following balanced condition:

$$\frac{W(\boldsymbol{R} \to \boldsymbol{R}')}{W(\boldsymbol{R}' \to \boldsymbol{R})} = e^{-[J(\boldsymbol{R}')-J(\boldsymbol{R})]}$$

Equation 4.6

Metropolis sampling becomes a particular case of the Monte Carlo algorithm embodied in Equation 4.6 where $J(E) = \beta E$. If on the other hand, $J(E) = S(E)$, a uniform distribution of internal energy is obtained. Therefore, finding the sampling probability that allows exploration of the entire energy space in one Monte Carlo will allow determination of an accurate $S(E)$. The entropy is then used to calculate the partition function and free energy profile. The procedure used by Lee to calculate the rough estimate of $S(E)$ in the weight-determining stage consisted of cycles of the following iterations:

(1) Initially $J(E) = S(E)$ is set to zero for all E.
(2) The histogram $H(E)$ in energy space is accumulated from a short Monte Carlo run where a trial move is accepted according to Equation 4.6. For the first iteration, this is just a random sampling.
(3) New estimate of $S(E)$ is given now as follows:

$$S(E) = \begin{cases} J(E)\ for\ H(E) = 0 \\ J(E) + \ln H(E)\ otherwise \end{cases}$$

Equation 4.7

The uniform sampling obtained in the above algorithms implies a sampling distribution that is the inverse of the density of states (Ω), where Ω is defined as the number of states per interval of energy that are available to be occupied, and is related to the partition function by:

$$Z(\beta) = \sum_i e^{-\beta V(\boldsymbol{R}_i)} = \sum_E \Omega(E)\, e^{-\beta E}$$

Equation 4.8

Density of states Monte Carlo methods introduce a biasing potential, which is directly the running best estimate of the inverse $\Omega(E)$. The Wang-Landau algorithm,[6,7] which has been applied to complex systems such as proteins, polymers and bulk liquid crystals,[8] comprises a series of stages over which the density of states is successively approximated with increasing precision by performing random

walks in energy space. Random changes of the energy are accepted with a probability that is proportional to the reciprocal of the density of states:

$$\frac{W(E_1 \to E_2)}{W(E_2 \to E_2)} = e^{-[\ln(\Omega(E_2))-\ln(\Omega(E_1))\,]}$$

Equation 4.9

At each step the density of states is modified by a multiplicative factor $f > 0$ such that $\ln(\Omega(E_2)) = \ln(\Omega(E_2)) + \ln(f)$ if the Monte Carlo step is accepted and $\ln(\Omega(E_1)) = \ln(\Omega(E_1)) + \ln(f)$ otherwise. The updated density of states is then used to perform a further random walk in energy space. In this way, each iteration generates a flat histogram of the energy distribution with an accuracy proportional to f. The modification factor is initially chosen to be large enough to reach all possible energy levels quickly but without introducing large statistical errors. Once a flat histogram in energy space is achieved, the histogram is reset and f is reduced. At the end of the simulation, f should be very close to 1, which corresponds to the ideal case of the random walk with the true $\Omega(E)$.[8]

4.1.2 MD Implementations

Comparative non-Boltzmann sampling schemes have been developed for MD simulations where ensemble configurations are connected in time. Thus, the Hamiltonian or Lagrangian is suitably modified to achieve equiprobable sampling for each state defined by ξ. These methods have been used to explore enzyme reactions.[2,9,10] A memory-dependence algorithm as first applied by local elevation[11] is central to the metadynamics[12] method. In this 'hills' approach, adaptive Gaussian functions informed from past simulations, are added at regular time intervals to improve sampling across high-energy regions in a manner similar to the Wang–Landau flat histogram. The Gaussian terms are dependent on the current configuration of the reaction coordinates $S(S_1, S_2, \cdots, S_d)$, also known as the collective variable (CV). Thus, as terms are added, the system is prevented from sampling previous configurations and explores new areas of phase space:

$$V_G(S(\boldsymbol{R}), t) = h \sum_{\substack{t'=0,\tau_G,2\tau_G,\cdots \\ t'<t}} \exp\left(-\sum_{\alpha=1}^{n} \frac{\left(S_\alpha(\boldsymbol{R}) - s_\alpha(t)\right)^2}{2\delta s_\alpha^2}\right)$$

Equation 4.10

In Equation 4.10, the repulsive Gaussian is centered on the current configuration of the CV, h specifies the Gaussian height, and δs_α is the Gaussian width specified for each S_α. Once equal sampling is achieved and the system becomes diffusive in CV-space, the PMF is formulated as the negative of the sum of repulsive functions. The relation $\lim_{t\to\infty} V_G(S,t) \approx -W(S)$ does not derive from any standard

identity for the free energy, such as the umbrella sampling equality or the perturbation free energy formula, and was postulated heuristically.

The accuracy of the scheme in reconstructing the free energy is strongly correlated to the dimension of the CV and the Gaussian width.[13] Importantly, all the relevant slow varying variables must be taken into account. If any slow relaxing degree of freedom coupled to the reaction is not taken into account, the biasing potential does not converge. However, the time required to escape from a local minimum is proportional to $(1/\delta s)^d$. Therefore, the CV dimension and the values of δs have to be carefully chosen to strike the best balance between accuracy and sampling efficiency. There are different flavors of metadynamics, and new developments concentrate on finding new and more general CVs, alongside merging the metadynamics algorithm with available sampling techniques to compensate for the limitation on the number of the composite variables.[14]

4.2 FEARCF

In the earliest implementation of the flat histogram strategy by an MD technique, the FEARCF algorithm sought to evolve the partition function, in multiple geometric dimensions, from an ensemble of simulations.[2,3,15,16] In contrast to the biasing potential constructed from repulsive Gaussians in metadynamics, FEARCF modifies the classical nuclear motion by including the gradient of the current PMF estimate as an external driving force in the potential function (Figure 4.1). As a result of applying the numerically differentiated driving forces directly in 3-D Euclidean space the need for a Jacobian correction is made redundant.

4.2.1 Running a Single FEARCF Iteration

The reaction coordinate space is a discretized *n*-dimensional grid where the sampling frequency for each k^{th} bin site is recorded. To run a single i^{th} iteration, an ensemble of j individual reaction dynamics trajectories are evolved under the influence of the external potential, U_i. The bias is defined as the inverse of the current estimate of the PMF:

$$U_i(\xi) = -W_{i-1}(\xi) = k_B T \ln P_{i-1}(\xi)$$

Equation 4.11

The sampling frequency generated from the ensemble of reactive trajectories, typically 30 ps in length, is recorded to construct probability histograms, p_k. The contributions to the unbiased probability distribution are weighted and the individual histograms from the current iteration i and all previous iterations $(1, 2, \cdots, i-1)$ are combined using WHAM. This is an extension of the histogram procedure developed by Ferrenberg and Swendsen,[17,18] where probability distribution histograms for all simulations overlap optimally to give an accurate, unbiased probability distribution $P_i(\xi)$:

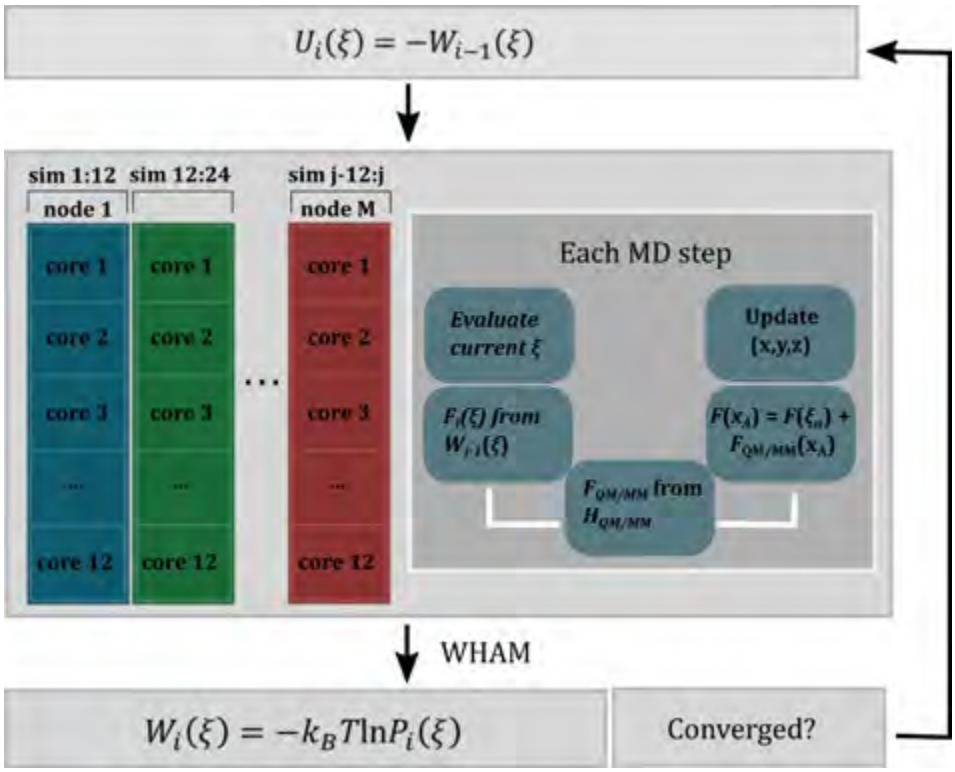


Figure 4.1 Schematic representation of the FEARCF algorithm. The central rectangular box represents the i^{th} iteration, wherein j MD simulations are run on K cores under the influence of the external potential, U_i. In this example K = M ×12 cores which is also equal to the number of simulations in each iteration. A successful pass of the convergence test terminates the FEARCF. Figure adapted from Naidoo.[9]

$$p_k = \frac{\sum_i n_{i,k}}{\sum_i N_i f_i e^{-\beta U_i(\xi_k)}}$$

Equation 4.12

In Equation 4.12 the total number of configurations N_i is obtained by summing over the number of configurations in all k bins of the discretized reaction coordinate:

$$N_i = \sum_k n_{i,k}$$

Equation 4.13

and the free energy weighting factors are defined as:

$$f_i = \frac{1}{\sum_k e^{-\beta U_i(\xi_k)} p_k}$$

Equation 4.14

Since the probability appears in the expression of the weighting factors, the set of equations are solved self-consistently. The resulting unbiased $P_i(\xi)$ is then used to update the external potential:

$$U_i(\xi) = k_B T \ln P_i(\xi)$$

Equation 4.15

The forces calculated from reaction coordinate histograms for all i iterations are then used to drive iteration $i+1$ toward the unexplored regions of reaction space. In this way, the full reaction coordinate space is explored in an unrestrained manner to iteratively yield an improved estimate of $W(\xi)$.

4.2.2 Obtaining a Converged Surface

In order to obtain accurate reaction free energy profiles (illustrated in Figure 4.2), the user defines a reaction coordinate appropriate to encompass the complexity of the reaction. Each bond that is formed or broken during a reaction can be allocated to a specific reaction coordinate or can be merged with other bonds (as chemically appropriate) in a combined reaction coordinate. The user also specifies the number and length of the j individual MD simulations that comprise each iteration.

Since the FEARCF external potential is unknown, no force is applied in the first iteration and the simulations are equivalent to equilibrium dynamics (Figure 4.2A). Histograms are constructed to gain an unbiased reaction coordinate probability density and an initial estimate of the PMF (Figure 4.2B). The estimate of the PMF is then refined over the next iterations in which the bins of the discretized

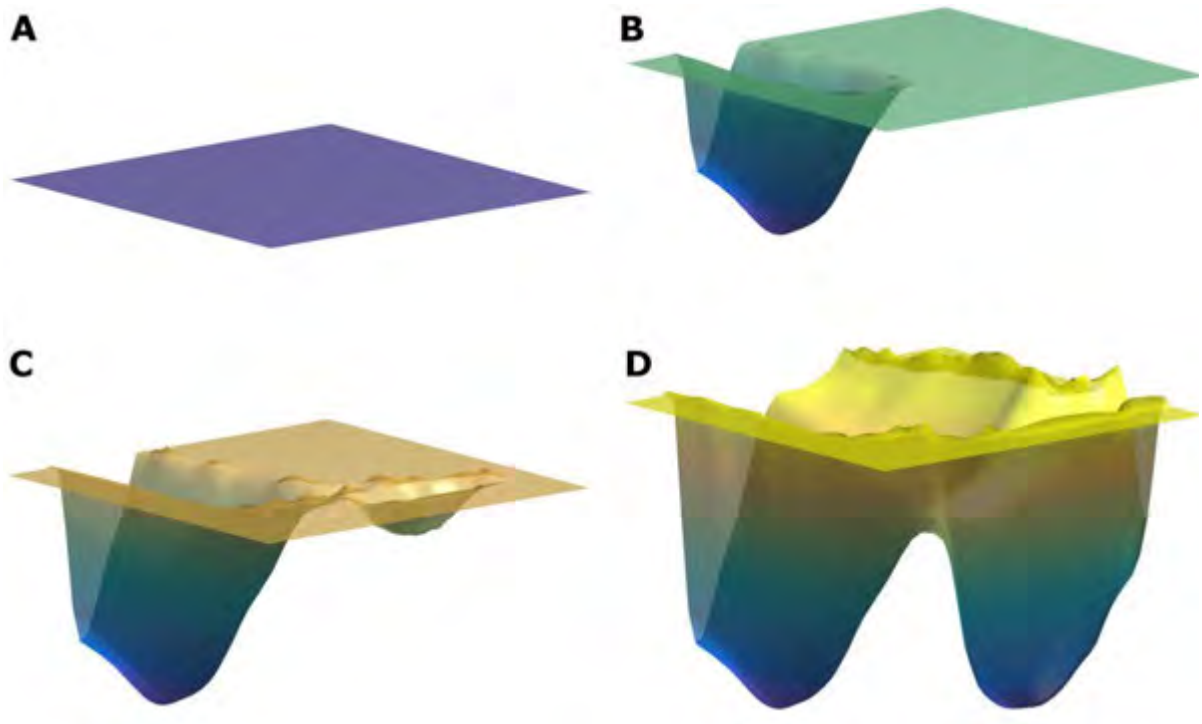


Figure 4.2 Snapshots of the FES resolved for $\xi(\xi_1, \xi_2)$. (A) The PMF is zero at all points in iteration zero, and is iteratively improved (B, C) until it converges (D).

reaction coordinate are cumulatively populated from the sampling of the reaction coordinate (Figure 4.2C and D). A flat histogram, and by definition a converged free energy surface, is achieved when equal ratio of sampling of the global minimum compared with that of the TS is observed – although practically, a convergence criterion of 1:50 has been used.[19-21] At this point the biasing forces are derived from the true PMF and there is effectively a flat, even potential with no barriers to cross. In this case, long trajectories will undergo multiple barrier recrossings between reactant and product wells. This reaction dynamics scheme possesses a number of desirable properties, including the observation of rare-event crossing trajectories and a simple extension to multiple dimensional progress variables including, but not limited to, distances and angles.

4.2.3 Representative Crossing Trajectories

The coupling of protein motions to the rearrangement of chemical bonds is included by averaging over an ensemble of reaction dynamics trajectories that do not restrain the protein or substrate. This is in contrast to metadynamics and umbrella sampling schemes. These methods inadvertently produce an equilibrated environment about the reacting substrate molecules. As the PMF is built from the reactant well in the FEARCF scheme, successful evolution from reactants to products is obtained. Electronic and conformational information can be extracted from these dynamic sketches to provide a representative description of the ensemble of physical reactive trajectories. In addition, collecting an ensemble of representative trajectories affords an ensemble of activated complexes that approximates the TS. In this way, the first objective of this thesis will be addressed in Chapter 5 by generating and analyzing the chemical nature of the TcTS deglycosylation TS.

4.2.4 Calculating Free Energy Volumes

Enzymatically catalyzed reactions are invariably dependent on a number of correlated events that need to be treated as components of a higher-dimensional reaction coordinate. The calculation of FEARCF driving forces, $F_i(\xi)$, as the gradient of $W_{i-1}(\xi)$ allows simple extension to a multi-dimensional geometric reaction coordinate, $\xi(\xi_1, \xi_2, \cdots)$. The value of ξ is calculated for each step of the reaction dynamics trajectories, and is used to compute driving forces $F(\xi_\alpha)$ from the negative partial derivative of the probability density, or simply the partial derivative of the PMF with respect to ξ_α:

$$F_i(\xi_\alpha) = \frac{\partial W_{i-1}(\xi)}{\partial \xi_\alpha}$$

Equation 4.16

where the gradient of the PMF is obtained using a cubic spline routine. Forces in Cartesian space, $F_i(x_A)$, are computed for atom A included in the definition of ξ_α using the chain rule:

$$F_i(x_A) = \frac{\partial W_{i-1}(\xi)}{\partial \xi_\alpha} \frac{\partial \xi_\alpha}{\partial x_A}$$

Equation 4.17

Thus, the free energy profile for complex reactions where multiple bonds are formed and broken can be directly computed to produce reaction free energy volumes/hypervolumes from which the reaction mechanism may be derived.[9]

4.3 Parallelized FEARCF Reaction Dynamics Trajectories

In principle, these forces are transferable to reaction dynamics run at alternative levels of theory. $F_i(x_A)$ will drive a second system along a crossing path. In this way, the semi-empirical PMF can be used as the initial estimate for refinement by a DFT/MM Hamiltonian, or forces derived from the PMF can be used to run single crossing trajectories in order to avail accurate molecular orbital interactions along the reaction pathway. SCC-DFTB/MM FEARCF simulations are currently run by averaging over ensembles of single trajectories, each run on a single core. However, the increase in task size associated with running DFT trajectories necessitates fine-graining of the parallelism scheme. That is, instead of an embarrassingly parallel distribution of each reaction dynamics trajectory to a single compute node, the trajectories themselves need to be parallelized for the *ab initio* FEARCF calculation to be tractable. For this reason, FEARCF was implemented as a Fortran 90 library that can be interfaced with NWChem to leverage the performance offered by the computational package on high-performance architectures.

4.3.1 Acceleration of DFT Calculations in NWChem

DFT/MM simulations in NWChem are run through a top-level interface between the DFT and MM modules that manages initialization, data transfer and various high-level operations. Optimization and dynamics calculations proceed by cycles of QM and then MM calculations in the presence of the fixed environment provided by the alternate level of theory. The QM/MM module therefore takes advantage of the efficiently parallelized implementations of QM and MM energy evaluations in NWChem. Communication between remote processors in each of these lower-level modules is carried out using the Global Arrays toolkit, which is a partitioned global address space (PGAS) programming model.

The most time-consuming operations of the DFT SCF calculation are the construction of the Fock and Exchange-correlation matrices and their diagonalization. NWChem implements a parallelized

calculation of explicit diagonalization using PeIGS software.[22] The separation of the treatment of Coulomb and exchange contributions to the DFT electronic energy is exploited to further accelerate the construction of the electronic matrices. Building the Fock and E_{xc} matrices are bound to O(N^4) by quartet integrals, although practically this is reduced by applying molecular and permutation symmetry, as well as integral screening.[23] Integrals are calculated employing a distributed-data approach outlined in Listing 4.1. Accordingly, they are organized into blocks with similar shell characteristics via a twofold blocking routine and computed simultaneously with the TEXAS integral package[24] using the Obara-Saika (OS) method.[25] The integrals with similar shell characteristics are calculated together to enable sharing and reuse of temporary data. Since the computational cost of integrals differs depending on their angular momentums, a dynamic task counter is used to distribute blocks of integrals to parallel workers to ensure that all parallel workers perform equal amounts of integral computations. Then, in a lower blocking level, the integrals are packaged into smaller blocks considering the cache size and the memory requirements of the integral shell characteristics.

```
my_task = global_task_counter(task_block_size)
current_task = 0
do ijkl = 2*ntype, 2, -1
      do ij = min(ntype, ijkl-1), max(1, ijkl-ntype), -1
            kl = ijkl – ij
            if (my_task .eq. current_task) then
                  call calculate_integral_block()
                  call add_integrals_to_Fock()
                  my_task = global_task_counter(task_block_size)
            endif
            current_task = current_task + 1
      enddo
enddo
```

Listing 4.1 Pseudo code, reproduced from Shane et al.,[23] outlining dynamic load balancing through a two-fold blocking scheme.

4.3.2 Acceleration of MM Calculations in NWChem

The bottleneck in MM calculations arises in the evaluation of energies and forces from nonbonded contributions. NWChem employs domain decomposition in order to leverage massively parallel architectures to accelerate this step. The molecular system is decomposed into rectangular prisms which are distributed as sub-grids to a logically arranged grid of processes (see Figure 4.3 for a two-dimensional representation). A pair list of sub-boxes is generated for only three faces of each prism based on the nonbonded cutoff radius and size of sub-box. Each processor then sequentially processes the sub-box pairs and accumulates force contributions to the appropriate array on the remote processor at each MD step. Dynamic load balancing is implemented both through sub-box resizing or sub-box pair list redistribution to reduce the workload of the busiest processor.

4.3.3 Interface of FEARCF Library with NWChem

During dynamics, the NWChem MD module calls the QM/MM module that carries out the accumulation of the atomic forces into the appropriate global arrays. These forces are then used by the MD module to evolve the next time step. FEARCF was written as a Fortran 90 library, and implemented in NWChem 6.5 as a high-level routine that modifies the QM/MM forces before each MD step. Coordinates and forces are collected into a global array instantiated in the MD module, and then passed to the FEARCF library. The library evaluates the current value of the reaction coordinate, ξ_α, and calculates atomic forces in the x-, y- and x-direction from the splined derivative $\partial W(\xi)/\partial\xi_\alpha$. The Cartesian forces are used to modify the original QM/MM forces, that are then recollected in the NWMD module and redistributed to the correct processors so that the simulation may continue in

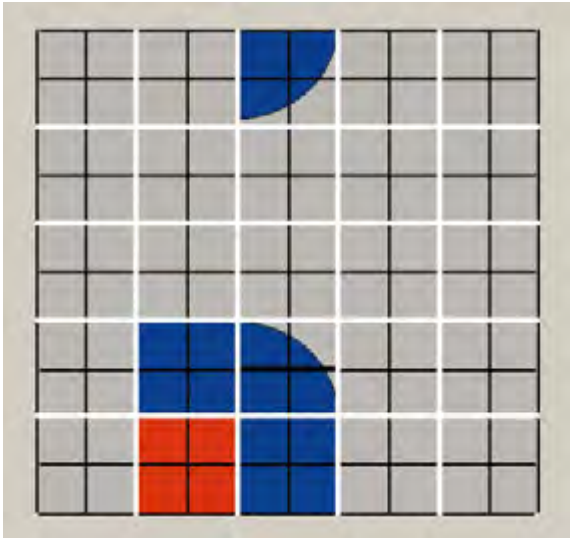


Figure 4.3 Two-dimensional representation of domain decomposition used to accelerate the evaluation of MM nonbonded interactions. The processor in red owns a subset of four sub-boxes and evaluates nonbonded interactions for neighboring atoms within the cutoff radius (two sub-boxes in this example). Interactions with only half of the neighboring sub-boxes need to be considered. Figure reproduced from Straatsma and McCammon.[26]

parallel. The advantage of this scheme is that each of the j simulations comprising the i^{th} FEARCF iteration is calculated in parallel. *Ab initio*/MM simulations can be run across multiple nodes using InfiniBand communication, that is characterized by high throughput and low latency (Figure 4.4).

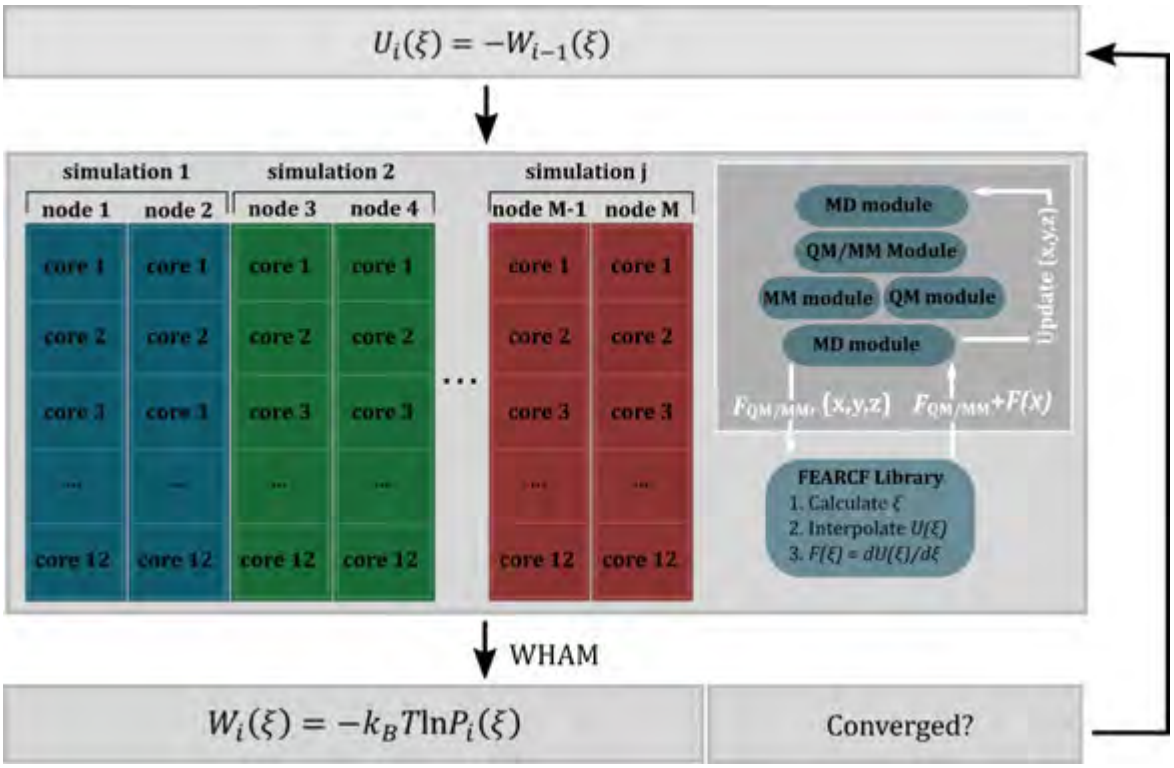


Figure 4.4 The modified FEARCF algorithm. The central rectangular box represents the i^{th} iteration, wherein j MD simulations are run on K cores under the influence of the external potential, U_i. In this example each reaction dynamics trajectory is parallelized over multiple processes so that while K = M ×12 cores, j = M/2.

4.4 References

(1) Masgrau, L.; Truhlar, D. G. *Acc. Chem. Res.* **2015**, *48*, 431.

(2) Naidoo, K. J. *Sci. China: Chem.* **2011**, *54*, 1962.

(3) Naidoo, K. J.; Johan, S. *J. Comp. Chem.* **2009**, *31*, 308.

(4) Lee, J. *Phys. Rev. Lett.* **1993**, *71*, 211.

(5) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9.

(6) Wang, F. G.; Landau, D. P. *Phys. Rev. E* **2001**, *64*.

(7) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.

(8) Landau, D. P.; Tsai, S.-H.; Exler, M. *Am. J. Phys.* **2004**, *72*, 1294.

(9) Naidoo, K. J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9026.

(10) Ensing, B.; De Vivo, M.; Liu, Z. W.; Moore, P.; Klein, M. L. *Acc. Chem. Res.* **2006**, *39*, 73.

(11) Huber, T.; Torda, A.; van Gunsteren, W. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 695.

(12) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A* **2002**, *99*, 12562.

(13) Laio, A.; Gervasio, F. L. *Rep. Prog. Phys.* **2008**, *71*, 126601.

(14) Sutto, L.; Marsili, S.; Gervasio, F. L. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 771.

(15) Naidoo, K. J.; Brady, J. W. *J. Am. Chem. Soc.* **1999**, *121*, 2244.

(16) Barnett, C. B.; Naidoo, K. J. *Mol. Phys.* **2009**, *107*, 1243.

(17) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1988**, *61*, 2635.

(18) Ferrenberg, A. M.; Swendsen, R. H. *Phys. Rev. Lett.* **1989**, *63*, 1195.

(19) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2010**, *114*, 17142.

(20) Bartels, C.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 865.

(21) Bartels, C.; Karplus, M. *J. Comp. Chem.* **1997**, *18*, 1450.

(22) Fann, G. I.; Littlefield, R. J.; Elwood, D. M. In *High performance computing 1995: Grand challenges in computer simulation*; Tentner, A., Ed.; Phoenix: Society for Computer Simulation: 1995, p 329

(23) Shan, H.; Austin, B.; De Jong, W.; Oliker, L.; Wright, N. J. et al. In *High Performance Computing Systems. Performance Modeling, Benchmarking and Simulation*; Jarvis, S. A., Wright , S. A., Hammond, S. D., Eds. 2014; Vol. 8551, p 261.

(24) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251.

(25) Obara, S.; Saika, A. *J. Chem. Phys.* **1986**, *84*, 3963.

(26) Straatsma, T. P.; McCammon, J. A. *IBM Syst. J.* **2001**, *40*, 328.

5 *Profiling Transition State Configurations on the* Trypanosoma cruzi trans-*Sialidase Free Energy Reaction Surface*[†]

Characterizing the mechanism of enzyme catalysis is at the frontier of computational biology and holds exciting potential for applications in TS analog-based inhibitors and enzyme engineering.[1] However, the basic principles of catalysis are still not fully understood and theories include stabilization of the TS, a spatial-temporal view (with the Near Attack Conformation [NAC] hypothesis as a notable proponent), and dynamic motion of the enzyme promoting transition from reactant to product.[2] The predominant theory, first proposed by Pauling, attributes the remarkable efficiency of enzymes to the preferential binding of the TS, encompassing both electrostatic and geometric complementarity.[3] The enzymatic lowering of the free energy of activation between the reactant and TS is often associated with the efficiency of catalysis. It is assumed that a Boltzmann distribution of states of the equilibrated enzyme:substrate complex exists and that the distribution contains a greater number of configurations at the TS. The increase in population of TS configurations is surmised to ultimately result in a faster reaction rate. This relationship between reaction rate and the free energy activation barrier is the framework within which TST is set.

The Eyring formulation is derived by assuming that i) a quasi-equilibrium between the reactant and TS exists and that ii) every TS configuration converts to product. However, a realization that not all events crossing the barrier contribute to product formation required the introduction of a correction factor for failed crossings, the so-called transmission coefficient (κ). It is the combination of the lowering of the activation free energy and changes in the transmission coefficient (recrossing of the TS, tunneling, and non-equilibrium contributions) that improves the accuracy of enzyme reaction rate prediction.[2] The first order catalytic rate constant is then given as:

$$k_{cat} = \kappa \frac{k_B T}{h} \exp\left[\frac{-\Delta G^{\ddagger}}{RT}\right]$$

Equation 5.1

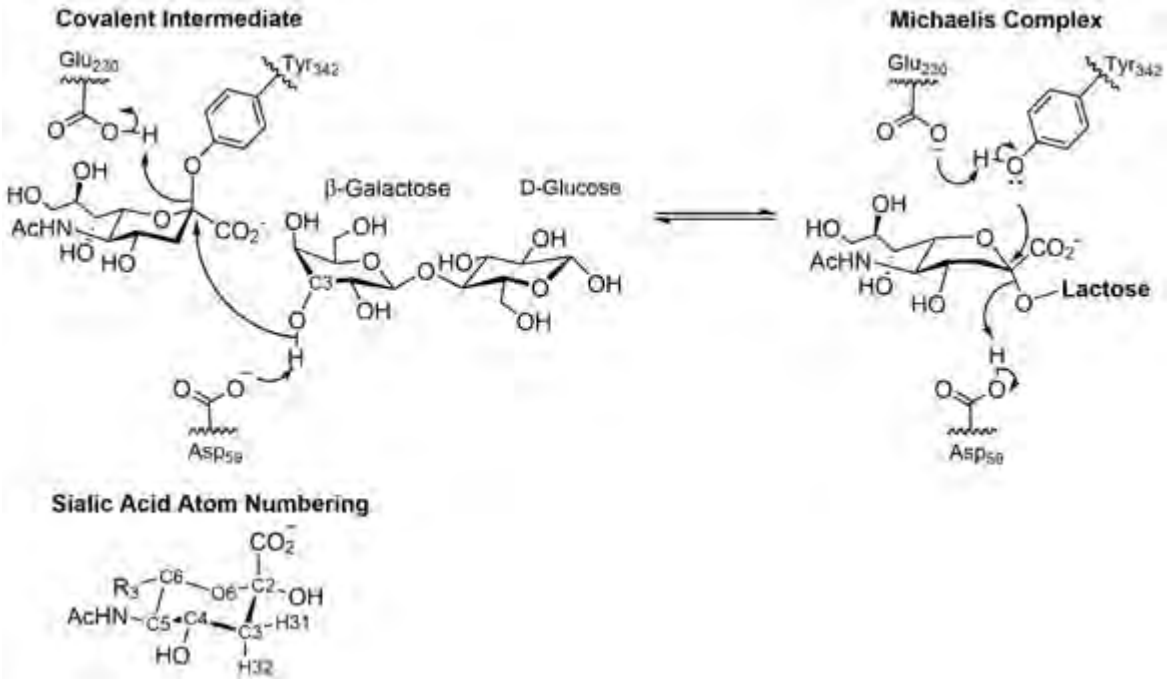
where $\Delta G^{\ddagger}$ is the free energy difference between the reactants and TS. The generalized TS is the free energy bottleneck for the reaction, and is located at the reaction coordinate where the free energy path is a maximum. The TS structure is found here by optimizing the remaining $6N - \xi$ degrees of

[†] The work presented in this chapter has been published in Rogers, I. L.; Naidoo, K. J. *J. Phys. Chem. B* **2015**, *119*, 1192.

freedom to a first-order saddle point on the potential energy surface. It is within this context that the TS profile is analyzed for the glycosyltransferase reaction catalyzed by TcTS.

Chapter 1 has discussed the wealth of experimental data available for TcTS built from the enzyme's significance as a potential drug target against Chagas disease. The sialidase appears to fulfill a number of important functions in the course of *T. cruzi* infection, one of which is to coat the parasite's membrane with sialic acid units.[4-6] TcTS does this by catalyzing the transfer of α(2,3)-linked sialic acid from host sialyl glycoconjugates to β-galactosyl glycoconjugates. While there is room for debate on the nature of the TcTS catalytic mechanism for efficient sugar transfer,[6] the current opinion leans in favor of a transferase reaction undergoing a classical Ping-Pong mechanism as is illustrated by recent umbrella sampling simulations.[7] In that restrained dynamics study, lactose was used as the sialic acid donor/acceptor molecule and a $\Delta G^{\ddagger}$ of 20.80 ± 0.7 kcal/mol was computed for the free energy path connecting the Michaelis complex and covalent intermediate wells. A dissociative A_ND_N pathway, which agrees with experimental kinetic isotope effect (KIE) results, was obtained for the formation of the covalent intermediate by averaging the bond lengths comprising two conflated reaction coordinates. The chemical step for deglycosylation of TcTS is viewed as the reverse of covalent intermediate formation as presented in the Ping-Pong mechanism. Thus, the deglycosylation reaction profile was compared to an increased barrier of 26.1 kcal/mol for the competing covalent intermediate hydrolysis to explain an observed preference for the transferase reaction.[8]

Here, the sialic acid transfer from the sialylated TcTS (covalent intermediate) to a lactose acceptor molecule is considered to be the forward reaction (Scheme 5.1). The complete FES for the deglycosylation step is constructed by following breaking of the (SA C-2)…(Tyr_{342} O) bond and forming of the (Gal O-3)…(SA C-2) bond. As such, the equilibrated sialyl-TcTS:lactose system was perturbed with driving forces obtained from the current estimate of the PMF to freely explore unsampled regions of reaction phase space. A detailed analysis of the TS structures from failed, crossing and recrossing events was performed. The characteristics of a successful crossing configuration are defined based on ring pucker and the conformations of the catalytic binding site.



Scheme 5.1 Reaction mechanism for the deglycosylation reaction that results in the transfer of SA from TcTS to a lactose acceptor molecule The galactose C-3 label marks the location for the α(2,3) bond to SA in sialyllactose. The SA atom numbering is given for reference.

5.1 Computational Methods

Reaction dynamics occurs in a concert of molecular motion taking place across short time and length scales in the presence of high noise to signal ratios. A complete characterization of enzymatic reactions requires an understanding of all the significant degrees of freedom that contribute to the transition coordinate, and not only in the free energy difference between initial (reactant) and final (product) states of the system. It has been recently pointed out that the changes to the protein conformation are key to the understanding of the drivers that evolve the reaction.[9] Considering that the enzyme's action is to lower the free energy of activation barrier as well as improve the transit of well configured structures and conformations about the TS, a computational dynamics method is used that neither restrains the protein nor the substrate in the successful evolution of the reaction from reactants to products via the TS.

The effects of the rotational, translational and importantly vibrational contributions to the partition function can be calculated for the stationary points of the potential energy function within the framework of the rigid-rotor, harmonic oscillator approximation.[10] A reaction path finding scheme can be used to locate the stationary points for a single protein structure and the entropic contributions can be included to calculate the free energy.[11-14] The free energy differences can then be related to experimental values using TST. While this strategy has the advantage that it can be used together with high level *ab initio* methods, it also has several limitations discussed in Chapter 1. These are primarily

associated with the neglect of multiple protein conformations along the enthalpic pathway and across the barrier separating reactants and products. These changes in the protein conformation significantly affect the FES if there is an accompanied change in protein-ligand interaction mode.[15] In addition TS structures are optimized structures in a constrained environment while the free energy bottleneck for passage from the reactant to product state includes entropic contributions from the protein and substrate itself.

It is therefore more desirable to monitor free energy changes for the complete reaction phase space, which can be achieved from probability distributions within the framework of classical statistical mechanics. Methods that enhance the sampling of enzyme reaction space include restrained dynamics schemes[16,17] and flat histogram methods,[18,19] often combined with histogram reweighting methods,[20,21] such as WHAM,[22,23] to obtain an estimate of the unbiased free energy. In umbrella sampling, the sampling of the reaction phase space is enhanced by adding a biasing potential to the system's Hamiltonian.[24-26] If the negative value of the PMF is used as a biasing potential, an accurate estimate of the free energy would yield a uniform distribution of configurations along the reaction coordinate. Since the PMF is not known before the start of the calculation, the use of windowed harmonic potentials is employed to restrain the system's reaction coordinates along the reaction path. This procedure inadvertently produces an equilibrated environment about the reacting substrate molecules that is not representative of the rapidly progressing reaction and the slower enzyme conformational motions that may not be in equilibrium with the substrate.

An alternative approach is the use of flat histogram methods[19,27] that explore numerous physical states in a single run. Since the sampling does not follow pre-defined paths, these methods have the advantage of not needing an initial estimate of the energy landscape. In the FEARCF method, introduced in detail in Chapter 4, the potential function is modified by including the gradient of the PMF as a driving force in an iterative process. At each iteration of the simulation the driving forces, $F(\xi_\alpha)$ for $\alpha = 1, 2, \cdots, n$, are applied to the atoms involved in the definition of the reaction coordinate.[28] For multi-dimensional surfaces the reaction coordinate force is computed from the negative partial derivative of the probability density or simply the partial derivative of the PMF with respect to the reaction coordinates, ξ_α:

$$F(\xi_\alpha) = \frac{\partial W(\xi)}{\partial \xi_\alpha}$$

Equation 5.2

which is being adapted with each sampling update. An equal ratio of sampling of the global minimum compared with that of the TS (the physically rarest event) implies a flat histogram has been achieved

and, by definition, a converged FES. The FEARCF routine therefore consists of a number of iterations for which the PMF gradient from the previous iteration is applied to the reaction coordinate as a perturbing force driving the system 'uphill' to unsampled regions of phase space. This has the advantage that single trajectories are observed that start at the reactant and cross the TS to the product.

5.1.1 Michaelis Complex Preparation

The forward (deglycosylation) and reverse (glycosylation) reactions catalyzed by TcTS were simulated using equilibrated systems derived from the experimental Michaelis complex crystal structure. TcTS has been crystallized with sialyllactose in the active site by mutation of the putative catalytic acid/base residue Asp_{59} to Ala_{59} (PDB accession code 1S0I).[29] The Michaelis complex structure was resolved at 1.6 Å. This structure was analyzed using the MolProb[30] and PropKa[31,32] web servers to assess the orientation of the Asn, His and Gln residues, and assign the protonation state of the titratable amino acid side chains. The TcTS:sialyllactose complex was initialized in CHARMM35b2,[33,34] which was used for all simulations, and then solvated in a truncated octahedral box of TIP3P water molecules[35] to a density of 1 g/cm^3. The all-atom CHARMM22/CMAP force field parameters[36] were used to treat protein atoms, and the CHARMM36 force field parameters[37] were used to model the sialyllactose molecule. Long-range Coulomb interactions were calculated using Ewald summations,[38,39] while a force shifting function applied on an atom-by-atom basis was used to zero the van der Waals forces between atoms further than 12 Å apart (see Appendix A for more background information on the simulation methods used).

Prior to MD simulations, the solvated Michaelis complex structure was minimized over a number of steps during which initial restraints on all heavy atoms were removed firstly from sialyllactose and side-chain atoms, and then from backbone atoms. Next, random velocities were assigned according to a Gaussian distribution to simulate a temperature of 248.15 K. The system was heated to 298.15 K by slowly reassigning velocities every 1 ps so as to avoid any large forces arising from possible steric clashes in the crystal structure. The Michaelis complex structure was then subjected to 300 ps of classical NPT dynamics using a time step of 1 fs and employing the SHAKE[40] algorithm to dampen unstable X – H bonds where X = C, O, N. The enzyme equilibrated as seen from the protein backbone RMSD (Figure D1, Appendix D).

5.1.2 QM/MM Simulations

SCC-DFTB[41] as implemented in CHARMM was used with the MIO parameter set and including both hydrogen bonding[42] and dispersion interaction[43] corrections. SCC-DFTB is formalized as a second-order expansion of the DFT total energy functional with respect to the charge density fluctuations $\delta\rho(r)$

around the reference density $\rho(r)$. The scheme treats polarization between atoms with different electronegativities, an effect neglected in earlier DFTB. SCC-DFTB implemented in a QM/MM algorithm has comparable computational speed albeit generally slower than other popularly used semi-empirical methods such as the MNDO, AM1, and PM3. Mostly of interest here is that SCC-DFTB has been validated against *ab initio* calculations for the current reaction.[7] Furthermore, SCC-DFTB been shown to produce reliable carbohydrate ring puckering behavior,[44] which is a particularly important characteristic for the reactions under investigation. The active site was equilibrated with SCC-DFTB/MM MD under stochastic boundary conditions using a 40 Å sphere cut from the equilibrated water box. The QM region included the side chain atoms of Tyr_{342}, Glu_{230} and Asp_{59}, as well as the substrate sialic acid and galactose residues. The modeling scheme used here moves the QM-MM boundary suitably far from the reaction site at the glycosidic bond. The QM-MM boundary atoms were treated as GHOs.[45] The system was again heated before being simulated for 500 ps using Langevin dynamics (active site RMSD shown in Figure D1, Appendix D).

5.1.3 FEARCF Simulations

The deglycosylation starting point was selected from the 500 ps SCC-DFTB/MM equilibration of covalent intermediate conformations (active site RMSD shown in Figure D1, Appendix D) extracted from the glycosylation reaction. The (SA C-2)…(Gal O-3) and (SA C-2)…(Tyr_{342} O) bonds were chosen as reaction coordinates ξ_1 and ξ_2 (Figure 5.2A). Protonation of the glycosidic linkage and proton transfer between Tyr_{342} and Glu_{230} requires only the available thermal energy and no assistance from a biasing force. Consequently, these bonds were excluded from the reaction coordinate definitions. The PMF had the dimensions 0.5-6.5 Å × 0.5-6.5 Å and both reaction coordinates were partitioned into discrete bins comprising a spacing of 0.1 Å. In each iteration i (comprising an ensemble of simulations), driving forces derived from the PMF (computed from collective $P(\xi_1, \xi_2)$) gained from iterations 1 through $i - 1$, were used in every (30 ps) simulation.

5.1.4 Analysis of TS Ensemble

TS conformations found at the free energy bottleneck for the passage from covalent intermediate to Michaelis complex were analyzed. The bottleneck for a 2-D PMF describing an idealized reaction is given by the *col*, drawn as a bold black line in Figure 5.1A. The TS ensemble is formed from the system reaching different positions along this dividing line, rather than from an equilibrium distribution as assumed by TST. Three event types can be described at the *col*: the system falls back down to the reactant (Figure 5.1B), the system crosses to the product state (Figure 5.1C), and the system crosses the *col* but then recrosses to the reactant state at a later time (Figure 5.1D).

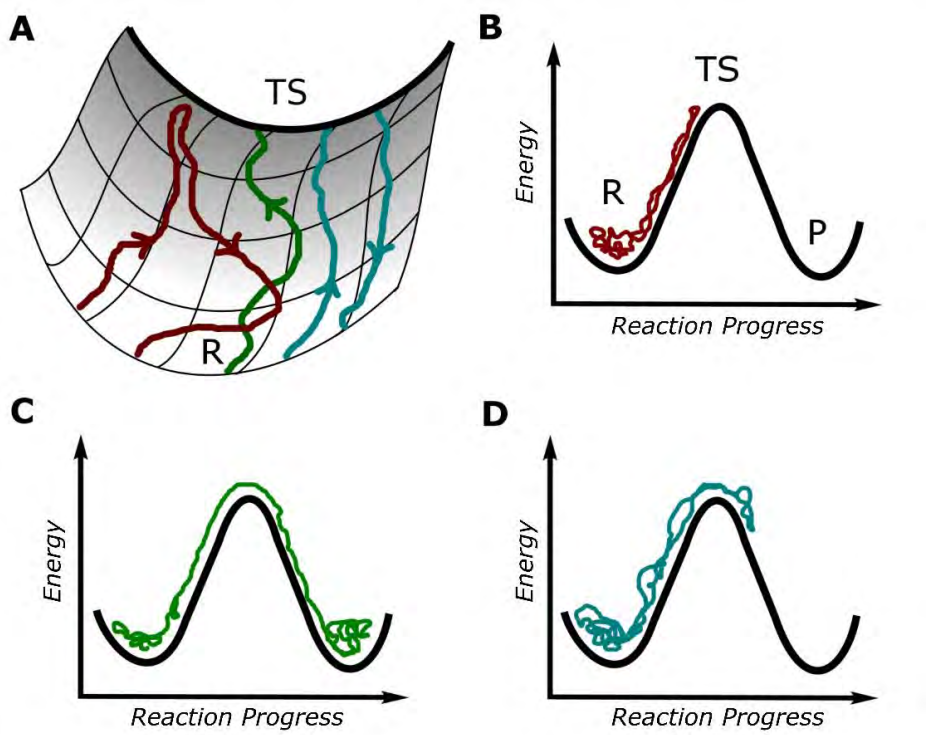


Figure 5.1 (A) Illustration of an idealized reaction col *(bold line) adapted from Anslyn and Dougherty,[46] and shown with example failed (red), crossing (green) and recrossed (cyan) trajectories. The events are shown in profile in (B)-(D).*

In the present work bluemoon-type simulations[47] were used to determine which structures from an ensemble of *col* structures cross and proceed to product formation (i.e., covalent intermediate to Michaelis complex). The *col* structures gained from FEARCF simulations that were initiated from the covalent intermediate (forward reaction) and Michaelis complex (reverse reaction) wells were extracted and used as starting conformations for these unbiased rare event (bluemoon) dynamics. Random velocities taken from Gaussian distributions centered at 298.15 K were assigned to each system configuration to initiate the trajectories. Starting structures for bluemoon trajectories ending up in the product (Michaelis complex) wells were assigned to *col* bins separated by 0.1 Å if their reaction coordinate values were 0.05 Å from the corresponding values of the closest bin. This gave 150 *col* structure configurations with 40 of these structures populating the lowest free energy bin (i.e., TS structure). The bin ensembles were analyzed for pucker, angle of nucleophilic attack and hydrogen bonding. In the case of the latter the criteria used were a D – A distance of 3.0 Å and a D – H – A angle of 150°.

5.2 Results and Discussion

The FES for the deglycosylation of TcTS, which is the transfer of sialic acid from sialylated TcTS to lactose acceptor, was resolved. The starting structure was selected from the QM/MM equilibration of final structures extracted from the simulation of the glycosylation step (Michaelis complex to covalent intermediate). Parallel trajectories were then driven by FEARCF forces through the nearest free energy bottleneck to sample the Michaelis complex state. The transition is achieved by breaking the covalent bond between SA C-2 and the Tyr_{342} leaving group, and forming a glycosidic C – O bond between SA C-2 and the lactose acceptor O-3 nucleophile (Figure 5.2A). During the course of this process the sialic acid ring changes from a $^2C_5$ pucker in the covalent intermediate to adopt a $B_{2,5}$ pucker in the Michaelis complex. In addition, an associated change in the C-4 hydroxyl interaction was observed. While our simulations show that this group is hydrogen bonded to deprotonated Asp_{59} in the covalent intermediate, this interaction is replaced by hydrogen bonding to Asp_{96} in the Michaelis complex (Figure 5.2A and B; D-A distance and D-H..A angle time series are shown in Figure D2 in Appendix D for the SCC-DFTB/MM equilibration of the two complexes). Reweighting the probability distributions obtained from the parallel FEARCF trajectories yields the reaction FES shown in Figure 5.2C. The free energy maximum for the reaction is found at a (SA C-2)…(Tyr_{342} O) bond length of 2.5 Å and a (SA C-2)…(Gal O-3) bond length of 2.2 Å with an associated energy of 19.8 kcal/mol. Thus, the reaction mechanism falls somewhere between A_ND_N and $D_N{*}A_N$ under the IUPAC nomenclature,[48] with the SA – Leaving Group length smaller than 3 Å but significantly longer than 2.0 Å. Locating this point on a More O'Ferrall-Jencks plot [49,50] (Figure 5.2D) gives a graphical interpretation of the dissociative nature of the reaction in which the (SA C-2)…(Tyr_{342} O) bond order is only ~0.2 at the TS.

Within the framework of TST, the Eyring equation (Equation 5.1) can be used to compare the experimental rate constant, k_{cat}, with computational results. Apparent k_{cat} values have been determined for the two-step TcTS-catalyzed reaction using PNP – SA and CF_3MU – SA donor substrates by inferring a Ping-Pong Bi-Bi kinetic mechanism.[51] If it is assumed that all TS complexes are converted to product, the experimental values of 6.77 s^{-1} and 15.77 s^{-1} correspond to free energy barriers of 15.8 kcal/mol and 16.3 kcal/mol respectively, which compare favorably with our computed barrier of 19.8 kcal/mol for the rate-limiting step using lactose as the acceptor molecule. In reality, not all configurations reaching the TS are converted to product, and characterization of features that ensure success should reveal important insight into the mechanism of action of TcTS. These factors conceivably include the geometric properties of the atoms constructing the reaction coordinate as well as protein-substrate interactions.

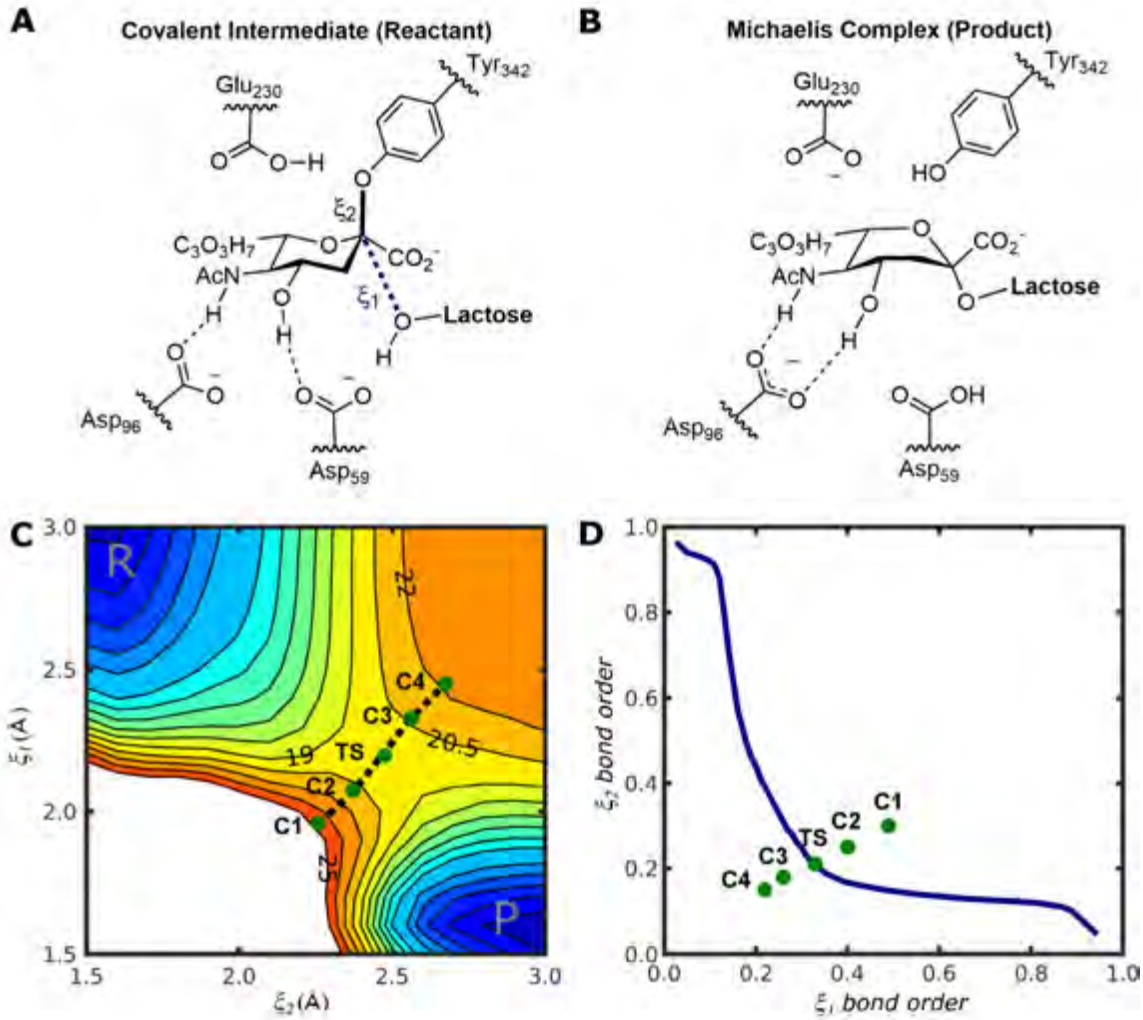


Figure 5.2 The reaction coordinates are shown for the covalent intermediate in (A) while the product complex, Michaelis complex, is given in (B). The FES (C) is shown as a contour map with energy levels spaced by 1.5 kcal/mol. The TS col *is drawn as a black line passing through the free energy maximum, and* col *bin centers displayed in green and labeled C1-C4. In (D) the bin centers are plotted together with the MEP calculated from the gradient of the free energy in a More O'Ferrall-Jencks diagram.*

A *col* ensemble for analysis could be extracted from the reaction trajectories generated in calculating the FES. These trajectories will include paths that approach the *col*, move about the *col* or cross to the Michaelis complex and yield a unique sampling of TS conformations not obtainable from a single trajectory. Furthermore, early FEARCF trajectories, which cross on the femtosecond time scale, will include trajectories for which the slower-moving protein environment has not fully equilibrated (as assumed by TST). In an initial analysis, to probe the effect of the protein environment, structures that failed to cross the *col* were equilibrated while fixing atoms involved in bond breaking and formation. It was found that after constrained equilibration a selection of the failed structures consistently progressed to the Michaelis complex via an unbiased MD protocol. Inspection of these structures showed very subtle differences in protein environment, but a change in interaction mode with the SA C-8 hydroxyl was observed in at least one case. This is illustrated in Figure 5.3 which shows a TS

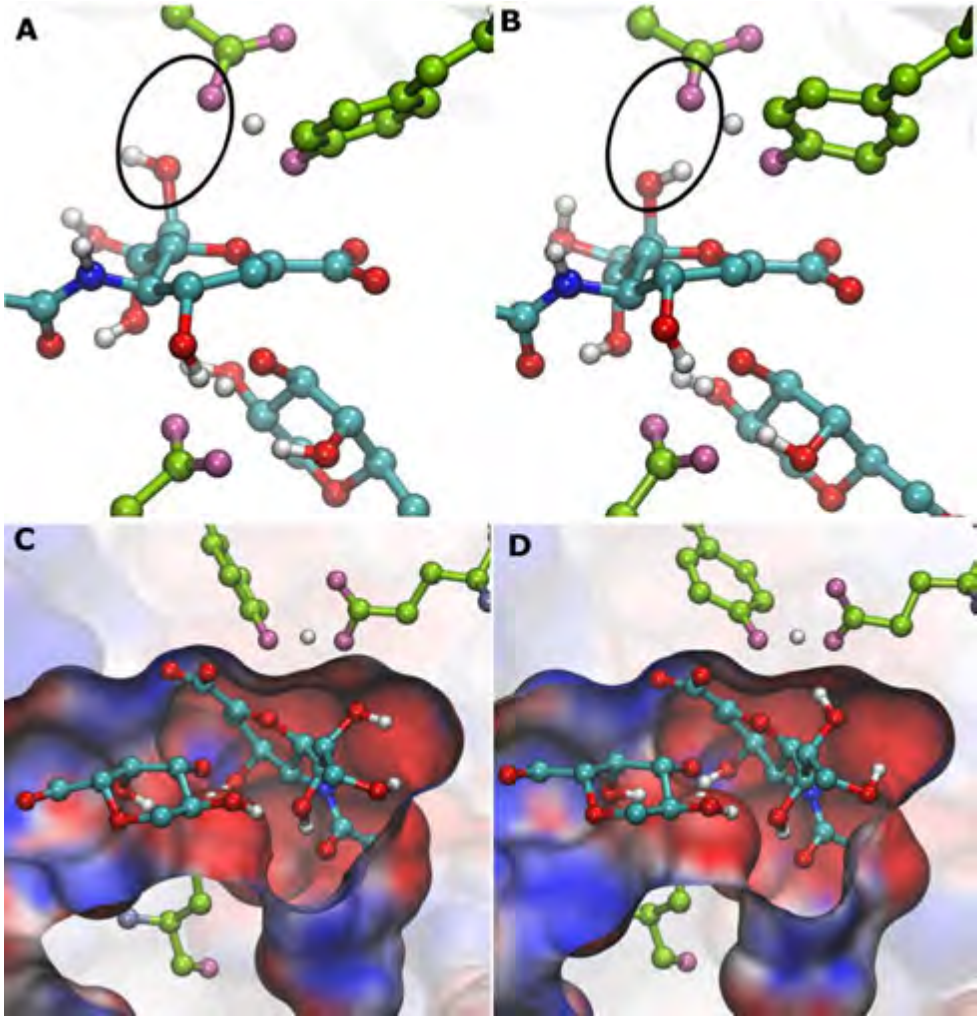


Figure 5.3 An illustrative TS complex shown before (A) and after (B) constrained equilibration. The comparison shows a difference in orientation of the SA C-8 hydroxyl group, either away or towards Glu_{230} of the nucleophile pair. The corresponding electrostatic surface potentials, as calculated by the Adaptive Poisson-Boltzmann Solver (APBS),[52] for appropriate MM and SCC-DFTB Mulliken partial charges, are displayed for the active site in (C) and (D) from -5 (red) to 5 (blue) k_BTe^{-1}.

configuration before and after equilibration. The RMSD of the protein 6 Å from the QM region was 0.65 Å, however one can see a difference in interaction mode with the sialic acid. After equilibration, the C-8 hydroxyl hydrogen points towards Glu_{230} of the nucleophile pair. This interaction conceivably stabilizes proton transfer from Glu_{230} in the covalent intermediate to the Tyr_{342} leaving group.

This observation prompted profiling of the protein-substrate interactions and associated pucker for TS structures along the PMF dividing line or reaction *col* (described in detail in Section 5.1.4). The pucker distribution of the conformations that progress to the Michaelis complex is dominated by E_5, $^4H_5$ and $^6H_5$ (Figure 5.4). These structures are closely related and occur on the same latitude of the Crèmer-Pople sphere denoting the canonical pucker conformers for pyranose rings. In the highlighted pucker conformations, the SA C-5 is positioned below the plane of the pyranose ring. Such a distribution is consistent with the transformation from $^2C_5$ (covalent intermediate) to $B_{2,5}$ (Michaelis complex), where

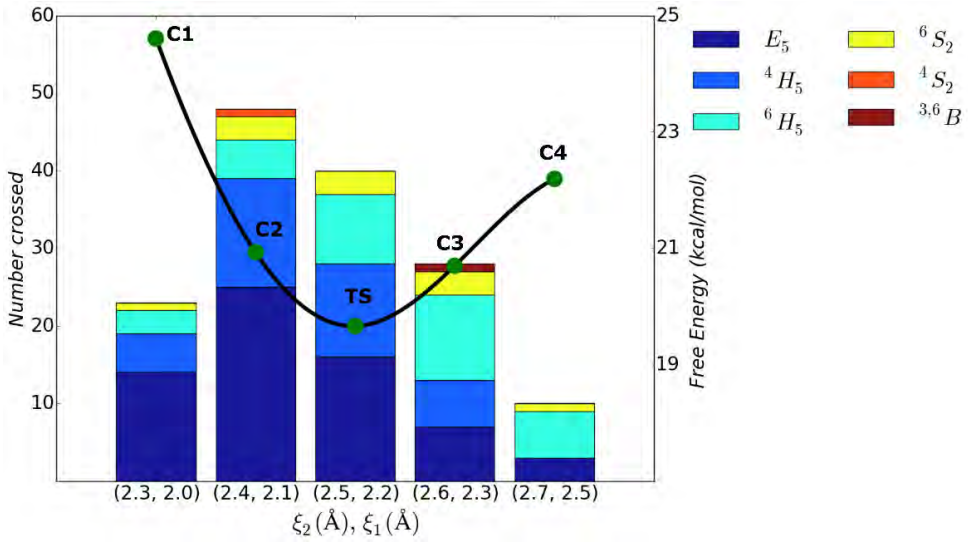


Figure 5.4 Pucker distribution for TS conformations that crossed to the Michaelis complex are shown as stacked bars. The black line represents the reaction col *and the location of the bin centers are labeled.*

the C-5 acetylamido group is held by hydrogen bonding to Asp_{96} and reaction proceeds with C-2 migration. The $^4H_5$ conformation has previously[7] been identified as the TS conformation for the transition to Michaelis complex to covalent intermediate, as well as for other enzyme-catalyzed glycosylation reactions.[21,53] In addition, the $^6S_2$ pucker conformation, closely related to $B_{2,5}$, has been highlighted as a plausible TS or intermediate conformation for glycosylation reactions.[54]

Hydrogen bonding interactions were characterized for the most prevalent puckers in the minimum bin (Figure 5.5A), and the results given in Figure 5.5B. The profiles show a difference in important interactions for sialic acid with the Glu_{230}/Tyr_{342} pair and catalytic base Asp_{59}. There is a high frequency of SA O-4 – Asp_{96} hydrogen bonding amongst the E_5 and $^4H_5$ pucker conformations. On the other hand, $^6H_5$ puckers do not display this interaction, but do show increased hydrogen bonding between SA O-8 and Glu_{230}. The profile of TS structures which cross to the Michaelis complex shows a conformation between the $^2C_5$ and $B_{2,5}$ puckers with one or both of the highlighted hydrogen bonding interactions being satisfied. It is noted that the relative frequency of the $^6H_5$ puckers along the *col* bins increases with (SA C-2)…(Tyr_{342} O) bond lengths. One possibility is that the hydrogen bonding between sialic acid and TcTS reduces the basicity of Asp_{59} perturbing the reaction to a later TS. The identification of different hydrogen bonding patterns points to the importance of dynamic protein-substrate interactions in defining the reaction free energy profile for the covalent intermediate to Michaelis complex transformation.

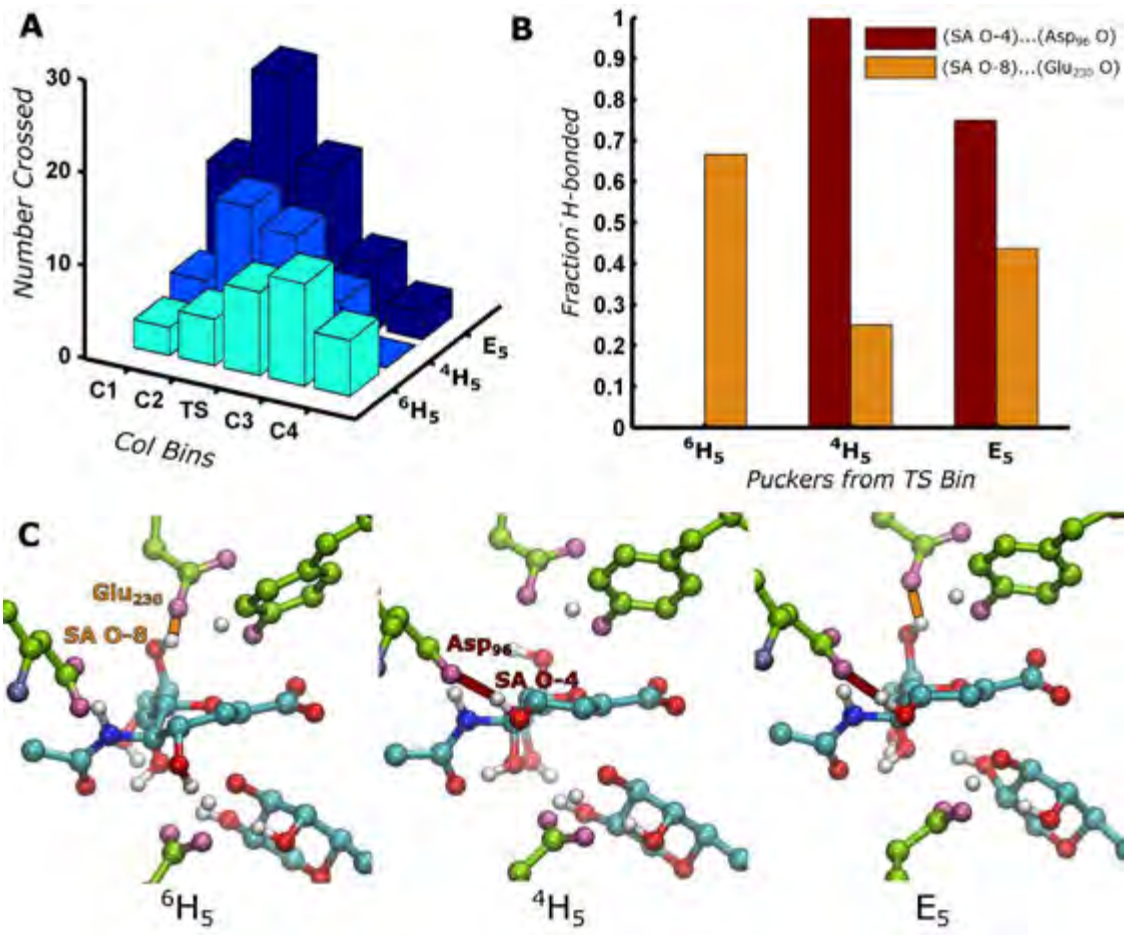


Figure 5.5 The dominant SA ring puckers for the TS ensemble shown as 3D bars in (A). The different hydrogen bonding profiles associated with the puckers in the TS bin are highlighted in (B) and representative snapshots for each are shown in (C). The red and orange colors of the hydrogen bonds displayed in the molecular view correspond to those in the bar graph of (B).

Finally, monitoring the pucker over part of a representative FEARCF trajectory (Figure 5.6) delineates the relationship between pucker and the molecular events. The (SA C-2)…(Tyr_{342} O) dissociation and (Gal O-3)…(SA C-2) bond formation occurs within 1 ps with no intermediate observed. Plotting the ring puckering frequency alongside that of the bonds shows the transition from $^{2}C_{5}$ to $B_{2,5}$. A TS region is highlighted in gray and shows population of E_5, $^{4}H_5$ and $^{6}H_5$ puckers with $^{6}S_2$ present at the final stage of transition to the Michaelis complex. Fluctuation of hydrogen bond D – A distances in Figure 5.6B shows the dynamic nature of the protein-substrate interactions during reaction progress. It can be observed that the TS region is dominated by the $^{6}H_5$ pucker. Consistent with the ligand-protein interaction observed for this sialic acid ring conformation, the successful crossing is accompanied by the formation of a (SA O-8)…(Glu_{230} O) hydrogen bond that is apparent from the shortening of the D – A distance from > 4.0 Å to < 3.0 Å (orange line in Figure 5.6B).

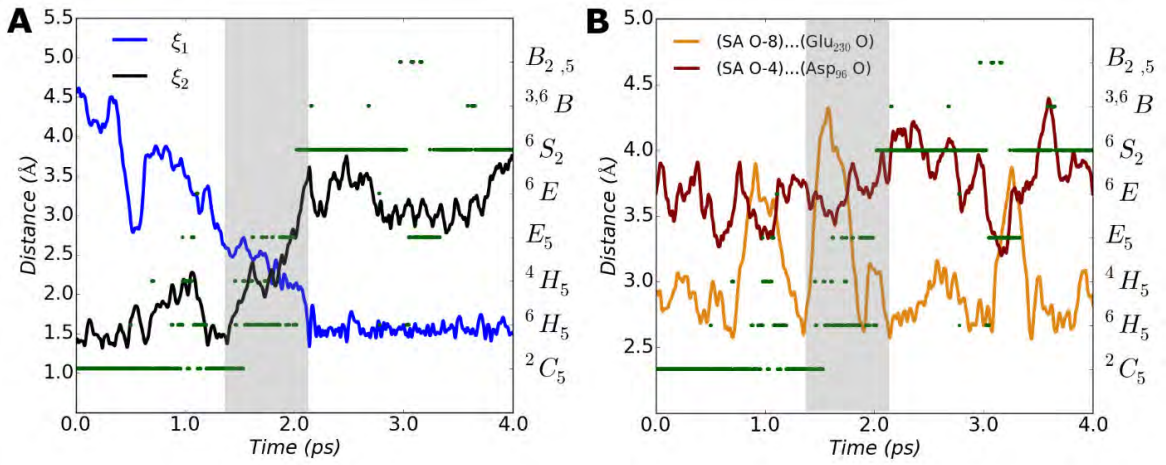


Figure 5.6 A representative FEARCF crossing trajectory. (A) Time series of the reaction coordinates alongside SA ring puckers (green markers) show the transition from $^{2}C_5$ to $B_{2,5}$. The gray block marks the transition from covalent intermediate to Michaelis complex. (B) Hydrogen bond D – A distances show the interaction of SA O-8 with Glu_{230} O during reaction progress.

5.3 Concluding Remarks

The free energy of activation has been calculated for the deglycosylation TS and phenomenologically determines k_{cat} within the context of canonical TST. In addition, the complete reaction FES was resolved as a function of the breaking, (SA C-2)…(Tyr$_{342}$ O), and forming, (SA C-2)…(Gal O-3), bonds. Using a method that constructs the free energy from ensembles of reactive trajectories has allowed chemical profiling of the ensemble of activated complexes. In contrast to the TST assumption that all TS configurations progress to products, these trajectories include paths that approach the *col*, move about the *col* or cross to the Michaelis complex and yield a unique sampling of TS conformations not obtainable from a single trajectory. Trajectories that successfully transform from the covalent intermediate to the Michaelis complex in TcTs aided catalysis were investigated, and it is found that there are multiple pathways that pass over the reaction *col*. The successful crosses follow a pattern where the breaking and the forming bonds each vary 0.1 Å about the TS configuration. Further, the sialic acid takes on one of three puckers (E_5, $^{4}H_5$ or $^{6}H_5$) in which each of the ring conformers are hydrogen bonded to Asp$_{96}$ (E_5 and $^{4}H_5$) or Glu$_{230}$ ($^{6}H_5$). Subtle changes in the protein-substrate interactions are needed to facilitate a successful crossing. The combination of specific pucker, substrate hydrogen bonds and relaxed protein conformations at the *col* can be forced through an equilibration process. Methods employing restrained equilibration may not observe the combination of events detailed here. This detailed study of a progression of configurations along the reaction pathways is aided by the use of dynamic flat histogram methods such as FEARCF.

5.4 References

(1) Schramm, V. L. *J. Biol. Chem.* **2007**, *282*, 28297.

(2) Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, *303*, 186.

(3) Pauling, L. *C&EN* **1946**, *24*, 1375.

(4) Schenkman, S.; Jiang, M. S.; Hart, G. W.; Nussenzweig, V. *Cell* **1991**, *65*, 1117.

(5) Paris, G.; Cremona, M. L.; Amaya, M. F.; Buschiazzo, A.; Giambiagi, S. et al. *Glycobiology* **2001**, *11*, 305.

(6) Freire-de-Lima, L.; Oliveira, I. A.; Neves, J. L.; Penha, L. L.; Alisson-Silva, F. et al. *Front. Immunol.* **2012**, *3*, 356.

(7) Pierdominici-Sottile, G.; Horenstein, N. A.; Roitberg, A. E. *Biochemistry* **2011**, *50*, 10150.

(8) Bueren-Calabuig, J. A.; Pierdominici-Sottile, G.; Roitberg, A. E. *J. Phys. Chem. B* **2014**, *118*, 5807.

(9) García-Meseguer, R.; Martí, S.; Ruiz-Pernía, J. J.; Moliner, V.; Tuñón, I. *Nat. Chem.* **2013**, *5*, 566.

(10) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Sausalito, California, 2000.

(11) Glowacki, D. R.; Harvey, J. N.; Mulholland, A. J. *Nat. Chem.* **2012**, *4*, 169.

(12) Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J. et al. *Angew. Chem. Int. Ed.* **2006**, *45*, 6856.

(13) Lodola, A.; Mulholland, A. J. In *Methods in Molecular Biology*; Humana Press: Totowa, NJ, 2013; Vol. 924.

(14) Senn, H. M.; Thiel, W. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198.

(15) Hammes, G. G.; Benkovic, S. J.; Hammes-Schiffer, S. *Biochemistry* **2011**, *50*, 10422.

(16) Kästner, J.; Senn, H. M.; Thiel, S.; Otte, N.; Thiel, W. *J. Chem. Theory Comput.* **2006**, *2*, 452.

(17) Dimelow, R. J.; Bryce, R. A.; Masters, A. J.; Hillier, I. H.; Burton, N. A. *J. Chem. Phys.* **2006**, *124*, 11413.

(18) F., L. *J. Stat. Phys.* **2006**, *122*, 511

(19) Wang, F. G.; Landau, D. P. *Phys. Rev. E* **2001**, *64*.

(20) Ensing, B.; De Vivo, M.; Liu, Z. W.; Moore, P.; Klein, M. L. *Acc. Chem. Res.* **2006**, *39*, 73.

(21) Barnett, C. B.; Wilkinson, K. A.; Naidoo, K. J. *J. Am. Chem. Soc.* **2011**, *133*, 19474.

(22) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011.

(23) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1995**, *16*, 1339.

(24) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578.

(25) Patey, G. N.; Valleau, J. P. *J. Chem. Phys.* **1975**, *63*, 2334.

(26) Torrie, G. M.; Valleau, J. P. *J. Chem. Phys.* **1977**, *66*, 1402.

(27) Wang, F. G.; Landau, D. P. *Phys. Rev. Lett.* **2001**, *86*, 2050.

(28) Naidoo, K. J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9026.

(29) Amaya, M. F.; Watts, A. G.; Damager, I.; Wehenkel, A.; Nguyen, T. et al. *Structure* **2004**, *12*, 775.

(30) Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M. et al. *Acta Crystallogr. Sect. D* **2010**, *66*, 12.

(31) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, *61*, 704.

(32) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. *J. Chem. Theory Comput.* **2011**, *7*, 2284.

(33) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J. et al. *J. Comput. Chem.* **2009**, *30*, 1545.

(34) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. et al. *J. Comput. Chem.* **1983**, *4*, 187.

(35) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

(36) Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400.

(37) Guvench, O.; Mallajosyula, S. S.; Raman, E. P.; Hatcher, E.; Vanommeslaeghe, K. et al. *J. Chem. Theory Comput.* **2011**, *7*, 3162.

(38) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H. et al. *J. Chem. Phys.* **1995**, *103*, 8577.

(39) Ewald, P. P. *Ann. Phys.* **1921**, *369*, 253.

(40) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(41) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569.

(42) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861.

(43) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.

(44) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2010**, *114*, 17142.

(45) Amara, P.; Field, M. J.; Alhambra, C.; Gao, J. *Theor. Chem. Acc.* **2000**, *104*, 336.

(46) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books: Sausalito, California, 2005.

(47) Ciccotti, G.; Ferrario, M.; Laria, D.; Kapral, R. *Progress of Computational Physics of Matter: Methods, Software and Applications*; World Scientific Publishing: River Edge, NJ, 1995.

(48) Guthrie, R. D.; Jencks, W. P. *Acc. Chem. Res.* **1989**, *22*, 343.

(49) O'Ferrall, R. A. M. *J. Chem. Soc. B* **1970**, 274.

(50) Jencks, W. P. *Chem. Rev.* **1972**, *72*, 705.

(51) Damager, I.; Buchini, S.; Amaya, M. F.; Buschiazzo, A.; Alzari, P. et al. *Biochemistry* **2008**, *47*, 3507.

(52) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037.

(53) Davies, G. J.; Planas, A.; Rovira, C. *Acc. Chem. Res.* **2012**, *45*, 308.

(54) Whitfield, D. M. *Carbohydr. Res.* **2012**, *356*, 180.

6 Multi-dimensional Reaction Dynamics Reveal How the TcTS Enzyme Suppresses Competing Side Reactions and their Side Products[†]

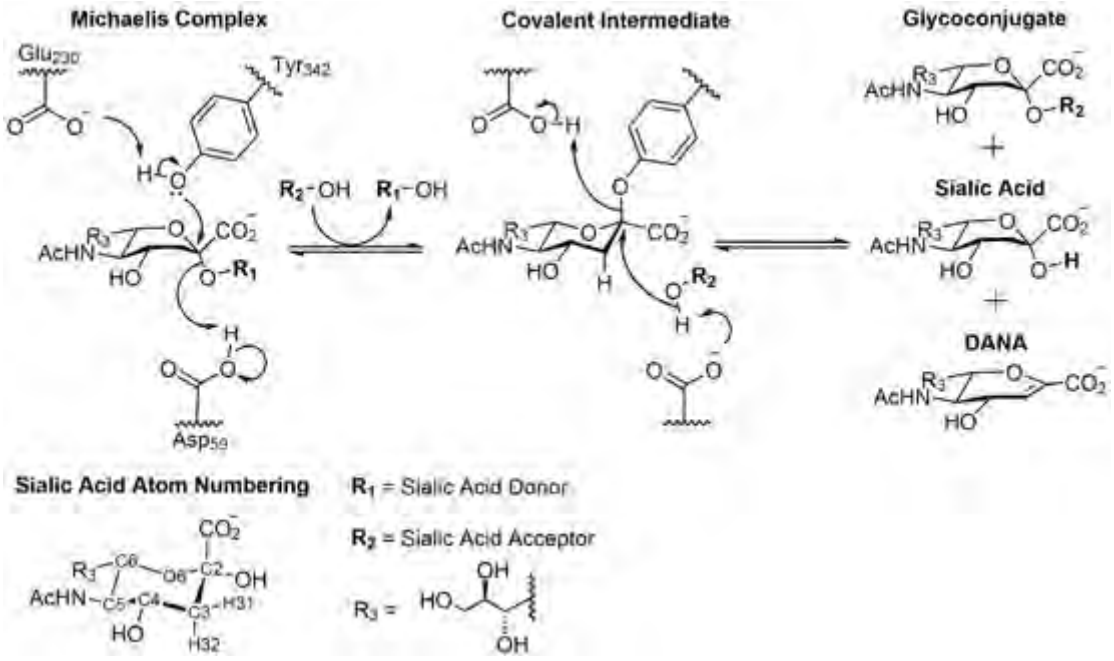
In Chapter 5, the TcTS deglycosylation FES was resolved as a function of the bond breaking and forming bonds, namely (SA C-2)…(Tyr_{342} O) and (Gal O-3)…(SA C-2). An ensemble of TS structures, defined in relation to the deconvoluted 2-D reaction coordinate, was sampled and profiled based on ring pucker and the conformations of the catalytic binding site. Here, the appropriate treatment of the reaction coordinate is addressed in characterizing the selectivity observed for TcTS activity. The dynamics of both steps of the overall TcTS-catalyzed transferase reaction are simulated, where glycosylation comprises the first step and deglycosylation the second step of the putative Ping-Pong mechanism.

The activation free energy is central to the phenomenological understanding of both rate acceleration and selectivity observed for catalyzed reactions compared with reference reactions in solution. In selectivity, kinetic considerations distinguish between pathways restricting the possible reaction paths that result in singular products. The TcTS-catalyzed reaction is susceptible to two side reactions, rooted in the native uncatalyzed reactions in solution (Scheme 6.1; refer to Section 1.5.2 for more detail). In the first side reaction, hydrolysis of the covalent intermediate by an active site water molecule competes with the primary transfer of sialic acid to an acceptor glycoconjugate.[1] The second side reaction is the elimination of an SA H-3 proton to DANA that has been detected when excess TcTS is incubated in the presence of sialyllactose.[2] More generally, this competition between elimination side reactions and displacement main reactions has been observed in other sialidases such as those expressed in *Influenza B* virus[3], *Vibrio cholera*[4] and *Streptococcus pneumonia*.[5] Elimination ostensibly proceeds either via H-3 abstraction by a mediating active site water molecule,[6] or by an appropriately located catalytic acid.[5] In contrast to the above enzymatic reactions, where the elimination reaction is observed to be in minor competition to the main displacement reaction, DANA formation readily reduces the success of sialic acid glycosylation in solution.[7] Here, the anomeric carboxylate group destabilizes the oxocarbenium ion through an inductive electron withdrawing effect to further yield a hindered electrophilic site presented by the bulky anion, and so promotes competing DANA formation via elimination. Therefore, the detection of DANA in only trace amounts for excess TcTS incubated with sialyllactose points to the ability of the *trans*-sialidase to mitigate the unfavorable side reaction.

Characterization of the elimination reaction at an atomistic level allows a unique opportunity to examine how enzymes diminish the success of competing reactions. The putative Ping-Pong

[†] The work presented in this chapter has been accepted for publication and will appear shortly in *ACS Catalysis*.

mechanism, in which Asp_{59} serves as the catalytic acid/base and TcTS is glycosylated to form the covalent intermediate, is presented in Scheme 6.1. As previously discussed in this thesis, experimental studies[1,2,8-10] have characterized substrate binding, identified enzymatic side products, determined the rate limiting step and calculated reaction rates for the overall TcTS-catalyzed reaction. Computational methods can be used to corroborate the proposed mechanism and are required for a complete characterization of both the thermodynamics and atomic motions governing the reactions under investigation. A detailed description of the reaction itinerary using only kinetic data is not possible. For instance, KIEs measured in a competitive fashion characterize only up to the first irreversible step[11] and NMR spectroscopy is limited to time averages and requires deconvolution of relevant peaks from the complex protein environment. Furthermore, deciphering the mechanistic details of reactions from these measurements is complicated by the relatively gross nature of the experimental spatial and temporal resolution. Similarly, analyses of X-Ray diffraction patterns are cumbersome involving the multistep synthesis of carbohydrate-related substrates that will inhibit the enzyme, and in so doing provide representative snapshots along the native reaction path.



Scheme 6.1 The putative catalytic mechanism for TcTS follows a catalytic Ping-Pong mechanism with the formation of a covalent intermediate. The primary transferase activity is susceptible to two side reactions: hydrolysis of the covalent intermediate to SA, and H-3 elimination to DANA. The atom numbering of the SA ring is given for reference (bottom left).

Computational models, for example MD simulations, have been used to show that TcTS mitigates the hydrolysis side reaction by excluding water from the active site[12] and only adopts a catalytically competent Tyr_{342} – Glu_{230} conformation in the presence of the sialic acid ligand.[13] QM/MM free energy calculations revealed that selectivity is further explained by the ~7 kcal/mol difference in energy barrier between these chemical steps.[13] Despite experimental observations of the DANA side product in some sialidases, there is an absence of a detailed molecular understanding of this competition. Here, the molecular details of TcTS' suppression of viable competing reactions is interrogated for physiological conditions to assert the precision and selectivity synonymous with enzyme catalysis. In the presence of a lactose acceptor, the β-galactose will be hydrogen bonded to Asp_{59}, and abstraction of the bottom-facing SA H-3 in the deglycosylation step will be by the acceptor molecule's O-3 oxygen, analogous to the water-mediated mechanism recorded by Jongkees and Withers[6] (refer to Scheme 1.3 in Chapter 1).

Multi-dimensional reaction dynamics are used to map out the TcTS-catalyzed Ping-Pong reaction as well as the competing reactions leading to DANA formation. While it is common to plot the free energy as a function of a (conflated) 1-D or 2-D reaction coordinate, the simplified picture derived from such models results in the loss of physically and chemically important information.[14] To eliminate these shortcomings that prevent the simultaneous monitoring of multiple and diverging reaction paths possible in an enzyme, 4-D FEVs are used to explore the TcTS-catalyzed *trans*-sialidase activity on sialyllactose.

6.1 Computational Methods

6.1.1 Protein Preparation

The Michaelis complex crystal structure was prepared and equilibrated first using an MM force field under periodic boundary conditions and then SCCDFTB/MM under stochastic boundary conditions as reported in Chapter 5. The equilibrated TcTS systems derived from the experimental Michaelis complex crystal structure were then used to commence computational reaction dynamics simulations.

6.1.2 FEVs

TcTS' catalytic itinerary was explored using free energies calculated from probability histograms along an appropriate reaction coordinate. In this work, each breaking or forming bond is mapped to an element, ξ_α, of the reaction coordinate, where a 2-D (ξ_1, ξ_2) results in a FES while a 3-D (ξ_1, ξ_2, ξ_3) gives a FEV. The FEARCF method was used to apply forces to atoms comprising ξ in order to drive the system to equally visit each state of the reaction coordinate space. The reaction coordinate-space for the first and second steps of TcTS-catalyzed hydrolysis is defined by three bonds: SA C-2 – Nucleophile,

SA H-3 – Base and SA C-2 – Leaving Group. The 3-D reaction coordinates are comprised of discretized bins that are populated by points taken from multiple reaction dynamics trajectories run in parallel. The addition of H-3 abstraction among the geometric variables allows the reaction free energy to be followed along the elimination path, and independent assessment of the reactions leading to product (covalent intermediate or sialyllactose) and side product (DANA) in each step. Protonation of the glycosidic linkage and proton transfer between Tyr_{342} and Glu_{230} is possible using only thermal energy and so requires no biasing force to drive this reaction coordinate. Consequently, these bonds were excluded from the reaction coordinate definitions. The PMF dimensions are 0.5-6.5 Å × 0.5-6.5 Å × 0.5-6.5 Å and all elements of the 3-D ξs were partitioned into discrete bins comprising a spacing of 0.1 Å.

The individual bond and angle sampling statistics describing the postures adopted by reacting substrates are conflated within the 3-D reaction coordinates that compose the FEV. Unpacking the mechanistic details required the deconvolution of the population distributions of selected internal coordinates from the equilibrium distributions that are derived from representative reactive trajectories. Consequently, the geometric relationship between Tyr_{342} or galactose and sialic acid was analyzed from the data taken from a 500 ps SCC-DFTB/CHARMM equilibrium trajectory.

The generalized TS was located from the maximum energy regions on the FEV as the point that has the smallest numerical gradient. The minimum pathway was then computed by following the steepest gradient down to the stationary points. TS volumes were calculated by a Voronoi tessellation of free energy contours about the TS (see Appendix C for details).

6.1.3 Reaction Trajectory Analysis

Representative crossings were selected from the first crossing trajectories and then compared with the minimum free energy path gained from the gradient of the FEV. Discrete Frèchet distances were calculated between the trajectories and the reaction path. The trajectory with the smallest distance was selected whereupon sialic acid ring puckers were calculated for each step using in-house code.

In addition, TS structures were extracted from the representative FEARCF trajectories and optimized using SCC-DFTB/CHARMM and M06-2x/6-31G(d)/CHARMM. The CHARMM MD engine does not implement robust local optimization methods for TS optimizations. Therefore, while the semi-empirical TS structures were optimized using the POLYRATE module,[15] the system was transferred to NWChem 6.5[16] in order to optimize the snapshots onto a DFT/CHARMM potential energy surface. Different implementations of the CHARMM force field in the two MD engines necessitated conversion of the CHARMM topology and parameters to the NWChem format. The code written to automate this conversion will be made available in the near future at the Scientific Computing Research Unit repository (https://bitbucket.org/scientificomputing). Calculations were run by optimizing structures

taken from a dynamic snapshot while restraining the atoms comprising the reaction coordinate. Thereafter the single negative frequency was followed to the TS structure (see Appendix C for details). Normal modes were computed for the converged structures to confirm the existence of a single negative frequency corresponding to reaction progress.

A natural bond orbital analysis (see Appendix C for details) was run on the DFT structures, using the NBO program as implemented in Gaussian 09[17,18], including all background charges. Bond orders reported are the Wiberg bond index. Primary ^{13}C and β-dideuterium KIEs were also calculated for Displacement 1A (Scheme 6.2) for comparison against experimentally calculated values. Ratios were calculated from the numerical Hessians of the optimized TS and Michaelis complex structures using the Bigeleisen-Mayer equation.[19]

6.2 Results and Discussion

The FES describing covalent intermediate formation resolved along the ξ_1 = (Tyr_{342} O)…(SA C-2) attack and ξ_2 = (SA C-2)…(Gal O-3) leaving distances, displayed a local free energy minimum located at a long glycosidic C – O bond length. The sampling of the oxocarbenium ion species is expected at this location of the FES as the result of a higher energy dissociative reaction path. Surprisingly, some structures from trajectories visiting this region showed the formation of DANA, in so providing evidence for the presence of a competing side reaction. Supporting this observation is the previous detection of trace amounts of DANA in kinetic studies on TcTS, albeit under hydrolyzing conditions. Although DANA is a weak inhibitor of TcTS,[20] further evidence demonstrating that the *trans*-sialidase is capable of dehydrogenating sialic acid is provided by a TcTS:DANA crystal structure that corresponds to monoclinic crystals soaked with sialic acid.[21]

To uncover the TcTS catalytic action that results in an exclusive transfer of sialic acid from host donors, it is best to monitor the elimination and displacement reactions simultaneously in both steps. Consequently, two FEVs that define the free energy as a function of three bonds were computed. The FEVs reveal the mechanistic drivers of the elimination reactions that are competing with the dominant displacement reactions within a catalytic itinerary under physiological (non-hydrolyzing) conditions. The structural and thermodynamic results are summarized in Table 6.1.

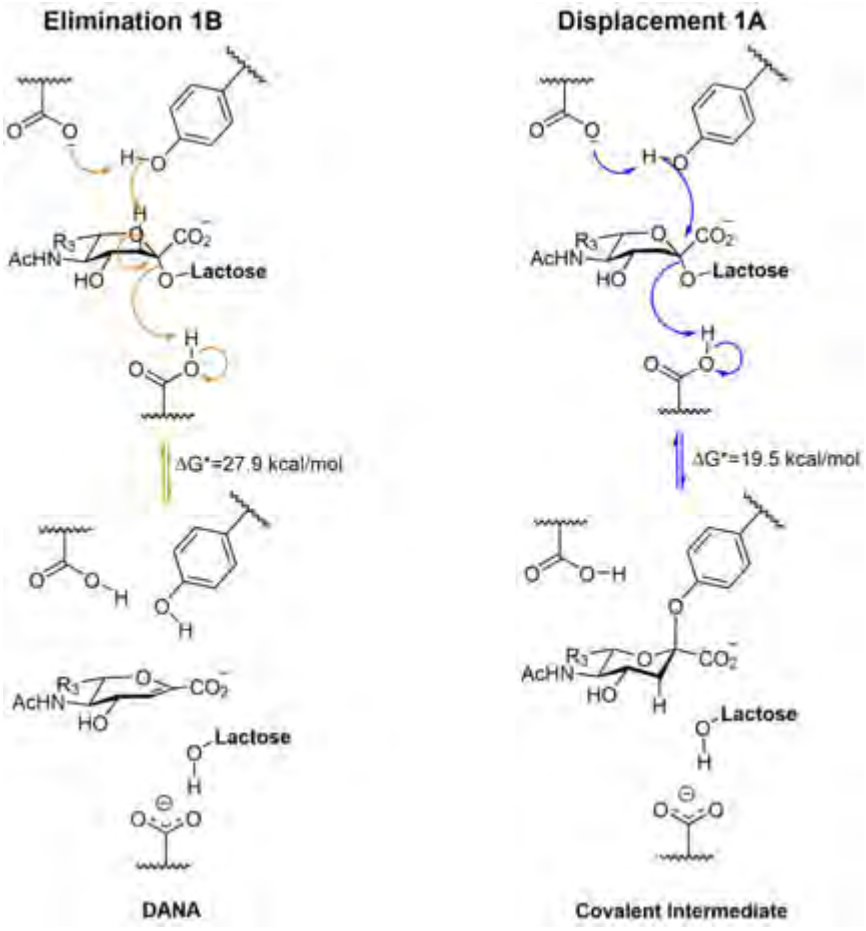
Table 6.1 Summary of structural and thermodynamic results.

Reaction	$\Delta G^{\ddagger}$ (kcal/mol)	FEV TS (Å)[a,b]	FEV TS Volume (Å^3)	PE TS Structure[a,b]	
				SCC-DFTB (Å)	M06-2x (Å)
Displacement 1A: Michaelis complex → covalent intermediate	19.5	(2.2,2.3)	0.021	(2.18, 2.22)	(2.57,2.10)
Elimination 1B: Michaelis complex → DANA	27.9	(1.3, 2.1)	0.001	(1.26, 2.15)	(1.34, 2.44)
Displacement 2A: covalent intermediate → Michaelis complex	20.6	(2.1,2.4)	0.024	(2.27, 2.18)	(2.08, 2.57)
Elimination 2B: covalent intermediate → DANA	27.4	(1.3,2.3)	0.006	(1.30, 2.12)	(1.31, 2.34)

a Parameters for SCC-DFTB were the MIO set, and a 6-31G(d) basis set was used for DFT calculations.
b Reaction coordinates are (SA – Nucleophile, SA – Leaving Group) for substitution reactions and (SA – Base, SA – Leaving Group) for elimination reactions.

6.2.1 Step 1: Glycosylation

The competing reactions explored for the glycosylation step, Displacement 1A and Elimination 1B, are shown in Scheme 6.2.



Scheme 6.2 The primary reaction, Displacement 1A, that leads to TcTS glycosylation is shown alongside the competing Elimination 1B reaction to DANA.

Covalent intermediate formation is characterized by an isosurface of 19.5 kcal/mol that is carved out of the FEV resolved for ξ_1 = (Tyr_{342} O)…(SA C-2), ξ_2 = (Tyr_{342} O)…(SA H-31) and ξ_3= (SA C-2)…(Gal O-3) (Figure 6.1). The isosurface extends from the Michaelis complex and covalent intermediate regions to connect at ξ_1 = 2.2 Å and ξ_3 = 2.3 Å. This barrier compares directly with umbrella sampling results ($\Delta G^{\ddagger}$ = 20.8 kcal/mol) where average values of 2.52 Å for ξ_1 and 2.19 Å for ξ_3 can be computed from the TS bin of the conflated reaction coordinates.[22] In comparison, an isosurface of 27.9 kcal/mol connects the Michaelis complex and DANA stationary points at ξ_2 = 1.3 Å and ξ_3 = 2.1 Å (Figure 6.1). In both the displacement and elimination reactions, dividing TS volumes separate the reactant and product stationary points suggestive of bimolecular mechanisms. The difference of 8.4 kcal/mol in activation energies shows that covalent intermediate formation via Displacement 1A is kinetically more favored.

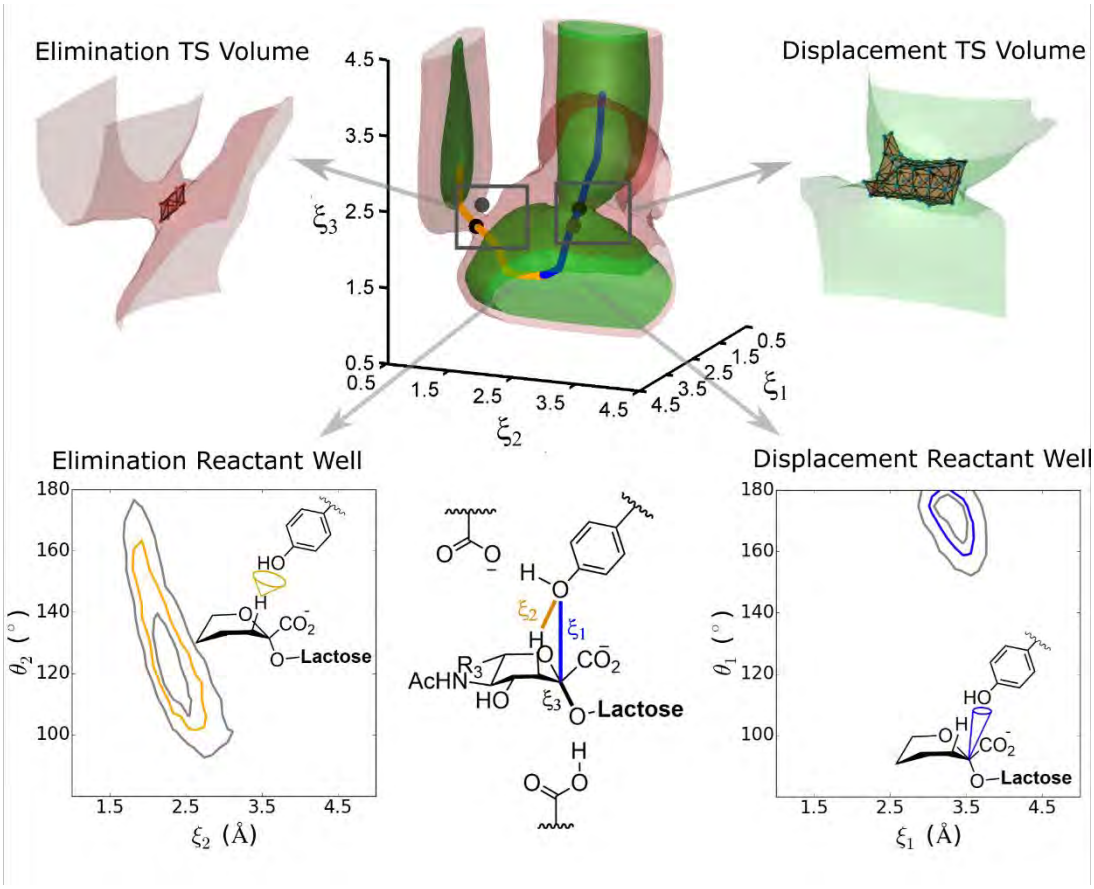


Figure 6.1 FEV for glycosylation step with ξ_1 = (Tyr$_{342}$ O)…(SA C-2), ξ_2 = (Tyr$_{342}$ O)…(SA H-31) and ξ_3= (SA C-2)…(Gal O-3) illustrated in the molecular drawing. Isosurfaces are shown at 21.0 kcal/mol (green) and 29.4 kcal/mol (red), and corresponding Displacement 1A (blue) and Elimination 1B (yellow) free energy paths are traced. SCC-DFTB generalized TSs are indicated by black dots, while the M06-2x/6-31G(d) PE TS structures are shown in gray. TS volumes are shown flanking the FEV. Probability contour plots for the geometrical analysis of the 500 ps Michaelis complex equilibrium trajectory are shown below at confidence intervals of 30%.

The free energy paths include entropic contributions that can be qualitatively analyzed by comparing the free energy landscape local to the TS and reactant. Isoenergetic surfaces connecting points 1.5 kcal/mol greater than the respective free energies of activation depict the entry channels to the TS (Figure 6.1). The volume encompassed by the surfaces are defined by the entropic degrees of freedom orthogonal to the reaction coordinate.[23] Relative narrowing of the channels indicates an unfavorable entropy change and diminished flux of activated complexes for both reactions. However, the TS volume for the primary displacement reaction is some 20 times larger than the elimination reaction in so registering a much faster rate of reaction for covalent intermediate formation (Table 6.1).

An NAC-type analysis was performed to probe the mechanistic detail underlying the free energy results. The geometric relationship between Tyr$_{342}$ and sialic acid was analyzed from the data taken from a 500 ps SCC-DFTB/CHARMM equilibrium trajectory. For the competing reactions, the probability contours were constructed as a function of the distance of the forming bond and a proxy for the angle

of attack (Figure 6.1). In the case of covalent intermediate formation this angle is θ_1 = (Tyr_{342} O)…(SA C-2)…(Gal O-3) and in the case of DANA formation it is θ_2 = (Tyr_{342} O)…(SA H-31)…(SA C-3). The ξ_1 bond length and θ_1 angle in the Michaelis complex remain tightly clustered over the 500 ps, within van der Waals contact distance and occupying a linear angle of nucleophilic attack with favorable stereoelectronic alignment. Conversely, the ξ_2 bond and θ_2 angle show a more diffuse distribution.

6.2.1.1 Mechanistic Details of Displacement 1A

The FEV is built from the total of all reaction dynamics trajectories making up each of the FEARCF iterations. The reactant well evolves over FEARCF iterations to the product well, which allows electronic and conformational information to be extracted from the reaction trajectories. The crossing trajectory that most closely follows the Displacement 1A free energy path shows near-synchronous lengthening of the glycosidic bond and shortening of the (Tyr_{342} O)…(SA C-2) distance on a timescale shorter than 1 ps (Figure 6.2A). This profile and the location of the free energy TS at ξ_1 = 2.2 Å and ξ_3 = 2.3 Å is consistent with a mechanism positioned between A_ND_N and the more dissociative $D_N{*}A_N$.

The M06-2x/6-31G(d) TS structure was gained from optimizing a dynamic snapshot selected from the reactive trajectory (ξ_1=2.52 Å, ξ_3=2.03 Å; Table 6.1). The primary ^{13}C KIE associated with this structure was 1.025 which compares favorably with the experimental value of 1.021 ± 0.014.[24] On the other hand, the secondary β-dideuterium KIE was calculated at 1.107 and shows some variation from the experimental value of 1.053 ± 0.010. The DFT refinement implies an earlier TS with an estimated bond order of 0.21 for the breaking glycosidic bond. Nucleophilic participation is seen through the donation of electron density from the Tyr_{342} OH lone pair into the (SA C-2)…(SA O-6) antibonding natural bond orbital (NBO) that has a second-order interaction energy of 5.4 kcal/mol. An oxocarbenium ion-like TS is evident from the anomeric carbon that has a valence electron configuration approximating an idealized sp^2 hybridization, $2s^{0.94}2p^{2.39}$. Further supporting evidence for an oxocarbenium ion comes from the increased (SA C-2)…(SA O-6) bond order (1.27 cf. reactant 0.95) and reduced charge on O-6 (-0.47 cf. reactant -0.59). The ring conformation starts at the Equator ($B_{2,5}$ and $^4S_2$) of the Crèmer–Pople sphere and travels longitudinally through E_5 and $^4H_5$ (TS) to $^2C_5$ (the covalent intermediate pucker conformer) at the pole. To achieve a flattened TS, the SA C-5 is held below the plane of the pyranose ring by a hydrogen bond between its acetylamido group and Asp_{96}, and C-2 is drawn down onto the plane.

6.2.1.2 Mechanistic Details of Elimination 1B

The representative Elimination 1B crossing trajectory is less synchronous and shows a greater degree of fluctuation in the forming bond between Tyr_{342} OH base and SA H-31 (Figure 6.2B). The glycosidic bond stretches to 2.0 Å early in the trajectory with the sialic acid ring adopting a flattened E_5 pucker

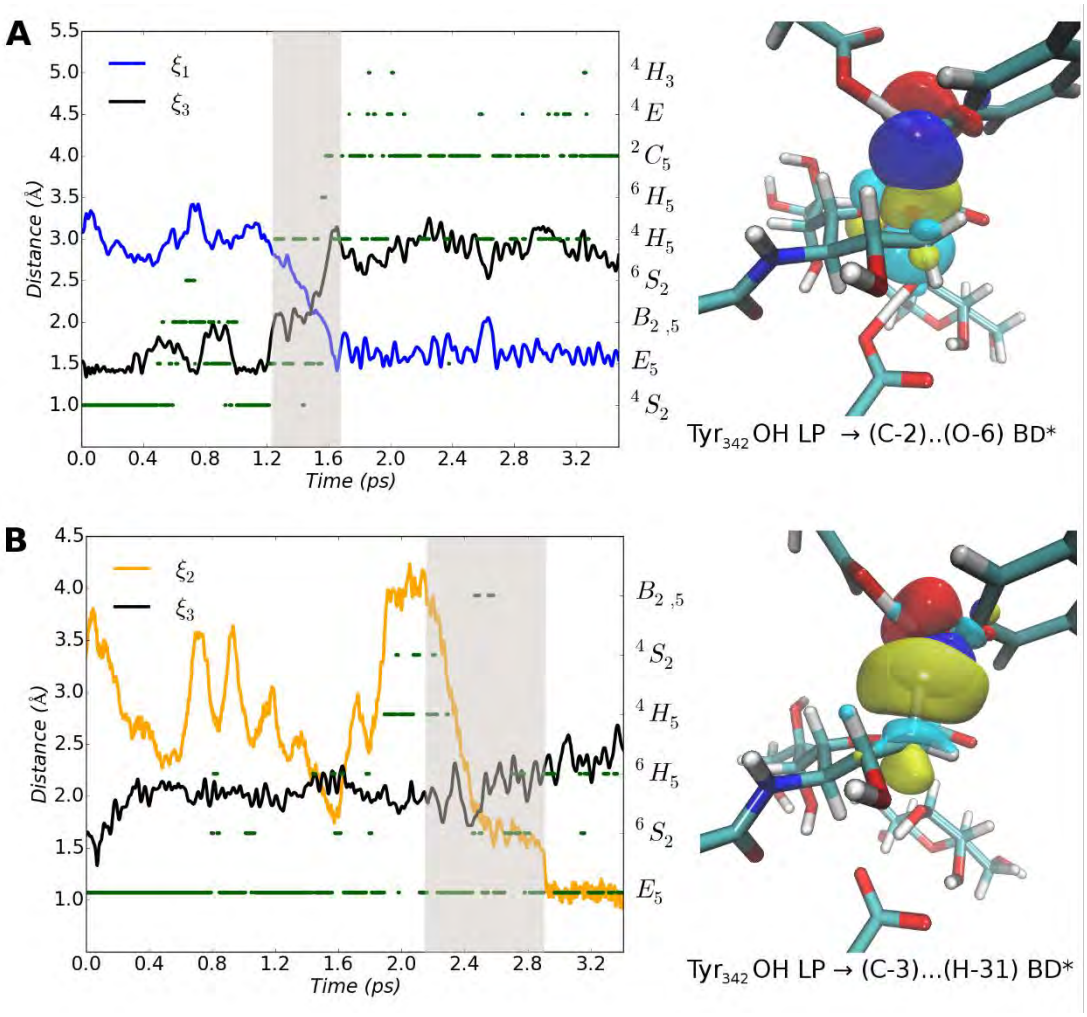


Figure 6.2 Mechanistic analysis of representative (A) Displacement 1A and (B) Elimination 1B crossings. Geometric analysis of the breaking and forming bonds comprising ξ is plotted on the left along with pucker. The forming bond is colored to match the ξ definition in Figure 6.1, and grayed sections illustrate the TS region from which the TS structure was optimized. NBOs for this structure are shown on the right at an electron density of -0.02 and 0.02.

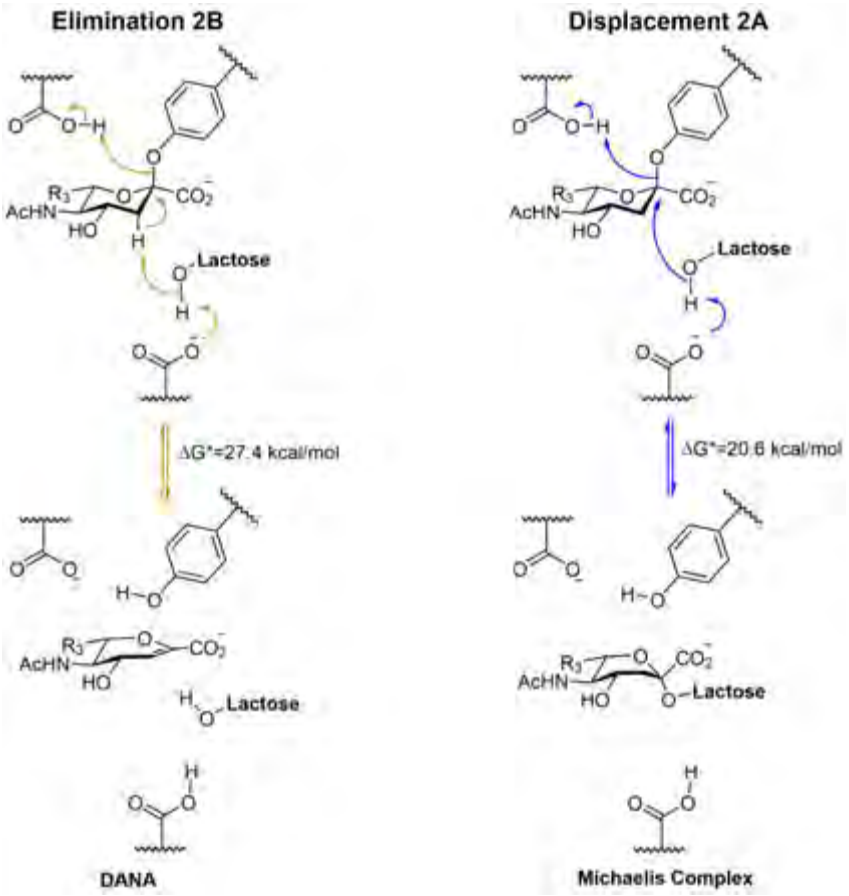
due to dissociation of the leaving group. In the absence of nucleophilic attack, Tyr_{342} OH forms contact with SA H-31. There is significant fluctuation in this contact distance with a variance of 0.18 Å over the first 1 ps of the trajectory, compared to 0.03 Å for (Tyr_{342} O)…(SA C-2) in the displacement reactive trajectory. The positional instability of the proton hinders the system from meeting the selective stereoelectronic requirement for proton abstraction. The effect of the proton's positional instability on the failure of the elimination reaction to progress is quantifiably evident from an examination of the diffuse equilibrium distribution of proton abstraction distance and angle (Figure 6.1). The probability of Tyr_{342} OH successfully approaching SA H-31, to form an NAC, is far less than Tyr_{342} OH approaching the SA C-2. However, when the former does happen, and Tyr_{342} OH comes within 2 Å of SA H-31, the pair fluctuates about this configuration for a further ~0.4 ps before abstracting the proton, thereby implying a dissociative $D_N{*}A_{xh}D_H$ mechanism.

On closer inspection, when the tyrosine oxygen approaches SA H-31, the ring puckers through a $^{6}S_{2}$ conformation displaying a pseudo-axial glycosidic bond. This approximate anti-periplanar geometry of the glycosidic C – O bond relative to H-31 is favorable for an $A_{xh}D_{H}D_{N}$ elimination reaction. An additional impediment to the successful completion of the elimination reaction is the narrowing of the free energy channel into a well-defined TS volume (Figure 6.1) therefore supporting a bimolecular reaction mechanism. The combination of the restricted TS volume, the positional instability of the SA H-31 and the associated ring puckering suggests that the mechanism derived from the SCC-DFTB level of theory is somewhere between $D_{N}{*}A_{xh}D_{H}$ and $A_{xh}D_{H}D_{N}$.

The DFT TS structure (ξ_2 = 1.34 Å, ξ_3 = 2.44 Å) supports a $D_{N}{*}A_{xh}D_{H}$ mechanism with an estimated bond order of only 0.08 for ξ_3 (cf. 0.23 for Displacement 1A TS structure). In addition, animation of the single negative frequency shows major contribution from the proton transfer, but little movement along the glycosidic bond stretching vibration. The electron deficiency on the anomeric carbon is quenched by donation of electron density into the (SA C-2)…(SA O-6) bond, giving the structure some oxocarbenium character (O-6 charge -0.48, (SA C-2)…(SA O-6) bond order 0.97), as well as donation of electron density into the (SA C-3)…(SA C-2) bond. The bond order for the latter increases to 1.13 compared to 0.98 in the displacement TS structure. The DFT TS structure reveals that the Tyr_{342} OH lone pair donates into the C-3 – H-31 antibonding orbital with an interaction energy of 89.2 kcal/mol. The protonation of both Glu_{230} and Tyr_{342} residues leads to a new interaction mode in which Tyr_{342} forms a hydrogen bond with the SA C-2 carboxylate group.

6.2.2 Step 2: Deglycosylation

A second possible source of DANA formation in the TcTS-catalyzed transferase reaction is the deglycosylation step (Scheme 6.3).



Scheme 6.3 The primary reaction, Displacement 2A, that leads to TcTS deglycosylation is shown alongside the competing Elimination 2B reaction to DANA.

The FEV (Figure 6.3) for the proposed competing reactions of the deglycosylation step was resolved as a function of ξ_1 = (Gal O-3)…(SA C-2), ξ_2 = (Gal O-3)…(SA H-32) and ξ_3 = (SA C-2)…(Tyr_{342} O). The results show that the primary displacement reaction is favored by both barrier height and larger TS volume. The displacement TS is located at ξ_1 = 2.1 Å and ξ_3 = 2.4 Å with an associated free energy of 20.6 kcal/mol. There is a degree of asymmetry in the FEARCF free energy barrier height and TS location for the forward and reverse displacement reactions. This may arise from the greater degree of translational freedom of galactose O-3 compared to the Tyr_{342} OH nucleophile of the reverse reaction (Displacement 1A). The elimination reaction crosses a TS barrier that is 6.8 kcal/mol higher. This isoenergy contour extends along ξ_3 through a bottleneck TS volume that is smaller than the displacement reaction TS volume by a factor of 4 (Table 6.1). The TS volume (top left frame in Figure 6.3) is flat in appearance and implies that elimination proceeds at a range of leaving group distances,

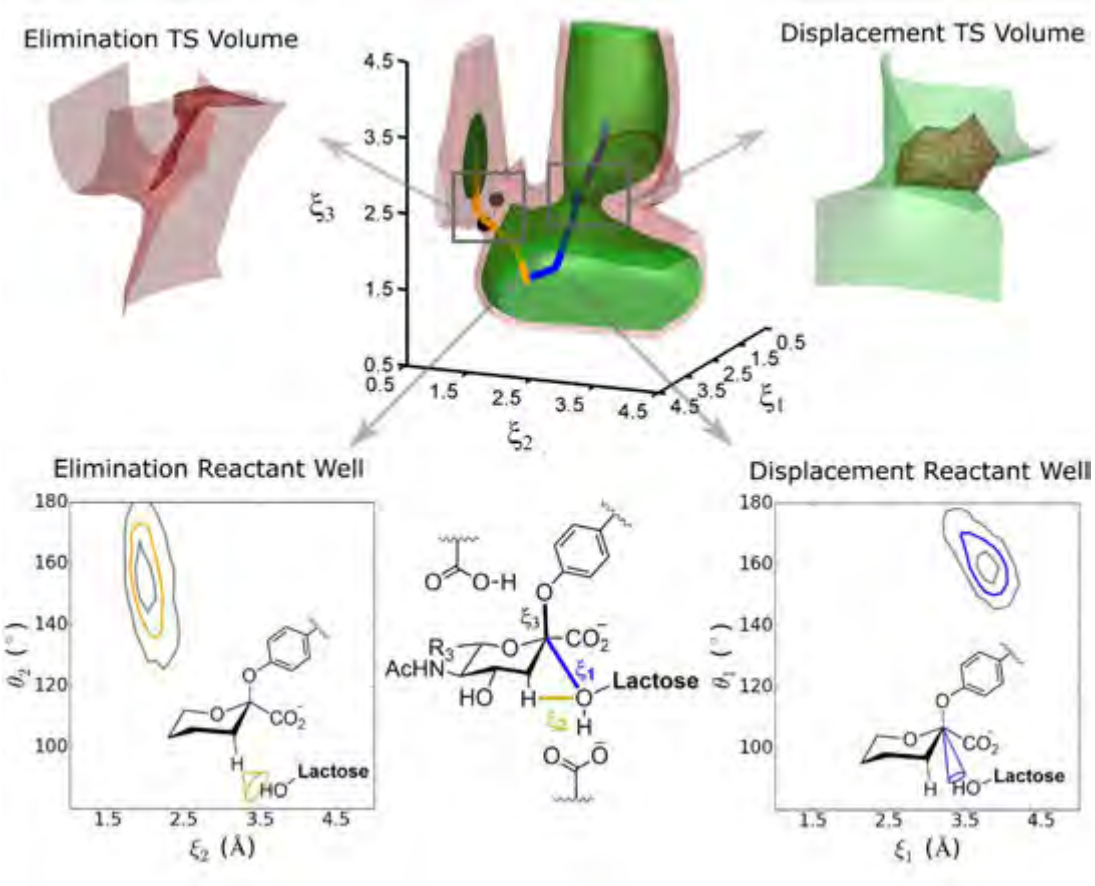


Figure 6.3 FEV for deglycosylation step with ξ_1 = (Gal O-3)…(SA C-2), ξ_2 = (Gal O-3)…(SA H-32) and ξ_3 = (SA C-2)…(Tyr$_{342}$ O) illustrated in the molecular drawing. Isosurfaces are shown at 22.1 kcal/mol (green) and 28.9 kcal/mol (red), and corresponding Displacement 2A (blue) and Elimination 2B (yellow) free energy paths are traced. SCC-DFTB generalized TSs are indicated by black dots, while the M06-2x/6-31G(d) PE TS structures are shown in gray. TS volumes are shown flanking the FEV. Probability contour plots for the geometrical analysis of the 500 ps Michaelis complex equilibrium trajectory are shown below at confidence intervals of 30%.

although the FEV grid point with the smallest gradient was found at ξ_2 = 1.3 Å and ξ_3 = 2.3 Å. While the Michaelis complex equilibrium sampling about the Displacement 1A NAC is significantly more concentrated than that about the Elimination 1B NAC, this difference in sampling area is not as pronounced in the second step. In Displacement 2A, the nucleophilic attack distance, ξ_1, shifts to longer bond lengths, with the probability contours centered just outside the van der Waals contact distance for carbon and oxygen, 3.77 Å. On the other hand, the postures adopted by the reacting substrates become more favorable for elimination. The ξ_2 distribution narrows and the θ_2 = (Gal O-3)…(SA H-32)…(SA C-3) angle becomes more linear, which enhances stereoelectronic alignment of the lone pair on galactose O-3 and the (SA C-3)…(SA H-32) antibonding orbital.

6.2.2.1 Mechanistic Details of Displacement 2A

Representative crossing trajectories for the deglycosylation step show that the dominant Displacement 2A reaction is more associative (Figure 6.4A) than competing Elimination 2B (Figure 6.4B). In the former reaction, the C – O bond to Tyr$_{342}$ fluctuates between 1.5 and 2.0 Å with the sialic

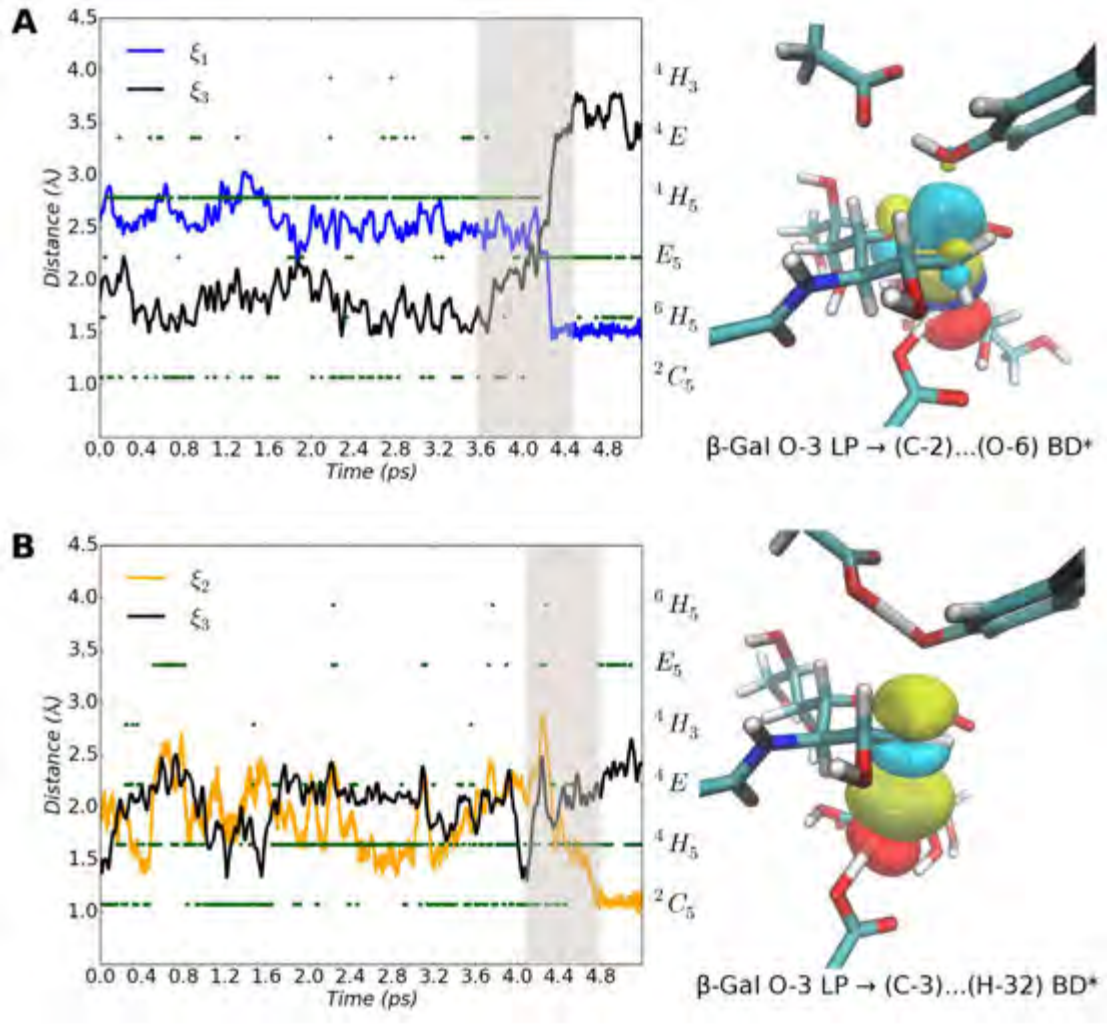


Figure 6.4 Mechanistic analysis of representative (A) Displacement 2A and (B) Elimination 2B crossings. Geometric analysis of the breaking and forming bonds comprising ξ is plotted on the left along with pucker. The forming bond is colored to match the ξ definition in Figure 6.3, and grayed sections illustrate the TS region from which the TS structure was optimized. The NBOs for this structure are shown on the right at an electron density of -0.02 and 0.02.

acid ring moving from $^{2}C_{5}$ to adopt a $^{4}H_{5}$ pucker. The SA – Tyr_{342} bond then starts breaking at 3.6 to 4.0 ps triggering nucleophilic attack by the galactose O-3 nucleophile and passage through a planar TS (E_5 pucker). As in Chapter 5, this profile and the location of the free energy TS at ξ_1 = 2.1 Å and ξ_3 = 2.4 Å is in line with a dissociative A_ND_N mechanism. Analyzing the electronic nature of the DFT TS structure, ξ_1 = 2.08 Å and ξ_3 = 2.57 Å, reveals that the galactose O-3 lone pair donates into the (SA C-2)…(SA O-6) antibonding NBO, localized on C-2, with a stabilizing energy of 50.0 kcal/mol. Since the TS structure is the same as Displacement 1A, but approached from the opposite end, the oxocarbenium chemical character is the same as described above, i.e. SA C-2 electron configuration $2s^{0.95}2p^{2.39}$, SA O-6 charge -0.47 and the (SA C-2)…(SA O-6) estimated bond order 1.27.

6.2.2.2 Mechanistic Details of Elimination 2B

On the other hand, the elimination reaction comprises a dissociating SA – Tyr_{342} bond fluctuating around 2.0 Å for significant periods while the sialic acid ring adopts a $^4H_5$ pucker. Here the lactose base approaches 1.5 Å from SA H-32 but does not immediately abstract the proton. The (Gal O-3)…(SA H-32) bond shows a variance of 0.12 Å over the first 1 ps of the shown trajectory (cf. 0.04 Å for (Gal O-3)…(SA C-2) in the representative Displacement 2A trajectory). Elimination eventually proceeds through $^4H_5$ that has a favorable anti-periplanar geometry.

The M06-2x/6-31G(d) TS structure (ξ_2 = 1.31 Å, ξ_3 = 2.34 Å), shows overlap of the galactose O-3 lone pair and the (SA C-3)…(SA H-32) antibonding NBO with an interaction energy of 105.3 kcal/mol. Electron deficiency on C-2 is quenched by increased density in the (SA C-2)…(SA O-6) and (SA C-2)…(SA C-3) bonds giving a chemical profile for the sialic acid ring very similar to the elimination TS in the first step: (SA C-2)…(SA C-3) estimated bond order 1.30, SA O-6 charge -0.47 and (SA C-2)…(SA O-6) estimated bond order 1.20. The generalized TS, the DFT TS structure, and the positional instability of the SA H-32 all point to a $D_N{*}A_{xh}D_H$ mechanism.

6.3 Concluding Remarks

Multi-dimensional computations that included the distance between the nucleophile (of the primary displacement reaction) and SA C-3 hydrogen in the ξ set for both steps have allowed simultaneous monitoring of the displacement and elimination pathways. This approach has the advantage of directly comparing the fundamental thermodynamics defining the elimination side reactions producing DANA with that of the displacement reactions. While the elimination reaction paths require crossing comparatively higher energy barriers ($\Delta\Delta G^{\ddagger}$ ~7 kcal/mol), the importance of controlling biased conformational and configurational sampling is a key observation that assists in our understanding of enzyme catalytic mechanisms. Deconvolution of the statistics underpinning the free energy for the equilibrium states provide the spatial-temporal details of the TcTS catalyzed transformation. In both the glycosylation and deglycosylation steps the probability of the nucleophile (alternatively Tyr_{342} OH or galactose O-3) successfully approaching SA C-2 to form an NAC, through which the ground state must pass to become the TS, is significantly greater than the same group approaching the SA C-3 hydrogen to form the elimination NAC. The nucleophile is held in a position with favorable stereoelectronic alignment for attack on the nucleophilic site, while the positional instability of the SA C-3 proton hinders the system from meeting the selective stereoelectronic requirement for proton abstraction.

6.4 References

(1) Scudder, P.; Doom, J. P.; Chuenkova, M.; Manger, I. D.; Pereira, M. E. A. *J. Biol. Chem.* **1993**, *268*, 9886.

(2) Todeschini, A. R.; Mendonça-Previato, L.; Previato, J. O.; Varki, A.; van Halbeek, H. *Glycobiology* **2000**, *10*, 213.

(3) Burmeister, W. P.; Henrissat, B.; Bosso, C.; Cusack, S.; Ruigrok, R. W. H. *Structure* **1993**, *1*, 19.

(4) Moustafa, I.; Connaris, H.; Taylor, M.; Zaitsev, V.; Wilson, J. C. et al. *J. Biol. Chem.* **2004**, *279*, 40819.

(5) Xu, G.; Kiefel, M. J.; Wilson, J. C.; Andrew, P. W.; Oggioni, M. R. et al. *J. Am. Chem. Soc.* **2011**, *133*, 1718.

(6) Jongkees, S. A. K.; Withers, S. G. *Acc. Chem. Res.* **2014**, *47*, 226.

(7) Ress, D.; Linhardt, R. *Curr. Org. Synth.* **2004**, *1*, 31.

(8) Amaya, M. F.; Watts, A. G.; Damager, I.; Wehenkel, A.; Nguyen, T. et al. *Structure* **2004**, *12*, 775.

(9) Yang, J. S.; Schenkman, S.; Horenstein, B. A. *Biochemistry* **2000**, *39*, 5902.

(10) Damager, I.; Buchini, S.; Amaya, M. F.; Buschiazzo, A.; Alzari, P. et al. *Biochemistry* **2008**, *47*, 3507.

(11) Bennet, A. J. *Curr. Opin. Chem. Biol.* **2012**, *16*, 472.

(12) Demir, O.; Roitberg, A. E. *Biochemistry* **2009**, *48*, 3398.

(13) Bueren-Calabuig, J. A.; Pierdominici-Sottile, G.; Roitberg, A. E. *J. Phys. Chem. B* **2014**, *118*, 5807.

(14) Naidoo, K. J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9026.

(15) Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; J. Villa; Hu, W.-P. et al. In *POLYRATE*; Version 9.0 ed.; University of Minnesota: Minneapolis.

(16) Aprà, E.; Bylaska, E. J.; Dean, D. J.; Fortunelli, A.; Gao, F. et al. *Comput. Mater. Sci.* **2003**, *28*, 209.

(17) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. In *NBO*; Version 3.1.

(18) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A. et al.; Gaussian, Inc.: Wallingford, CT, USA, 2009.

(19) Bigeleisen, J.; Mayer, M. G. *J. Chem. Phys.* **1947**, *15*, 261.

(20) Freire-de-Lima, L.; Oliveira, I. A.; Neves, J. L.; Penha, L. L.; Alisson-Silva, F. et al. *Front. Immunol.* **2012**, *3*, 356.

(21) Buschiazzo, A.; Amaya, M. F.; Cremona, M. L.; Frasch, A. C.; Alzari, P. M. *Mol. Cell* **2002**, *10*, 757.

(22) Oliveira, I. A.; Gonçalves, A. S.; Neves, J. L.; von Itzstein, M.; Todeschini, A. R. *J. Biol. Chem.* **2014**, *289*, 423.

(23) Anslyn, E. V.; Dougherty, D. A. *Modern Physical Organic Chemistry*; University Science Books: Sausalito, California, 2005.

(24) Yang, J.; Schenkman, S.; Horenstein, B. *Biochemistry* **2000**, *39*, 5902.

7 Evaluation of the Electronic Treatment of the TcTS QM Region

Reaction free energy profiles were calculated for the TcTS system in Chapters 5 and 6. In addition to the appropriate definition of the reaction coordinate and an adequate level of sampling, reliable free energies require an accurate treatment of the electronic and conformational degrees of freedom in the reaction space. Therefore, tractable computations need to balance computational efficiency and accuracy. Accordingly, the reaction coordinate is routinely conflated, and the electronic degrees of freedom are confined to a subset of atoms constituting the QM region. Here, the electronic treatment of the QM region is investigated for QM/MM reaction dynamics of TcTS.

Enzyme QM regions are often treated with a semi-empirical model to take advantage of the speed offered by these methods. The derivation of the approximate SCC-DFTB method from the Kohn-Sham DFT energy functional results in a scheme with few parameters that are obtained directly through computation (see Chapter 3). As a consequence, the parameters are largely transferable and SCC-DFTB has been successfully applied to the study of enzyme reactions, including those that chemically transform carbohydrate substrates.[1-4] The speed of SCC-DFTB, comparable to that of traditional semi-empirical methods such as the MNDO, AM1, and PM3 schemes, makes SCC-DFTB an attractive option to treat the electrons of the reacting atoms. Since pucker plays an important role in reactions that act on glycosidic bonds, it is noteworthy that SCC-DFTB has been shown to model pucker behavior more accurately than the aforementioned semi-empirical wavefunction methods.[5]

To quantify the effects of treating the electronic degrees of freedom in reaction space using a semi-empirical model, it is necessary to compare the potential energy function against a higher level of theory. This is due to the difficulties inherent in validating computational models against experiment. TST may be invoked to calculate a theoretical reaction rate for comparison against experimental values. However, cancellation of errors in an inaccurate model may yield the correct free energy barrier from an incorrect free energy path.[6] Path integral sampling can be used to 'correct' PMFs resolved using classical Newtonian mechanics and so generate comparative KIEs. Unfortunately, path integral methods are best suited to enzyme systems where only a few degrees of freedom involved in the chemistry need to be quantized.[7] A more thorough assessment of the computational model would include validating molecular structures sampled along the reaction coordinate. The ensemble of activated complexes holds central importance to such an analysis, but current and near-future experimental techniques cannot resolve their identity. Low occupation of TS structures that have a lifetime of femtoseconds within millisecond enzyme catalytic cycles, precludes characterization by spectroscopic or static techniques.[8] Although indirect approaches such as chemical precedent and TS

inhibitors are available, the only direct experimental determination of the TS complex is measurement of KIEs.[9] Importantly, KIE information requires interpretation (with intrinsic isotope effects) into TS structures.[8]

In this chapter, the effects of treating the TcTS QM region with SCC-DFTB are quantified when considering the catalyzed glycosylation reaction. The transfer of sialic acid from a β-galactose donor to TcTS is now considered to be the forward reaction, in that reaction dynamics are initiated from the equilibrated Michaelis complex. The previously reported FEARCF results for the transition between Michaelis complex and covalent intermediate states is summarized in Figure 7.1.

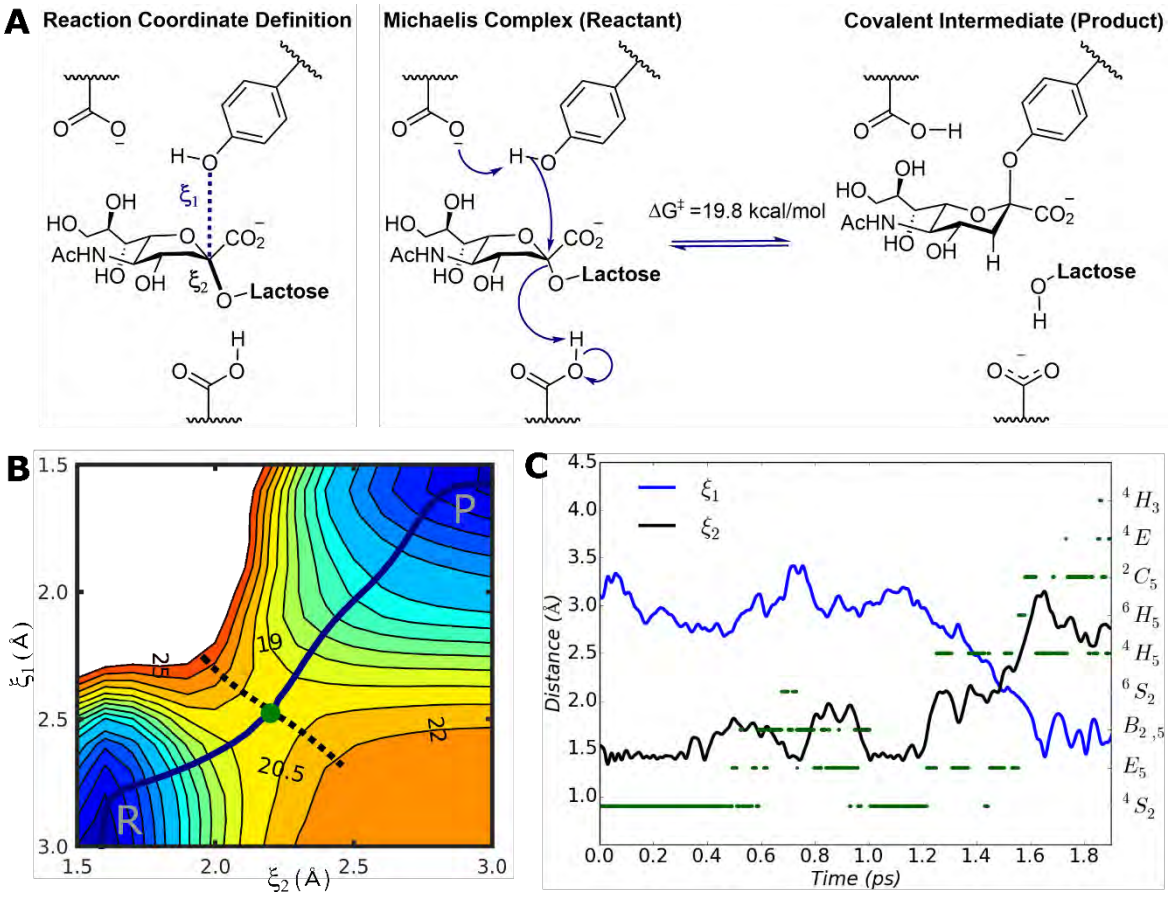


Figure 7.1 (A) The reaction scheme for the first (glycosylation) step of the TcTS catalytic itinerary, which is now treated as the forward reaction is shown alongside the reaction coordinate used to resolve the SCC-DFTB FES (B) A representative SCC-DFTB crossing trajectory is shown in (C).

The FES was calculated from reaction dynamics trajectories in which atoms comprising the reaction coordinate were perturbed by forces obtained directly from the free energy gradient.[10,11] In principle, the same forces can be used to enhance DFT sampling initialized at the Michaelis complex to obtain a high-level molecular orbital description of the chemical reactivity time series. Successful crossing depends on good overlap of the respective potential energy functions as illustrated in Figure 7.2. The B3LYP functional is used here for the DFT level of theory since the repulsive potential for the MIO

parameter set was fitted to B3LYP/6-31G(d). Initial efforts to evolve a crossing trajectory using the 6-31G(d) basis set were unsuccessful, which prompted us to compare the SCC-DFTB/MIO, B3LYP/6-31G and B3LYP/6-31G(d) Hamiltonians, from which molecular geometries are drawn during dynamics.

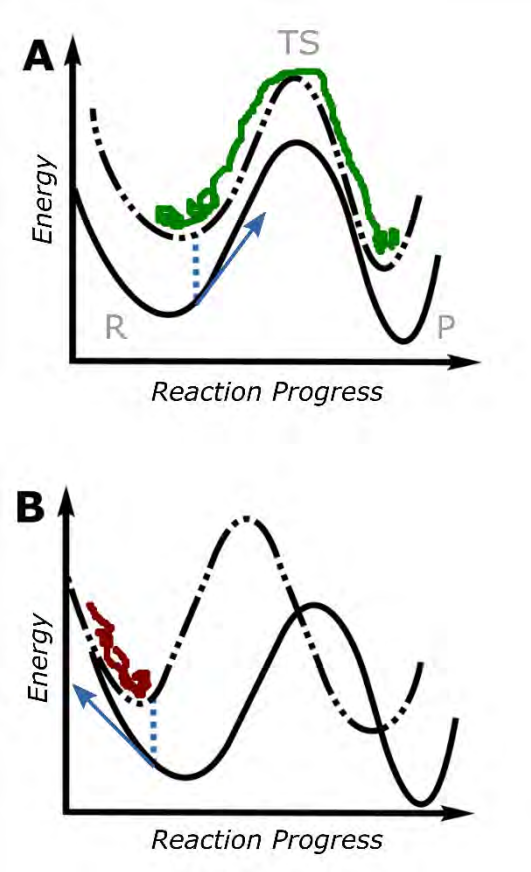


Figure 7.2 Reactive trajectories that are run on the ab initio *surface (dotted line) are perturbed with forces (blue arrows) derived from the semi-empirical PMF (solid line). In (A) the potential functions show sufficient overlap such that the forces drive the* ab initio *trajectory from reactant to product. In (B) the gradients along reaction progress are similar, but the shift in energy minimum results in the driving forces pushing the system away, rather than towards, an activated complex.*

7.1 Theory and Computation

7.1.1 DFT-GGA and Basis Set Dependence

The Kohn-Sham (KS) total energy functional solves the energy of a system of non-interacting electrons and subsumes deviations from the actual system into the exchange-correlation functional (E_{xc}):

$$E[\rho] = \sum_{i}^{n} n_i \left\langle \psi_i \middle| -\frac{1}{2}\nabla^2 - \sum_A^N \frac{Z_A}{|\boldsymbol{r} - \boldsymbol{R}_A|} + \frac{1}{2}\int \frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r}' \middle| \psi_i \right\rangle + E_{xc}[\rho] + \frac{1}{2}\sum_{A}^{N}\sum_{A\neq B}^{N} \frac{Z_A Z_B}{r_{AB}}$$

Equation 7.1

The electron density, $\rho(\boldsymbol{r})$, that minimizes the energy is then found. Practically, the density is constructed from KS molecular orbitals that are formulated as a linear combination of atomic orbital basis functions. Generally speaking, increasing the basis set imparts greater flexibility during the SCF process to find a more accurate electron distribution,[12] although DFT shows reduced basis set sensitivity compared to wavefunction methods.[13] Increasing the number of primitives used to treat split valence orbitals allows independent variation of the coefficients for the inner- and outer-shell basis functions. Adding polarization functions allows a more anisotropic electron distribution. However, greater accuracy is associated with computational cost resulting from the increase in four-center integrals. For the current system of 94 atoms, a basis set of 6-31G(d) was the largest practical basis set that could be employed. MD simulations parallelized over 96 cores completed 1 ps of dynamics over 34 hours.

7.1.2 Approximations Used to Derive SCC-DFTB

SCC-DFTB is derived from the KS total energy functional by defining the density as a reference density perturbed by some density fluctuation, $\rho(\boldsymbol{r}) = \rho_0(\boldsymbol{r}) + \delta\rho(\boldsymbol{r})$, and then expanding E_{xc} in a Taylor series:

$$E[\rho] = \sum_{i}^{n} n_i \langle \psi_i | \hat{H}[\rho_0] | \psi_i \rangle$$

$$+ \frac{1}{2}\sum_{A}^{N}\sum_{A\neq B}^{N} \frac{Z_A Z_B}{r_{AB}} - \frac{1}{2}\iint \frac{\rho_0' \rho_0}{|\boldsymbol{r} - \boldsymbol{r}'|} d\boldsymbol{r} d\boldsymbol{r}' - \int \frac{\delta E_{xc}[\rho_0]}{\delta \rho_0} \rho_0 d\boldsymbol{r} + E_{XC}[\rho_0]$$

$$+ \frac{1}{2}\iint \left[\frac{\delta\rho\delta\rho'}{|\boldsymbol{r} - \boldsymbol{r}'|} + \frac{\delta^2 E_{XC}}{\delta\rho\delta\rho'}\bigg|_{\rho_0} \right] d\boldsymbol{r} d\boldsymbol{r}'$$

Equation 7.2

The terms in the exact expression for molecular energy given in Equation 7.2 can be grouped as follows:

$$E^{Total} = E^{bnd} + E^{rep} + E^{2nd} + \cdots$$

Equation 7.3

where E^{rep} encompasses the second line Equation 7.2 and includes nuclear repulsions as well as the DFT 'double counting' contributions. The SCC-DFTB total energy functional is obtained by truncating the series expansion after the second-order term, and employing two fundamental approximations to accelerate the solution of the generalized eigenvalue problem: the molecular reference density and charge density fluctuation are represented by atomic components ($E[\rho_0, \delta\rho] = E\left[\sum_i^N \rho_a, \sum_a^N \delta\rho_a\right]$), which then forms the basis for evaluation of integrals as summations over pairwise potentials:

$$E = \sum_i^n n_i \langle \psi_i | \hat{H} | \psi_i \rangle + \frac{1}{2}\sum_{AB}^N V_{AB}^{rep} + \frac{1}{2}\sum_{AB}^N \Delta Q_A \Delta Q_B \gamma_{AB}$$

$$\gamma_{AB}^h = \frac{1}{R_{AB}} - S(R_{AB}, U_A, U_B) \times h(R_{AB}, U_A, U_B)$$

Equation 7.4

In the evaluation of E^{bnd} the KS molecular orbitals are expanded in an LCAO basis set such that the diagonal elements of $H_{\mu\nu}$ are calculated as the KS energy of the 'optimized' atomic orbital basis function in the neutral atom, and the off-diagonal matrix elements depend only on interatomic separation of A and B:

$$H_{\mu\nu} = \begin{cases} \varepsilon_\mu^{free\ atom} & if\ \mu = \upsilon \\ \left\langle \phi_\mu \left| -\frac{1}{2}\nabla^2 + \hat{V}_{eff}[\rho_A + \rho_B] \right| \phi_\upsilon \right\rangle & if\ A \neq B \\ 0 & if\ A = \mathrm{B}, if\ \mu \neq \upsilon \end{cases}$$

Equation 7.5

V_{AB}^{rep} in Equation 7.4 is calculated from DFT calculations for a set of reference molecules, and E^{2nd} introduces self-consistency in solving atomic orbital coefficients. The atomic partial charge, ΔQ_A, in the second-order term is calculated as the Mulliken charge, while γ_{AB} is an analytical function that serves as a screening factor for modulating the charge-charge interaction so that it interpolates the correct limiting behavior of the integral in E^{2nd}. At large interatomic distances, the overlap integral tends to 0 and E^{2nd} reduces to the Coulombic interactions between partial charges, while at the atom center γ_{AA}^h reduces to U_A, the second derivative of the energy with respect to charge density fluctuation.

Since $H_{\mu\nu}$ and V_{AB}^{rep} are independent of the specific chemical environment, they are pre-calculated once for every pair of atoms in a reference set of molecules, and then either fit to analytic functions or tabulated for future recall. The electronic parameters – the Hubbard values and compression radii

used for determining the optimized LCAO basis set and neutral atomic densities – allow only fine-tuning in performance, while the repulsive potential parameters are crucial for accuracy.[14] The values of the latter determine bond energies, bond distances and stretch vibrational frequencies. In the present work the MIO parameter is used for which pairwise repulsive terms were parameterized against B3LYP/6-31G(d).

7.1.3 Limitations of SCC-DFTB

Computational studies of enzyme-catalyzed reactions require accurate modeling of both the covalent bonds and intermolecular interactions. This is contingent on proper treatment of electrostatics, charge-transfer effects and dispersion interactions. In general, SCC-DFTB reproduces accurate geometries and predicts reaction energies reasonably well.[15] However, deficiencies arise from the method's derivation from DFT-GGA, as well as the approximations used to accelerate solution of the generalized eigenvalue problem. SCC-DFTB inherits a tendency to overbind covalent bonds and does not treat long-range dispersion interactions, although the latter can be corrected through an empirical correction factor. Additionally, use of a minimal basis and monopole approximation in E^{2nd} results in inaccurate electrostatic properties (polarization, dipole-moments and charge-transfer). As a result, charged systems are not treated particularly well. Specifically, proton affinities between negatively charged molecules are overestimated.[16] On the other hand, hydrogen bonding that relies on the combination of electrostatic and dispersion interactions, is typically underestimated by ~1-2 kcal/mol.[17,18]

7.1.4 *Ab Initio* Reaction Dynamics on Semi-Empirical Free Energy Volumes

In order to run DFT reactive trajectories on a semi-empirical PMF, FEARCF was implemented in NWChem to leverage its performance on massively parallel architectures. NWChem uses data and domain decompositions to accelerate QM/MM calculations in which communication between remote processors is carried out using the Global Arrays toolkit. In NWChem, a top level QM/MM module interfaces between the classical MD and DFT modules, managing initialization, data transfer, and various high-level operations. The QM and MM modules are implemented to accelerate the bottlenecks in evaluating each Hamiltonian over multiple nodes on high-performance clusters. NWChem carries out parallelized diagonalization of the Fock matrix using PeIGS,[19] and integrals are computed simultaneously with TEXAS[20] via a twofold blocking routine. The evaluation of nonbonded interactions in the MM Hamiltonian is accelerated using a domain decomposition. The molecular system is decomposed into rectangular prisms which are distributed as sub-grids to a logically arranged grid of processes. At each MD step, individual processors calculate the interaction energies for sub-box pairs and accumulate force contributions to the appropriate global array. FEARCF was implemented as a Fortran 90 library, and interfaced with NWChem 6.5 by modifying the NWChem MD

module (NWMD). Coordinates and forces are collected in NWMD after their evaluation by the QM/MM module, and then passed to the FEARCF library. The biasing potential is interpolated from the PMF at the reaction coordinate value, and is used to modify the atomic forces. Modified forces are then recollected in the NWMD module and the simulation is evolved.

7.2 Computational Details

7.2.1 DFT/MM Equilibration Trajectories

Here, B3LYP/CHARMM MD trajectories were run with either the 6-31G(d) or 6-31G basis set in NWChem. The QM region included the side chain atoms of Tyr_{342}, Glu_{230} and Asp_{59}, as well as the substrate residues sialic acid and galactose. This definition resulted in a net charge of -2. QM-MM links were treated with hydrogen link atoms and, with the exception of the link bond group, all background charges within 18 Å of the QM region polarized the wavefunction. Numerical integration was performed using the Euler-MacLaurin quadrature in the evaluation of E_{xc} and a grid size was defined to give a total energy accuracy of 1×10^{-6} Eh. The influence of the QM region on the MM chargers was calculated using electrostatic potential (ESP) charges for the QM region. Simulations were conducted in a 40 Å sphere using a leapfrog integrator to evolve the system for 12 ps. As in previous simulations, CHARMM22 was used to model protein residues. Atoms further than 25 Å from the QM region were held fixed, and MM-MM interactions were evaluated with a cutoff of 18 Å. Atom velocities were assigned from a random distribution centered around 298.15 K, and the temperature was then held constant using the Berendsen thermostat for which the temperature relaxation time was 0.1 ps for the initial 2.5 ps and 5 ps thereafter.

7.2.2 QM/MM TS Structures

TS structures were optimized by local optimization as described in Appendix C. Following this procedure, the normal mode corresponding to reaction progress was identified from the numerical Hessian after relaxing the starting structure while restraining atoms comprising the reaction coordinate. The system was then optimized along the negative frequency to a saddle point.

The B3LYP/6-31G(d)/MM TS structure obtained from the B3LYP/6-31G/MM reactive trajectory was additionally confirmed by checking that the structure had a near-equal chance of falling to reactant (Michaelis complex) and product (covalent intermediate). To do so, random velocities were drawn from a Gaussian distribution centered at 298.15 K to initiate unbiased rare event (bluemoon) trajectories from the TS. The trajectories were then monitored to see if they progressed to Michaelis complex or covalent intermediate. Finally, primary ^{13}C and β-dideuterium KIE's were calculated for the B3LYP/6-31G(d)/MM TS structure for comparison against experimentally calculated values. These

ratios were determined from the numerical Hessians of the optimized TS and Michaelis complex structures using the Bigeleisen-Mayer equation.[21]

7.2.3 DFT/MM FEARCF Crossing Trajectories

The Michaelis complex microscopic states prepared by B3LYP/MM equilibration were used as the starting point for simulating B3LYP/6-31G/MM and B3LYP/6-31G(d)/MM FEARCF reactive trajectories. Driving forces were derived from the SCC-DFTB/MIO/MM PMF for which crossing was first observed. Crossing to covalent intermediate was observed with the 6-31G basis set which allowed access to molecular orbitals along the reaction path. NBO analysis was used to map electron densities to chemically intuitive Lewis-type bonding and lone pair orbitals, as well as their antibonding counterparts, compatible with the concept of electron donor-acceptor interaction. This donor-acceptor interaction energy is quantified using second-order perturbation theory analysis of the Fock matrix (refer to Appendix C for details). Further, chemically meaningful NBO atomic charges and Wiberg bond indices can also be derived.[22] NBO analysis was run in Gaussian 09,[23,24] including all background charges in the calculation.

7.3 Results and Discussion

A comparison of the B3LYP/6-31G(d)/MM and SCC-DFTB/MIO/MM potential energy functions was conducted in order to illuminate why DFT reactive trajectories did not successfully evolve from Michaelis complex to covalent intermediate under the influence of external SCC-DFTB FEARCF driving forces. Noting that one of the primary approximations used to speed up SCC-DFTB is the use of a minimal basis set, B3LYP/6-31G/MM has been included for the purpose of qualifying B3LYP basis set dependence for the TcTS system. Both the equilibrium properties of the reactant state (Michaelis complex) and the chemical profile along the glycosidic C – O bond stretch were explored.

7.3.1 Effect of SCC-DFTB Approximations on the Potential Energy Function

In previous KIE[25] and computational[2] studies, as well as in this thesis, the glycosylation step has been found to proceed via a dissociative-type mechanism categorized between A_ND_N and $D_N{*}A_N$. The reaction is accordingly initiated with glycosidic C – O bond stretching accompanied by a degree of nucleophilic participation. The chemical profile of the glycosidic linkage is therefore an important consideration, and a potential energy scan was run along the (SA C-2)…(Gal O-3) bond (ξ_1 for the reaction coordinate defined in Figure 7.1). Extending the glycosidic C – O bond resulted in nucleophilic attack by Tyr_{342} OH for all systems, so we observe a reaction energy profile – which is not, however, the same as the MEP gained from optimization along the full 2-D ξ. The energy profile (Figure 7.3A) shows an obvious trend: improving the level of theory gives rise to a steeper gradient along the C – O bond stretch and a larger enthalpic barrier. Increasing the basis set by addition of polarization

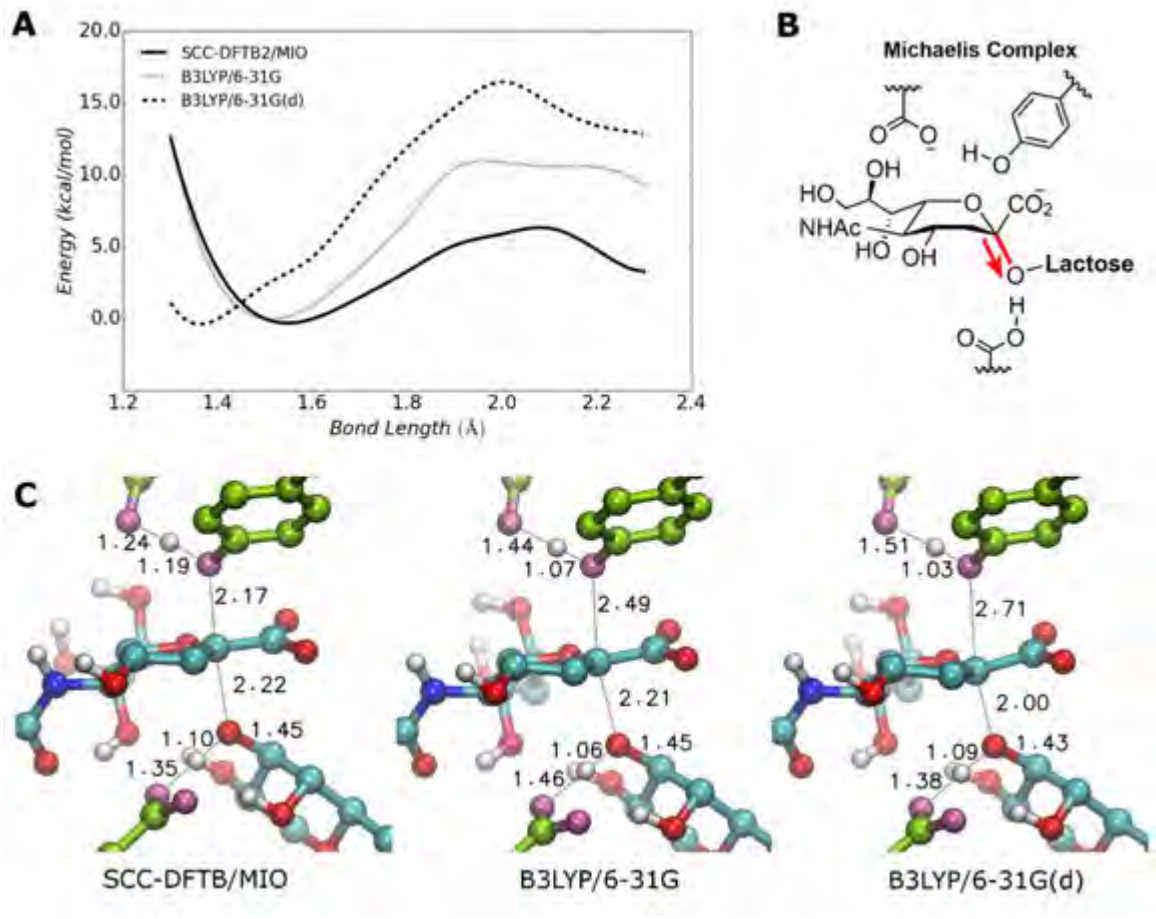


Figure 7.3 (A and B) QM/MM PE scan along the Michaelis complex glycosidic C – O bond for the indicated levels of theory; (C) TS structures obtained by removing the restraint on the glycosidic C – O bond and following the imaginary frequency to a saddle point.

functions further increases the barrier height, but a shift in the position of the energy minimum from ~1.5 Å to ~1.4 Å (cf. 1.42 Å for crystal structure) means that the 6-31G and 6-31G(d) bond stretch gradients are comparable.

Focusing on the energies at long bond distances (~2.0 Å in Figure 7.3A), it can be seen that the B3LYP/6-31G(d) saddle point is sharper and easier to identify than the flatter profiles presented for B3LYP/6-31G and SCC-DFTB. Additional insight is gained by removing the glycosidic bond restraint and optimizing the system along the imaginary frequency to TS structures shown in Figure 7.3C. As could be expected from the similarities in the potential energy scan, glycosidic C – O bond lengths for B3LYP/6-31G and SCC-DFTB are very close (2.22 Å and 2.21 Å respectively), while adding polarization functions to the DFT basis set results in significant shortening of the bond to the departing lactose moiety. On the other hand, improving the level of theory notably alters the reacting geometry of Tyr_{342}, pointing to reduced nucleophilic participation in the TS structure. The earlier stage of proton transfer that activates the Tyr_{342}/Glu_{230} nucleophile pair is reflected in the bond lengths of (Glu_{230} O)...(Tyr_{342} H) and (Tyr_{342} H)…(Tyr_{342} O) shifting from 1.24 Å and 1.19 Å in SCC-DFTB to 1.44 Å and 1.07 Å in B3LYP/6-

31G. Concomitantly, the nucleophilic attack distance increases from 2.17 Å to 2.49 Å. The addition of polarization functions in the basis set further enhances this effect with B3LYP/631G(d) bond lengths recorded at 1.51 Å, 1.03 Å and 2.79 Å for (Glu_{230} O)…(Tyr_{342} H), (Tyr_{342} H)…(Tyr_{342} O) and (Tyr_{342} O)…(SA C-2) respectively.

At the minimum, the B3LYP/6-31G(d) reactant well not only shifts to shorter glycosidic C – O distances, but is also narrower than the SCC-DFTB reactant well which suggests a stiffer molecular geometry. The chemical nature of the Michaelis complex is of particular interest since FEARCF forces are initially applied to the atoms comprising ξ in the reactant state. Therefore, performances of the three levels of theory in treating the Michaelis complex were compared by optimization to minima as well as by generating equilibrium distributions from MD simulations on the respective potential energy surfaces (Figure 7.4). The results for the freely optimized Michaelis complex structures are consistent with those reported above. SCC-DFTB and B3LYP/6-31G show identical glycosidic C – O bond lengths (1.54 Å; Figure 7.4A), while B3LYP/6-31G(d) registers a shorter length (1.50 Å). MD simulations at 298.15 K show similar fluctuations for each level of theory around their respective minimum C – O bond lengths: SCC-DFTB = 1.52 Å ± 0.04 Å, B3LYP/6-31G = 1.52 Å ± 0.04 Å and B3LYP/6-31G(d) = 1.49 Å ± 0.04 Å. Likewise, distributions of distance and angle of attack generated for the displacement reaction are tightly clustered around favorable ξ_1 = (Tyr_{342} O)…(SA C-2) lengths and linear θ_1 = (Tyr_{342} O)…(SA C-2)…(Gal O-3) angles for all levels of theory (Figure 7.4B). On the other hand, support for a stronger glycosidic C – O bond with more accurate treatment comes from the normal mode analysis of the Michaelis complex stationary structures. While the C – O bond stretching mode appears at 795.8 cm^{-1} in B3LYP/6-31G(d), this frequency is red shifted to 751.5 cm^{-1} with the use of a 6-31G basis set. In addition, improving the level of theory constricts the sialic acid ring pucker conformer distribution (Figure 7.4C). SCC-DFTB dynamics displays a flexible sialic acid ring for which ^5E, $^4S_2$ and $B_{2,5}$ pucker conformations are equally sampled with some $^6S_2$ conformations also observed. Improving the level of theory results in a loss of sampling of $^6S_2$, while distortion to $B_{2,5}$ is resisted, increasing observation of the $^4S_2$ pucker conformation.

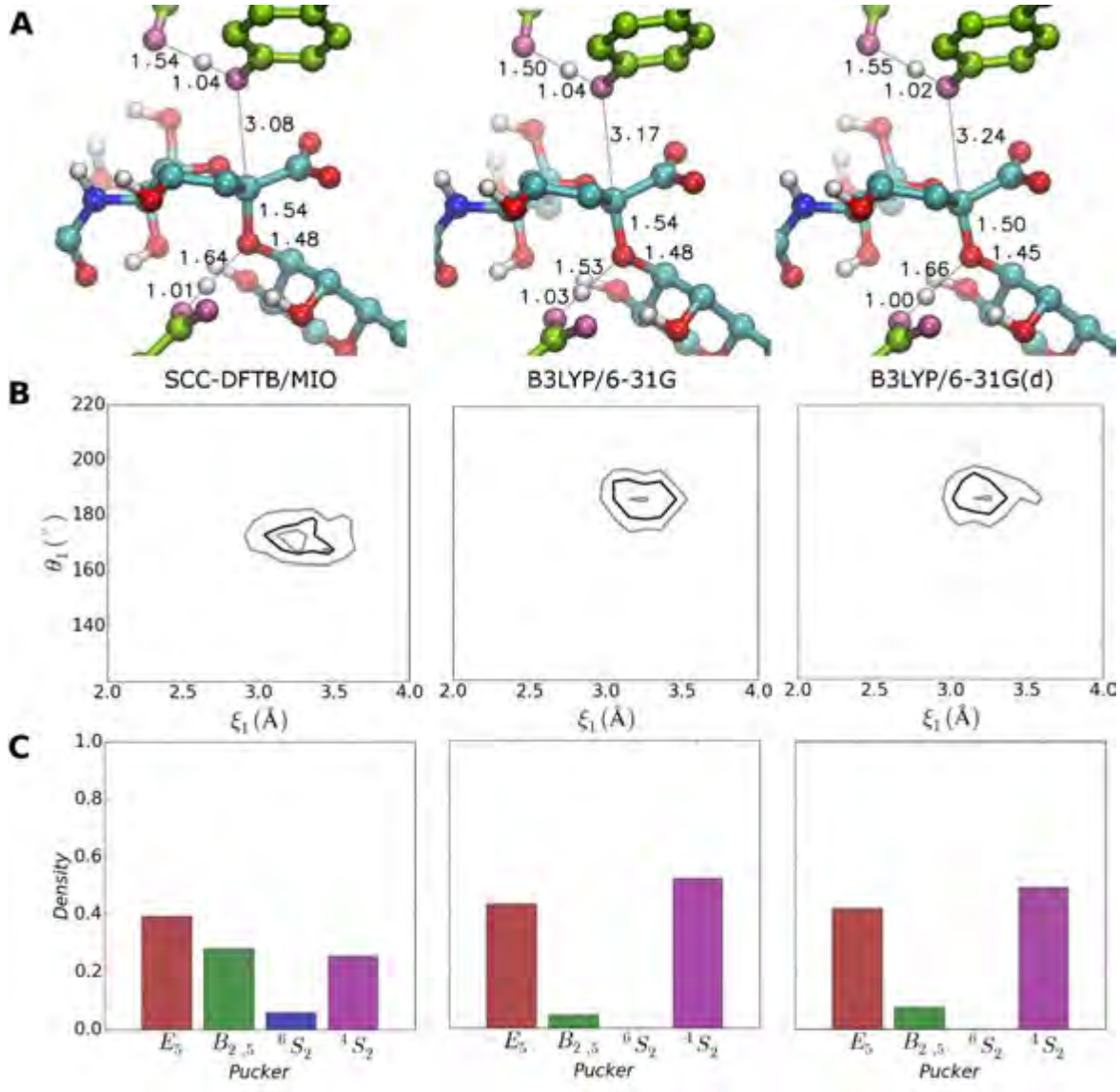


Figure 7.4 Effects of SCC-DFTB approximations on the treatment of the Michaelis complex are summarized in three panes: (A) freely optimized Michaelis complex structures; (B) probability contour plots for the ξ_1 = (Tyr_{342} O)…(SA C-2) bond and θ_1 = (Tyr_{342} O)…(SA C-2)…(Gal O-3) angle defining the displacement NAC; (C) SA ring pucker conformer distribution. Sampling taken from 12 ps QM/MM equilibrium dynamics.

7.3.2 DFT/MM Reactive Trajectory

Analysis of the potential energy functions reveals that improving the level of DFT theory and basis set affects both the glycosidic C – O bond and Glu_{230} – Tyr_{342} reacting geometry. B3LYP/6-31G shows behavior intermediate to that of SCC-DFTB and B3LYP/6-31G(d). Notably, the glycosidic C – O bond shows comparable behavior at the reactant and TS structures when treated with either B3LYP/6-31G or SCC-DFTB. This similarity suggests an overlap with SCC-DFTB/MIO performance that would allow B3LYP/6-31G FEARCF reactive trajectories to be run in the presence of driving forces derived from the SCC-DFTB PMF. Running FEARCF DFT trajectories on the SCC-DFTB free energy not only verifies use of the semi-empirical potential DFT energy function, but also affords access to a molecular orbital description of the reaction. Using the NBO approach, information stored in the molecular orbitals can

be transformed into a representation that is compatible with the key idea of chemical functional groups, and thus enables evaluation and comparison of the behavior of bonding units within distinct molecular environments along reaction progress. Successful crossing from Michaelis complex to covalent intermediate was observed and the representative trajectory is shown in Figure 7.5. Further geometric and NBO analyses were conducted for four snapshots extracted from the reactive trajectory at the Michaelis complex (t_0), at four points along reaction progress (t_1 – t_4) and at the covalent intermediate (t_f). The results are summarized in Figure 7.5, Table 7.1 and Table 7.2.

Like SCC-DFTB (Figure 7.1), the reactive trajectory shows near synchronous lengthening of the glycosidic C – O bond and shortening of the (Tyr_{342} O)…(SA C-2) bond distance. Electron donation into σ^*_{C2-O3} increasingly weakens the glycosidic C – O bond from t_0 (ξ_1= 3.16 Å, ξ_2 =1.49 Å) to t_2 (ξ_1= 2.74 Å, ξ_2 =1.90 Å). The antibonding orbital is stabilized by an additional ΔE2 energy of -35.9 kcal/mol (Table 7.2). Stabilization primarily stems from the endo-anomeric effect, with a nominal contribution from hyperconjugation with σ_{C3-H3} NBOs and electron delocalization of π_{C1-O11}. While the estimated bond order for the glycosidic C – O bond has decreased from 0.76 to 0.46, there remains very little nucleophilic participation. Instead, the interaction between Tyr_{342} OH and SA C-2 is predominantly electrostatic, where Tyr_{342} O and SA C-2 possess oppositely signed charges of -0.76 and 0.53 respectively. Donation of electron density from the SA O-6 lone pair into σ^*_{C2-O3} and concomitant lengthening of the (SA C-2)…(Gal O-3) bond is accompanied by initial sampling of the $^4H_5$ pucker. In this conformation the (SA C-2)…(SA O-6) bond has increased in bond order from 0.96 to 1.14 and SA O-6 has a reduced charge of -0.43.

The glycosidic C – O bond continues to lengthen from t_2 to t_3 (ξ_1= 2.59 Å, ξ_2= 2.25 Å) at the cusp of the TS region. The system fluctuates briefly around a leaving bond distance of ~2.2 Å which is similar to the optimized B3LYP/6-31G TS structure reported above. Consistent $^4H_5$ pucker conformation sampling and the NBO profile at t_3 indicates a TS species with an oxocarbenium-like nature. The glycosidic C – O bond is ostensibly broken with a bond order of only 0.16, and there are no significant donor-acceptor interactions with σ^*_{C2-O3}. This structure accumulates a charge of 0.58 on SA C-2 which compares to 0.49 at t_0. Charge development is quenched by donation of SA O-6 electron density into the (SA C-2)…(SA O-6) bond, such that the bond order has increased form 0.96 at t_0 to 1.23. The resulting π^* orbital is stabilized by the Tyr_{342} O lone pair density with an interaction energy of -6.1 kcal/mol.

Finally, overlap of the Tyr_{342} O lone pair with the developing p^*_{C2} orbital provides a stabilization energy of -141.6 kcal/mol that overcomes the competing stabilization by p_{O6} , π_{C1-O11} and σ_{C3-H3} densities (total ΔE2= -112.4 kcal/mol) to initiate bond formation. Progression of nucleophilic attack forms the covalent intermediate state marked by sampling of a $^2C_5$ sialic acid ring pucker conformation.

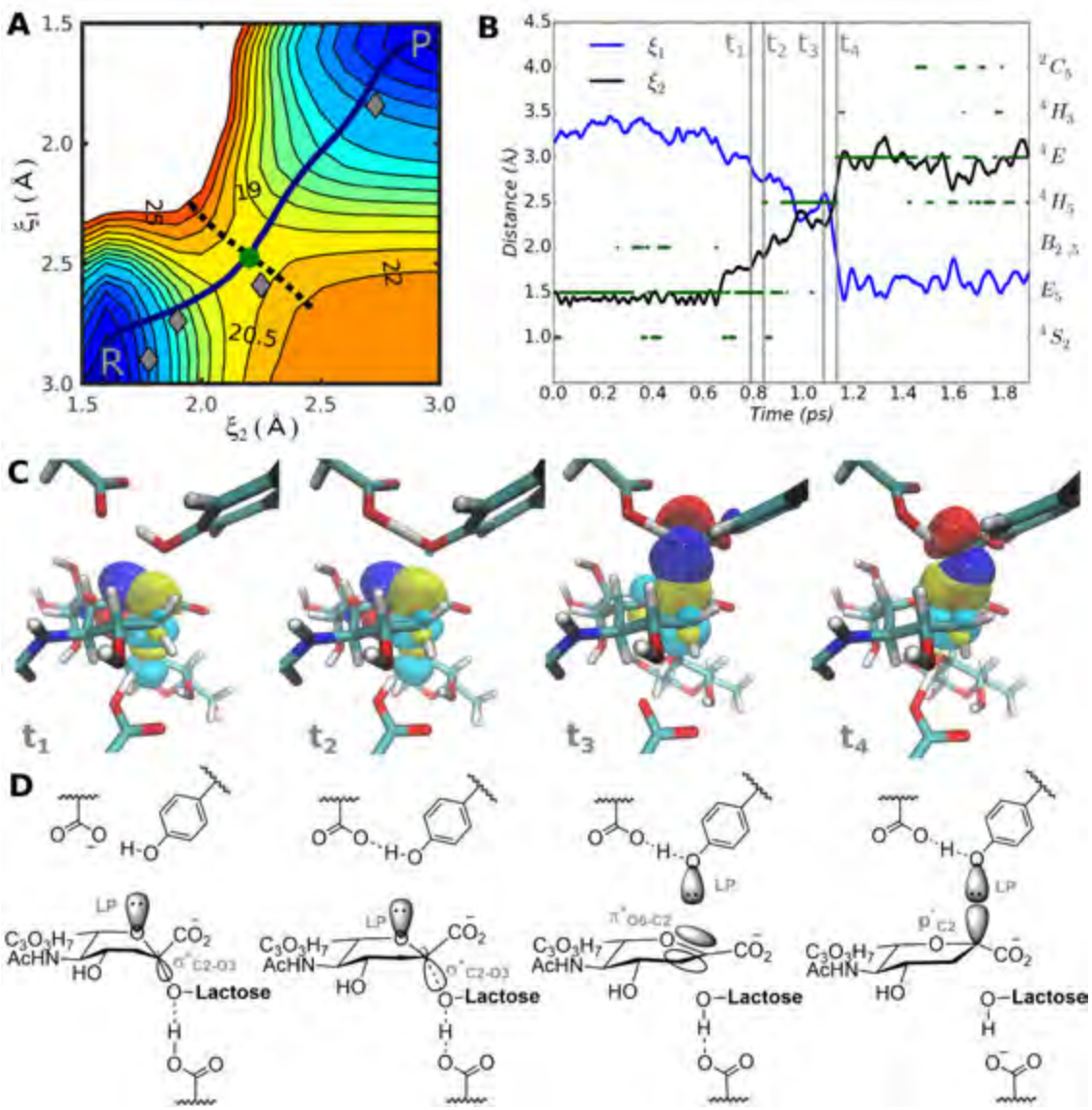


Figure 7.5 (A) Gray triangles on the SCC-DFTB FES indicate the sequential positions of NBO snapshots t_1 – t_4 from Michaelis complex, R, to covalent intermediate, P; (B) Representative B3LYP/6-31G crossing trajectory for which the locations of t_1 – t_4 are indicated with vertical gray lines; (C) NBOs illustrating the significant donor-acceptor interactions for each of t_1 – t_4 are drawn at 0.002 electrons Bohr^{-3} above corresponding chemical drawings in (D).

Table 7.1 Evolution of charge and bond orders calculated with NBO along the B3LYP/6-31G crossing trajectory.

Time	Charges				Bond Length (Å) (Wiberg Bond Order)		
	Tyr_{342} O	SA C-2	SA O-6	SA O-3	(Tyr_{342} O)… (SA C-2)	(SA C-2)…(Gal O-3)	(SA C-2)…(SA O-6)
t_0	-0.73	0.49	-0.55	-0.65	3.16 (0.00)	1.49 (0.76)	1.39 (0.96)
t_1	-0.78	0.51	-0.52	-0.69	2.90 (0.00)	1.78 (0.57)	1.38 (1.07)
t_2	-0.76	0.53	-0.48	-0.74	2.74 (0.02)	1.90 (0.46)	1.32 (1.14)
t_3	-0.74	0.58	-0.43	-0.78	2.59 (0.07)	2.25 (0.14)	1.34 (1.23)
t_4	-0.67	0.54	-0.48	-0.82	1.84 (0.41)	2.73 (0.02)	1.32 (1.17)
t_f	-0.61	0.52	-0.61	-0.81	1.46 (0.80)	3.19 (0.00)	1.45 (0.88)

Table 7.2 Evolution of electron donation – acceptor interactions along the B3LYP/6-31G crossing trajectory.

Frame	Charge Delocalization Interaction		ΔE2 (kcal/mol)
t_0	Hyperconjugation	$\pi_{C1-O11} \rightarrow \sigma^*_{C2-O3}$	-3.3
		$\sigma_{C3-H31} \rightarrow \sigma^*_{C2-O3}$	-3.1
		$\sigma_{C3-H32} \rightarrow \sigma^*_{C2-O3}$	-2.0
	Endo-Anomeric	$p_{O6} \rightarrow \sigma^*_{C2-O3}$	-23.7
	Nucleophilic Attack	$p_{TyrO} \rightarrow \sigma^*_{C2-O3}$	-0.2
t_1	Hyperconjugation	$\pi_{C1-O11} \rightarrow \sigma^*_{C2-O3}$	-6.1
		$\sigma_{C3-H31} \rightarrow \sigma^*_{C2-O3}$	-5.8
		$\sigma_{C3-H32} \rightarrow \sigma^*_{C2-O3}$	-2.4
	Endo-Anomeric	$p_{O6} \rightarrow \sigma^*_{C2-O3}$	-40.3
	Nucleophilic Attack	$p_{TyrO} \rightarrow \sigma^*_{C2-O3}$	-0.6
t_2	Hyperconjugation	$\pi_{C1-O11} \rightarrow \sigma^*_{C2-O3}$	-7.7
		$\sigma_{C3-H31} \rightarrow \sigma^*_{C2-O3}$	-3.8
		$\sigma_{C3-H32} \rightarrow \sigma^*_{C2-O3}$	-4.5
	Endo-Anomeric	$p_{O6} \rightarrow \sigma^*_{C2-O3}$	-55.5
	Nucleophilic Attack	$p_{TyrO} \rightarrow \sigma^*_{C2-O3}$	-2.1
t_3	Nucleophilic Attack	$p_{TyrO} \rightarrow \pi^*_{O6-C2}$	-6.1
t_4	Hyperconjugation	$\sigma_{C3-H32} \rightarrow p^*_{C2}$	-10.5
		$\pi_{C1-O11} \rightarrow p^*_{C2}$	-11.9
	Endo-Anomeric	$p_{O6} \rightarrow p^*_{C2}$	-90.0
	Nucleophilic Attack	$p_{TyrO} \rightarrow p^*_{C2}$	-141.6

7.3.3 Confirmation and Verification of the TS

A snapshot along the reactive trajectory was optimized to a TS structure using the 6-31G(d) basis set while incorporating all background charges. The structure displays bond breaking and forming distances of ξ_1= 2.78 and ξ_2 = 1.97 Å, and so indicates a slightly earlier transition state than TS structures previously isolated for M06-2x/6-31G(d) (ξ_1= 2.57 Å, ξ_2 = 2.10 Å), and SCC-DFTB/MIO (ξ_1= 2.17 Å, ξ_2 = 2.22 Å). The first-order saddle point is confirmed by the presence a single negative

frequency corresponding to motion along the transition coordinate. The saddle point nature of the molecular geometry can be further checked by initiating trajectories with random velocities drawn from a Gaussian distribution centered around 298.15 K, and then measuring the probability of the system falling to the Michaelis complex or covalent intermediate state. When 56 simulations were monitored, a probability of 0.36 was observed for crossing to the covalent intermediate. Although this single conformation does not show exactly equal probability of falling to reactant or product, the structure was verified against experiment through computational KIEs. Applying the Bigeleisen-Mayer equation to the normal mode frequencies returns primary ^{13}C and β-dideuterium KIEs of 1.019 and 1.045 which match, within the reported error, the experimental values of 1.021 ± 0.014 and 1.053 ± 0.010.[25]

7.4 Concluding Remarks

Use of a Hamiltonian that balances accuracy and speed is important for treating the reaction space in free energy methods and has implications for the reliability of extracted structural and thermodynamic data. In FEARCF, driving forces are obtained directly from the gradient of the previous PMF estimate. Therefore, in order to run a successful DFT crossing on an SCC-DFTB PMF, the potential energy functions need to be comparable so that forces differentiated from the PMF drive the system towards the activated complex. This condition was met using B3LYP/6-31G for the TcTS glycosylation reaction. The resulting DFT crossing trajectory shows that the reaction has a mechanism between A_ND_N and $D_N{*}A_N$, also observed for SCC-DFTB. Lengthening of the glycosidic C – O bond is facilitated by electron donation into the C – O antibonding orbital through the endo-anomeric effect and hyperconjugation. The decrease in glycosidic C – O bond order gives rise to a partial (SA C-2)…(SA O-6) double bond. Finally, full attack of Tyr_{342} O lone pair on the SA C-2 p^*_{C2} antibonding NBO results in covalent intermediate formation.

The deviations in the performance of SCC-DFTB from B3LYP/6-31G(d) for the TcTS active site highlights some of the limitations associated with the approximations used to derive the semi-empirical DFT method. The suggestion that the limitations arise primarily from use of a minimal basis set is supported by the intermediary behavior observed for B3LYP/6-31G. Mixing a small amount of *d*-character into the molecular orbitals allows the anti-symmetric combination of atomic basis functions. This will facilitate greater overlap of atomic orbitals to produce areas of high electron density between the atoms and enhance bonding interactions. This effect is apparent in the shorter minimum (SA C-2)…(Gal O-3) bond length and larger force constant of the C – O bond at the B3LYP/6-31G(d) level of theory. These results provide an explanation for not observing crossings from Michaelis complex to covalent intermediate when B3LYP/6-31G(d) FEARCF reactive trajectories are run on an SCC-DFTB PMF. Forces derived from the SCC-DFTB PMF at the DFT equilibrium position will push SA C-2 and galactose O-3

towards each other rather than apart, and a stronger bond may provide a further hindrance to reaction progress. A final consideration that is worth mentioning is the limitations of SCC-DFTB in treating hydrogen bonding and proton affinities for charged species. In SCC-DFTB the proton transfers involved in the glycosylation step require only the available thermal energy and no assistance from a biasing force. Consequently, these bonds were excluded from the reaction coordinate definitions. However, this may not be an appropriate reaction coordinate for the higher level of theory.

7.5 References

(1) Bueren-Calabuig, J. A.; Pierdominici-Sottile, G.; Roitberg, A. E. *J. Phys. Chem. B* **2014**, *118*, 5807.

(2) Pierdominici-Sottile, G.; Horenstein, N. A.; Roitberg, A. E. *Biochemistry* **2011**, *50*, 10150.

(3) Barnett, C. B.; Wilkinson, K. A.; Naidoo, K. J. *J. Am. Chem. Soc.* **2011**, *133*, 19474.

(4) Liu, J.; Wang, X.; Xu, D. *J. Phys. Chem. B* **2010**, *114*, 1462.

(5) Barnett, C. B.; Naidoo, K. J. *J. Phys. Chem. B* **2010**, *114*, 17142.

(6) Jir, S.; Kamp, M. W. V. D.; Mulholland, A. J.; Bana, P.; Otyepka, M. *J. Chem. Theory Comput.* **2014**, *10*, 1608.

(7) Vardi-Kilshtain, A.; Nitoker, N.; Major, D. T. *Archives of Biochemistry and Biophysics* **2015**, *582*, 18.

(8) Schramm, V. L. *Annu. Rev. Biochem.* **2011**, *80*, 703.

(9) Schramm, V. L. *Annu. Rev. Biochem.* **1998**, *67*, 693.

(10) Naidoo, K. J. *Sci China Chem.* **2011**, *54*, 1962.

(11) Naidoo, K. J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 9026.

(12) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*; Second ed.; John Wiley & Sons, Ltd: West Sussex, England, 2004.

(13) Bauschlicher, C. W.; Partridge, H. *Chem. Phys. Lett.* **1995**, *240*, 533.

(14) Elstner, M.; Seifert, G. *Phil. Trans. R. Soc. A* **2014**, *372*, 20120483.

(15) Gaus, M.; Chou, C.-P.; Witek, H.; Elstner, M. *J. Phys. Chem. A* **2009**, *113*, 11866.

(16) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931.

(17) Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861.

(18) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614.

(19) Tilson, J. L. *Int. J. Quantum Chem.* **1999**, *13*, 291.

(20) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251.

(21) Bigeleisen, J.; Mayer, M. G. *J. Chem. Phys.* **1947**, *15*, 261.

(22) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.

(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A. et al.; Gaussian, Inc.: Wallingford, CT, USA, 2009.

(24) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. In *NBO*; Version 3.1.

(25) Yang, J.; Schenkman, S.; Horenstein, B. *Biochemistry* **2000**, *39*, 5902.

8 Conclusion

The effect and limits of using current computational methods in enzymology have been explored in this thesis, with a focus on reaction free energy methods. Specifically, the computation of reaction mechanisms within the framework of TST, discovery of transition state structures and elucidation of the drivers of enzyme selectivity have been addressed. These objectives were undertaken by systematically exploring the enzyme TcTS. In doing so, free energy activation profiles were calculated for the glycosylation and deglycosylation reactions. In line with previous computational and experimental work, the primary displacement reactions were found to have an activation barrier of 20 kcal/mol at the SCC-DFTB level of theory, and proceed via a mechanism between A_ND_N and $D_N{*}A_N$ (with the deglycosylation reaction being more dissociative). Significantly, in using free energies to explore TcTS, some of the theoretical and methodological challenges facing the resolution of reaction profiles for glycoenzyme systems were tackled.

The reaction coordinate definition, and its sampling, is an important consideration in free energy methods. The appropriate level of sampling used to calculate the free energy profile along the *T. cruzi*-catalyzed reaction paths was interrogated by increasing the dimensionality of the reaction coordinate. Inclusion of the (Tyr_{342} O)...(SA H-3) distance in a deconvoluted reaction coordinate revealed competing reactions to DANA that have also been experimentally observed. The observation of DANA formation would not be clear using a conflated reaction coordinate that does not describe hydrogen abstraction. In that case, the sampling of the oxocarbenium ion and DANA species would equally contribute to probability histograms at the same region of the PMF (long glycosidic C – O bond lengths), leading to a loss of chemically and physically meaningful information. In this work, simultaneous monitoring of the hitherto unexplored competition between a minor elimination reaction and the dominant displacement reaction present in both steps of the catalytic cycle, revealed that the dominant displacement reactions display lower barriers (~ 7 kcal/mol). The difference in barrier heights, when seen in the context of the quantitative relationship between $\Delta G^{\ddagger}$ and the chemiflux, provides a kinetic explanation for the selectivity of the displacement products. Also significant to this thesis, is that sampling was conducted for the full reaction space defined by either the 2-D or 3-D reaction coordinates. This allowed free energy analysis of not only the stationary points on the reaction potential energy profile (importantly, the TS structure), but also the regions local to the TS. In Chapter 5 the reaction *col* for the deglycosylation FES was characterized, while in Chapter 6 analyses of the FEV TS volumes showed a greater number of possible transition paths leading to successful crossing reaction trajectories for nucleophilic attack than for SA H-3 proton abstraction.

Computational studies of enzyme reactions aim to balance accuracy and sampling to provide otherwise unobtainable insight into the chemical dynamics. In order to both calculate the thermodynamics governing the reaction pathways and explain the atomic motions that drive the system from reactant to product, it is desirable to use an enhanced sampling technique that allows dynamical and chemical analyses of the underpinning statistics. Free energy methods that artificially reduce the potential energy barrier restrain the system and provide an equilibrated environment within a single simulation. In contrast, constructing histograms from multiple reactive trajectories allows the detailed study of a progression of configurations along the reaction pathway. This approach was used to both characterize the chemical nature of the deglycosylation TS and provide insight into the spatial-temporal drivers of the selectivity observed for TcTS.

Applying FEARCF forces (Chapter 5) to the atoms comprising the (Tyr_{342} O)…(SA C-2) and (SA C-2)…(Gal O-3) bonds resulted in multiple crossing trajectories from the covalent intermediate to the Michaelis complex state. The reaction *col* passing through the generalized TS provided the geometric criteria for isolating an ensemble of TS configurations. Profiling the structures that successfully transited to the Michaelis complex under unbiased MD, and analyzing substrate interactions with the enzyme, offers direct chemical interpretation of the TS. Such a detailed description of the activated complex is inaccessible to experimental methods. In addition, profiling the TS configurations moves beyond a TST description of chemical dynamics where the TS is defined mathematically as a dividing surface, and successful product formation is restricted to the minimum energy path. The results presented provide new evidence to the prevailing premise that there are several pathways from reactant to product passing through the saddle. Similarly, statistical analyses of the dynamics and mechanistic processes (Chapter 6) underpinning the glycosylation and deglycosylation FEVs provides a spatial-temporal description of the selectivity observed for TcTS. Preferential formation of the displacement product over DANA formation is explained by the nucleophile (Tyr_{342} O or galactose O-3) being held in a position with favorable stereoelectronic alignment for attack on SA C-2, while the positional instability of the SA C-3 proton hinders the system from meeting the selective stereoelectronic requirement for proton abstraction.

In balancing accuracy and sampling to calculate the aforementioned results, the reaction spaces were simplified using a QM/MM partitioning scheme and by treating the electron distribution changes with the semi-empirical SCC-DFTB method. Consequently, it was important to verify the resultant reaction profiles. However, as highlighted throughout this thesis, validation is not straight forward and there is no common praxis for assessing the reliability of the information extracted from the free energy results. Ideally, computation should be benchmarked against observed behavior when that data is available for the chemical steps under scrutiny. The experimental reaction rates for TcTS-catalyzed

transfer of sialic acid from PNP – SA and CF_3MU – SA to lactose compare favorably with the barriers computed here of 19.8 (2-D ξ) and 20.6 kcal/mol (3-D ξ) for the rate-limiting deglycosylation step. Importantly, favorable comparison of the free energy of activation with observed reaction rates is not sufficient to comprehensively verify the computational model. It is preferable to verify the molecular complexes sampled in reaction space, and particularly the activated complexes. However, the very need for computational simulations arises from experiment's inability to directly access the TS. Instead, the generalized TS calculated with the semi-empirical electronic method can be evaluated against higher levels of theory. Here, structures from representative SCC-DFB/CHARMM crossings were relaxed onto the M06-2X/6-31G(d) potential energy surface. These structures were then optimized to a TS structure using a local optimization technique, and verified against experiment by computational KIEs. Importantly, the DFT calculations corroborate a dissociative A_ND_N mechanism for the TcTS-catalyzed displacement reactions. Results for the elimination reaction also point to a dissociative-type mechanism, although here the representative DFT TS structures suggest a more dissociative $D_N{*}A_{xh}D_H$ mechanism than SCC-DFTB.

The accuracy of semi-empirical applications and interpretations of electronic degrees of freedom in reaction space were more thoroughly investigated in Chapter 7. SCC-DFTB is formulated using physically motivated approximations to accelerate solution of the truncated DFTB energy function. Understandably, the resulting approximations – namely use of a minimal basis set, the monopole approximation and the neglect of three-center integrals – lead to deviations from full DFT calculations. The performance of SCC-DFTB/MIO in modeling the TcTS-catalyzed reactions was evaluated in order to gauge the level of electronic treatment required for reliable analysis of a glycoenzyme system. The potential energy function, from which molecular structures are drawn during FEARCF calculations, was compared with that of B3LYP since the MIO repulsive parameters were parameterized against B3LYP/6-31G(d). SCC-DFTB models the Michaelis complex state reliably with a somewhat shorter glycosidic C – O bond and more flexible sialic acid ring pucker displayed. Comparison of potential energy scans along the breaking glycosidic bond showed a similar profile for B3LYP/6-31G(d), B3LYP/6-31G and SCC-DFTB/MIO, but with DFT treatments showing a stronger glycosidic C – O bond. The intermediary results observed for B3LYP/6-31G point to the minimal basis set as the primary limitation in using SCC-DFTB when modeling enzymatic reactions transforming sialic acid substrates. The overlap of performances for SCC-DFTB/MIO and B3LYP/6-31G allowed successful evolution of reactive trajectories from Michaelis complex to covalent intermediate when applying atomic forces derived from the SCC-DFTB PMF. This, importantly, provided access to the molecular orbital changes along reaction progress and offers a new level of insight into the chemical dynamics. The reaction profile shows the dissociative nature of the mechanism, with development of a (SA C-2)…(SA O-6) partial

double bond and a small perturbation energy associated with the donation of electron density localized on Tyr_{342} O into (SA C-2)…(SA O-6) π^* and then SA C-2 p^* anti-bonding NBOs.

A technical innovation was necessary to make simulation of the DFT/MM reactive trajectories tractable. Consequently, I ported FEARCF as a Fortran 90 library that interfaces with NWChem. The library will be made available in the near future at the Scientific Computing Research Unit repository (https://bitbucket.org/scientificcomputing). These reactive trajectories mark an additional level of parallelization to the FEARCF scheme. Each j^{th} simulation in the ensemble of concurrent MD simulations comprising the i^{th} iteration is now also parallelized. With the advances in CPU and GPU technology, and optimization of scientific computing algorithms for these architectures, it will soon be possible to routinely run full DFT/MM free energy profiles for enzyme systems. For example, in DFT/MM reaction dynamics simulations of the chorismate mutase enzyme active site containing 24 atoms (not included in this thesis), 7 FEARCF iterations each comprising 10 simulations can be calculated. If each of the simulations is parallelized over 48 cores, 10 ps of dynamics per day is obtainable. Running the FEARCF calculation over a week will construct the weighted histograms along the reaction path using sampling from 700 ps of DFT/MM dynamics. This will allow improved treatment of the electronic degrees of freedom in reaction space when calculating the reaction free energy.

However, the $O(N^3)$ scaling of DFT with the number of electrons means that such calculations are not yet tractable for common glycoenzyme systems. In this thesis, timings of 0.7 ps a day were obtained when running B3LYP/6-31G(d)/MM MD simulations of the 94 atom TcTS QM region over 96 cores. Furthermore, the sialic acid unit in this system does not contain a bond to phosphorous that activates many glycoside donors. Until technological and algorithmic developments have accelerated QM calculations to allow routine treatment of these systems with *ab initio* methods, simplification of the reaction space will continue to be necessary. It is important that models are applied carefully. It has been shown that, considering its limitations, SCC-DFTB performs reliably for the TcTS system. Employing SCC-DFTB with a free energy method that allows chemical and dynamical analyses of the underpinning statistics has provided new insight into the activity of this glycoenzyme system.

Appendix A: Simulation Methods

Explicit Solvation Models

In periodic boundary conditions (PBC) the molecular system is placed inside a unit cell which is then replicated in all directions. The shape of the cell must be such that the central cell fills all of space by translation operations in three dimensions. The cubic cell is the simplest shape to fulfil this criterion, and its use in periodic boundary conditions is illustrated using a 2-D example in Figure A1. The number of particles within the central box is held constant by replacing any particle leaving the box with an image particle that enters from the opposite side. Although the cubic cell is the simplest, cells which have a more spherical geometry, such as the truncated octahedron and rhombic dodecahedron, will require fewer particles and are more computationally efficient. Thus, the truncated octahedron has been widely used in the all-atom simulations of globular proteins.[1] Nonbonded interactions under periodic boundary conditions are calculated using the minimum-image convention, in which the interactions of an atom are restricted to only the nearest image of the other atoms. This leads to the condition that the unit cell must be large enough so that the solute does not interact with itself and that no water molecules interact with the solute twice. In order to achieve this, the shortest edge of the unit cell should be greater than twice the length of the nonbonded cutoff employed.[1]

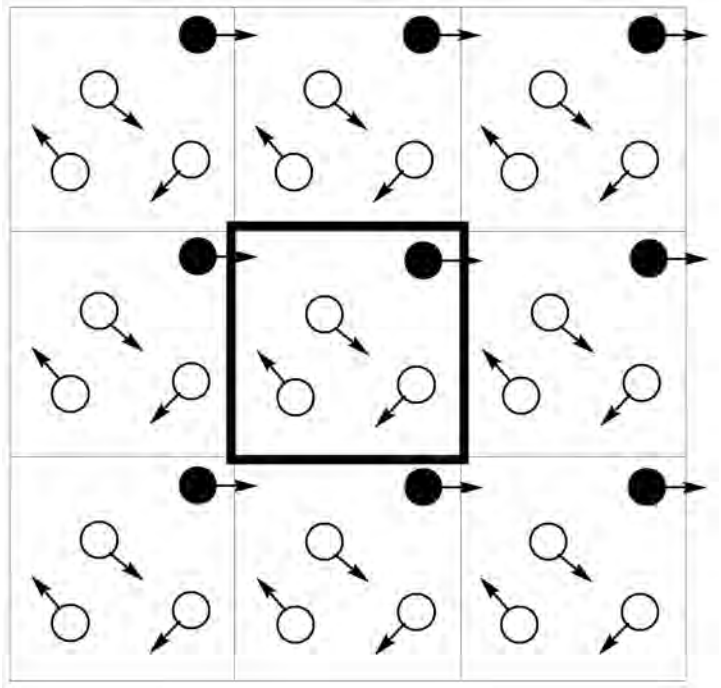


Figure A1 Illustration of periodic boundary conditions. The central cell has a bold border and is translated to fill the 2-D space with image cells. Any atom which leaves the cell (shown by the black atom) is replaced by its image from the opposite side. Illustration adapted from Leach.[1]

Enzyme reaction dynamics simulations focus on the protein active site, and more specifically on the electron redistribution processes at the catalytic site. Therefore, it is a defensible approach to model a small number of particles far away from the catalytic site as if they were in bulk fluid rather than undertake the computational expense of explicitly simulating the bulk solvent. In stochastic boundary conditions[2], a sphere of explicit environment molecules is centered on the active site (Figure A2; explicit water molecules are not shown). The enzyme and water molecules are then divided into three regions. All residues with an atom within the reaction region, $\boldsymbol{R_1}$, are subject to Newtonian dynamics, while all residues with an atom outside $\boldsymbol{R_2}$ constitute the reservoir region and are kept fixed. Atoms that fall between the reaction and reservoir regions (i.e. between $\boldsymbol{R_1}$ and $\boldsymbol{R_2}$) belong to the buffer region. The atoms in this region evolve in time according to an adapted form of the Langevin equation:

$$m_A \frac{d^2 x_A(t)}{dt^2} = \boldsymbol{F}_A\{x_A(t)\} - m_A \boldsymbol{\mathcal{B}}_A^2 \left[x_A(t) - x_A^{ref}\right] - \gamma_A \frac{dx_A(t)}{dt} m_A + \boldsymbol{\mathcal{F}}_A(t)$$

Equation A1

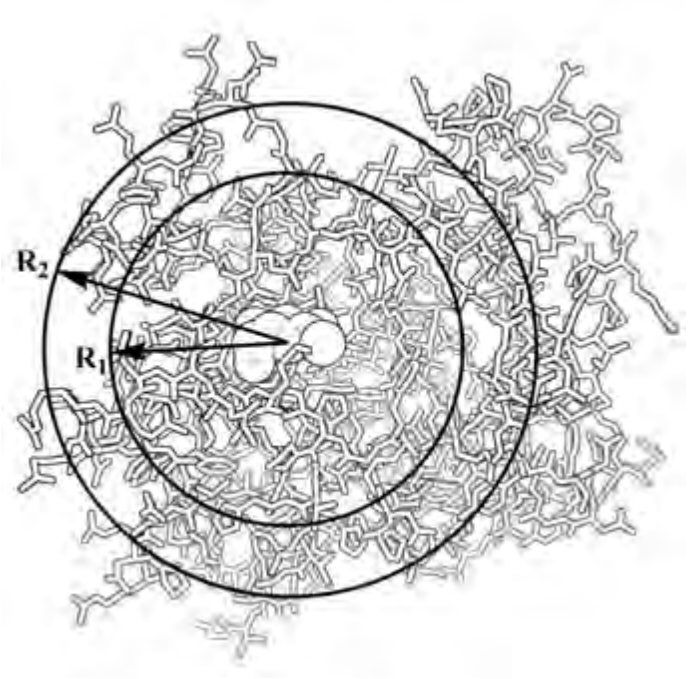


Figure A2 Illustration showing how stochastic boundary conditions are used to study enzyme reaction dynamics. This illustration does not display the explicit waters comprising the sphere of radius R_2.

In Equation A1, $\boldsymbol{\mathcal{B}}_i^2\left[x_A(t) - x_A^{ref}\right]$ are boundary forces applied to solute atoms derived from atomic mean-square fluctuations. These boundary forces are gradually scaled over the buffer region so that they are weakest at the boundary of the free moving reaction region and strongest near the boundary of the static reservoir region. Explicit water molecules are allowed to diffuse between the reaction and reservoir regions, and are acted upon by a deformable boundary potential which maintains the correct average distribution of water molecules and prevents water 'evaporating' into the vacuum. The remaining forces in Equation A1 originate from Langevin dynamics and simulate the effect of bulk water. $\boldsymbol{\mathcal{F}}_i$ are the forces of random collisions associated with energy and temperature, while $\gamma_A \frac{dx_A(t)}{dt} m_A$ simulates the frictional drag and removes energy from the system.[1,3]

Treating Long-range Electrostatic Interactions

The Ewald Summation

When using periodic boundary conditions, the electrostatic interaction between a charge and all its periodic images can be accurately computed using the Ewald summation.[4,5] For a box defined by $\left(n_x L, n_y L, n_z L\right)$, where n, is its position at a cubic lattice point and L is the length of the cube, the charge-charge contribution to the potential energy due to all pairs of charges can be calculated:

$$U_{coulomb} = \frac{1}{2}\sum_{|n|}^{\infty}\sum_{A}^{N}\sum_{B}^{N}\frac{Q_A Q_B}{4\pi\varepsilon_0|r_{AB}+n|}$$

Equation A2

where r_{AB} is the minimum distance between charges Q_A and Q_B and $\varepsilon_0$is the dielectric constant. Equation A2 converges slowly and depends on the order in which its terms are considered (conditionally convergent series). The Coulomb potential was reformulated by Ewald[4] as the summation of three terms each of which converge more rapidly than Equation A2:

$$U_{coulomb} = U_{real} + U_{fourier} + U_{self}$$

Equation A3

In the real space summation, U_{real}, each charge is considered to be surrounded by a neutralizing charge distribution of equal magnitude but of opposite sign. The interactions of a set of distributions added to cancel the initial neutralizing Gaussians are calculated in reciprocal space, $U_{fourier}$. Finally, a self-term, U_{self}, cancels the interaction of each of the introduced artificial counter-charges with itself.

Truncation Potentials

The evaluation of the nonbonded interactions for all atom pairs in the molecular system scales as N^2 and so is the most time consuming operation in the energy calculation.[1,3] The expense can be minimized by truncating the pair potential for the nonbonded terms. Thus, nonbonded interactions are only calculated between neighbor atoms lying within a specified cutoff distance. The assumption is that the forces are negligible at the cutoff boundary because the van der Waals and electrostatic potentials tend to zero as the distance between the atom pairs tends to infinity. This is appropriate for the short-ranged van der Waals interactions since the Lennard-Jones potential has a r^{-6} distance dependence. For a system with bare charges or with correlated dipoles, such as an α-helix, the Coulomb term is proportional to r^{-1}.[6]

However, discontinuity of the electrostatic interactions causes fluctuations in the molecular potential near the cutoffs.[1,3,6] One way to reduce this noise is to use group-based cutoffs. In these methods the distance between neutral groups of atoms are calculated and pairwise interactions are only calculated between the atoms of two groups if the groups fall within a cutoff distance. Since the electrostatic interaction between two neutral groups is proportional to r^{-3}, the range of the interaction is reduced considerably.[1,3,6]

The nonbonded potential near the cutoff value can also be smoothly attenuated by applying a switching or shifting function.[7] In the former method, the entire nonbonded potential energy is shifted so that the interaction potential is zero at the cutoff distance (Figure A3). On the other hand, switching leaves the potential unaltered until a first cutoff value. Between this first cutoff value and the last cutoff value the interaction potential is tapered to zero. The CHARMM22/CMAP and CHARMM36 force fields were used in modeling the protein and carbohydrate atoms respectively. The former was parameterized using a shifting potential to truncate the long-range van der Waals interactions and a switching function acting between 10.0 and 12.0 Å to attenuate the electrostatic interactions. In the CHARMM36 force field the same nonbonded parameters were used except the truncation potentials were updated to force-shifting and force-switching algorithms.

Computing all atom pair distances to evaluate whether they fall within the cutoff is still time consuming.[1] On the assumption that an atom's neighbors do not change significantly over 10 or 20 MD steps, a nonbonded neighbor list can be used to compute the interactions between previously identified neighbors.[1] This list stores atoms within the cutoff distance, together with all atoms that are slightly further away. This ensures that only the distances between the central atom and the atoms in its neighbor list have to be computed. The frequency of updating the list should be slow enough to be computationally efficient but fast enough to ensure accuracy.

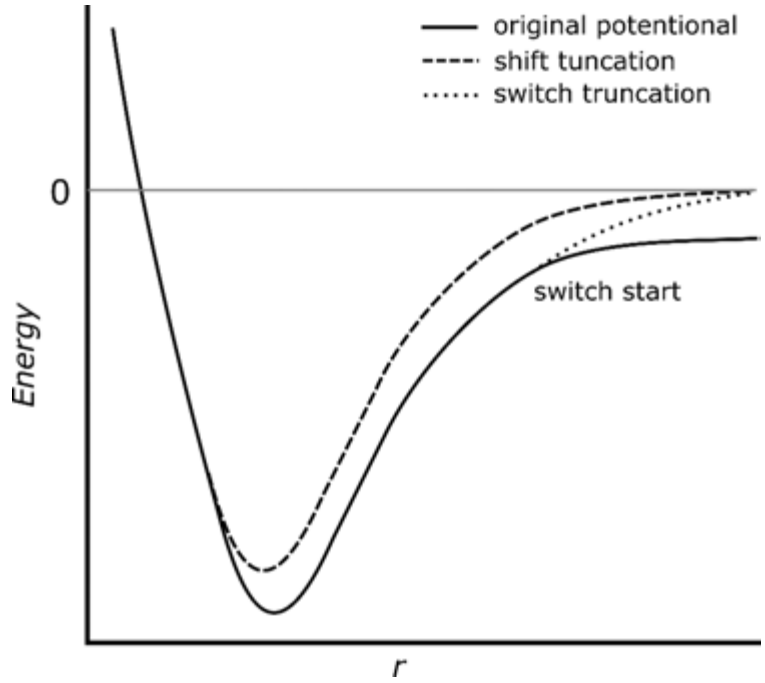


Figure A3 Illustration of the methods used for smoothing the nonbonded interaction potential to zero at a specified nonbonded cutoff distance.

Integrating the Equations of Motion

MD trajectories of an enzyme system are evolved by solving the differential equations embodied in Newton's second law:

$$\frac{d^2x_A(t)}{dt^2} = \frac{\boldsymbol{F}_A\{x(t)\}}{m_A}$$

Equation A4

Equation A4 describes the motion in the x-direction of particle A with a mass m that is being acted upon by force $\boldsymbol{F}$. The next configuration at any future time can be calculated from an initial configuration by solving this second order differential equation for every particle in the system and assuming constant acceleration. This deterministic method is applied to a series of time intervals, dt, to obtain the dynamic behavior of the system. The forces acting on each particle are calculated at a time t from the differentiation of the system's Hamiltonian. The accelerations of the particles are then determined from the forces and used to calculate the positions and velocities at a time $t + dt$.[1]

When solving Newton's second law of motion, the force acting on each particle will change as the particles move. This many-body problem cannot be solved analytically and the equations of motion embodied in Equation A4 are integrated using a finite difference method.[1] There are many algorithms for achieving this. These algorithms assume that the positions and dynamic variables can be approximated as Taylor series expansions:[1,3]

$$r(t+dt) = r(t) + dtv(t) + \frac{1}{2}dt^2a(t) + \frac{1}{6}dt^3b(t) + \frac{1}{24}dt^4c(t)\cdots$$

Equation A5

$$v(t+dt) = r(t) + dta(t) + \frac{1}{2}dt^2b(t) + \frac{1}{6}dt^3c(t)\cdots$$

Equation A6

$$a(t+dt) = r(t) + dtb(t) + \frac{1}{2}dt^2c(t)\cdots$$

Equation A7

In Equations A5 – A7, v is the velocity and a is the acceleration. CHARMM uses the Verlet,[8] Leap-frog[9] and Velocity Verlet[10] algorithms to numerically solve Newton's equations of motions for all atoms of the system.

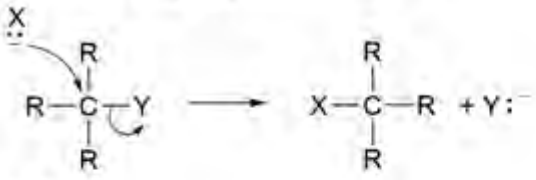
Appendix B: Guthrie-Jencks Mechanistic Nomenclature

The IUPAC 'System for Symbolic Representation of Reaction Mechanisms' is a mechanistic nomenclature detailing the covalent-bond-making and -breaking steps of elementary reactions comprising a reaction mechanism. These steps involve core atoms undergoing transformation, peripheral atoms within the substrate molecule, and carrier atoms that remain part of an external reagent molecule throughout the reaction. The advantage of the nomenclature, as stated by Guthrie and Jencks, is its ability to indicate 'whether processes are concerted or occur in separate steps and, if they are separate, whether the intermediate species have a sufficient lifetime to diffuse freely through the solvent before reacting further'.[11] The symbols used to describe the relevant steps of elementary reactions are summarized in Table B1, after which, example displacement and elimination transformations are given.

Table B1 Symbols used to describe the relevant steps of elementary reactions. An abbreviated version of the table presented in Guthrie and Jencks.[11]

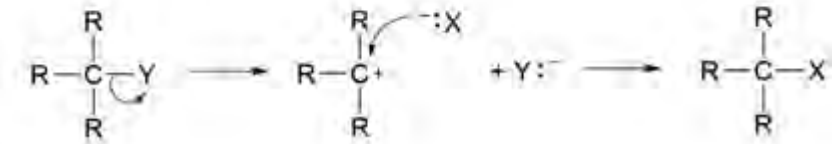
Symbol	Placement	Meaning
A	on the line	bond making (association)
D	on the line	bond breaking (dissociation)
+	on the line	stepwise process
*	on the line	same as +, but the intermediate is short lived i.e. steps occur faster than an intermediate reaches diffusional equilibrium with the bulk solvent
E	subscript	electrophilic (at core atom)
H	subscript	same as E, with hydrogen as electrophile
N	subscript	nucleophilic (at core atom)
xh	subscript	bond making or breaking between hydrogen and a hydrogen carrier reagent atom

Concerted S_N2-type Substitution



A_ND_N

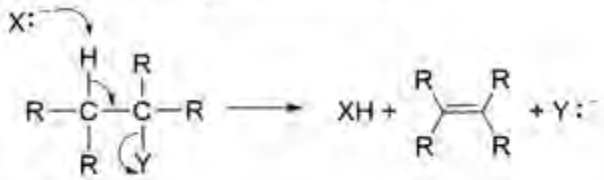
Stepwise S_N1-type Substitution



D_N+A_N

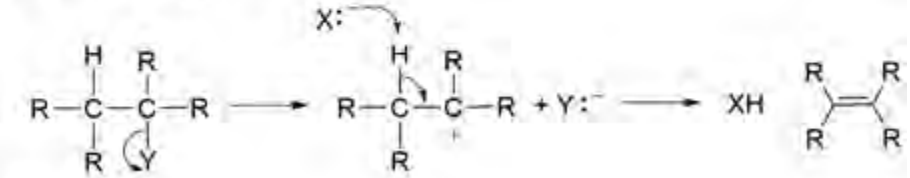
Scheme C1 Example displacement transformations are shown. 'A_N' indicates that the nucleophile X associates with the single core atom, while 'D_N' describes the dissociation of the leaving nucleophile group Y from the core atom. '+' in the stepwise substitution denotes that the carbocation intermediate reaches diffusional equilibrium with the bulk solvent.

Concerted E2-type Elimination



$A_{xh}D_HD_N$

Stepwise E1-type Elimination



$D_N+A_{xh}D_H$

Scheme C2 Example elimination transformations are shown. The use of lower-case subscripting in A_{xh} indicates that the bond undergoing change is between 'non-core atoms'.

Appendix C: Analysis Methods and Conditions

TS Volume Analysis

The TS volumes for the competing reactions in the TcTS catalytic itinerary were calculated from discrete sets of grid points. Points were selected for the glycosylation reaction on the criteria that they were within 0.3 Å of the TS locations, had a gradient smaller than 1.7 and had an energy value within ±1.5 kcal/mol of $\Delta G^{\ddagger}$ (i.e. 18.0 - 21.0 kcal/mol for displacement TS and 26.4 - 29.4 kcal/mol for elimination). The TS volumes for the deglycosylation reaction were calculated from grid points within 0.6 Å of the TS locations, had a gradient less than 1.2 and had an energy within ±1.5 kcal/mol of $\Delta G^{\ddagger}$ (i.e. 19.1 - 22.1 kcal/mol for the displacement reaction and 25.9 - 28.9 kcal/mol, for the elimination reaction).

3-D Voronoi tessellation diagrams were generated from the selected points in MATLAB.[12] This is achieved by decomposition of the space around each selected grid point $x(i)$ into a Voronoi regions, R_i. The result of the decomposition is that an arbitrary point within the Voronoi region R_i is closer to point i than any other point. Once the vertices are determined such that all Voronoi regions are defined, the enclosed volume can be calculated.

QM/MM TS Optimization

In the case of SCC-DFTB calculations, dynamic snapshots were minimized onto the PES while restraining the atoms comprising the reaction coordinate. Eigenvector following was then used to optimize the system to a TS structure with a single negative frequency. The optimization was carried out using POLYRATE[13] that interfaces with CHARMM through the CHARMMRATE[14] module. In my experience, the TS optimization would often end up following an eigenvector that did not represent the reaction coordinate. The system was therefore optimized to a minimum at each step of the eigenvector following scheme while keeping the bond lengths comprising the reaction coordinate restrained. The Hessian was then also recalculated at each step.

In the case of DFT calculations, snapshots from the SCC-DFTB/CHARMM dynamics were relaxed onto the DFT PES while restraining the atoms comprising the reaction coordinate. This was achieved using the DRIVER routine in NWChem 6.5 where QM/MM calculations proceed by cycles of QM calculations in the presence of a fixed MM region, followed by MM calculations keeping the QM region fixed. The numerical Hessian was then calculated once at this stage before eigenvector following was used to optimize the system to a TS structure. Numerical integration was performed using the Euler-MacLaurin quadrature in the evaluation of E_{xc} and a grid size was defined to give a total energy accuracy of

1×10^{-6} Eh. QM-MM links were treated with hydrogen link atoms and, with the exception of the link bond group, all background charges polarized the wavefunction. The influence of the QM region on the MM chargers was calculated using electrostatic potential (ESP) charges for the fixed QM region. Calculations were conducted in a 40 Å sphere using CHARMM22 to model protein residues. Atoms further than 25 Å from the QM region were held fixed, and MM-MM interactions were evaluated with a cutoff of 18 Å. The same conditions used for TS optimization were employed for normal mode analysis, i.e. all background point charges polarized the QM wavefunction.

Natural Bond Orbital Analysis

In NBO analysis, the diffuse electron density of the *n*-electron molecular wavefunction is made mathematically compatible with the electron donor-acceptor concept of interaction. Natural atomic orbitals (NAOs) are localized one-center orbitals that describe the electronic structure around each nucleus embedded in its local molecular environment. NAOs can be determined from a standard basis set of atom-centered orbitals. An orthonormal set of NBOs are then constructed from one-center Natural Hybrid Orbitals that are in turn an optimized linear combination of NAOs on the given center. In this process electron occupancies are maximized in particular directed regions in space to give a chemically intuitive Lewis-type description of the total *n*-electron density. The resulting NBOs are categorized as either bonding and lone pair orbitals, or as the antibonding and Rydberg counterparts. Electron delocalization from filled NBOs to vacant NBOs causes a variational lowering of the total energy that can be quantified by second-order perturbation theory analysis of the Fock matrix:[15]

$$E^2 = \Delta E_{ij} = q_i \frac{F_{ij}^2}{\varepsilon_i - \varepsilon_j}$$

Equation A8

where q_i is the donor orbital occupancy, F_{ij} are the Fock matrix elements between the NBO i and j, and ε_i and ε_j are the orbital energies.

Appendix D: Results

RMSD Analyses for TcTS Equilibration MD Trajectories

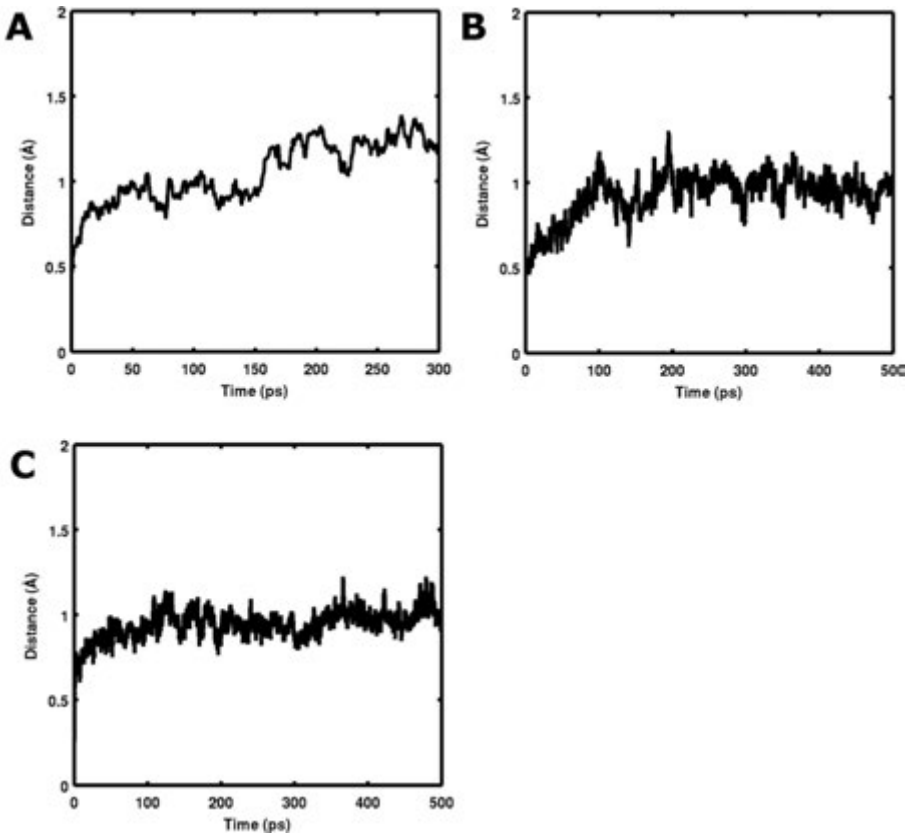


Figure D1. RMSD time series plots are presented for (A) protein backbone atoms over the 300 ps MD equilibration of the Michaelis complex under NPT (B) protein residues 8 Å from QM atoms for the SCC-DFTB/MM equilibration of the Michaelis complex in a water sphere and (C) protein residues 8 Å from QM atoms for the SCC-DFTB/MM equilibration of the covalent intermediate in a water sphere.

SA C-4 OH Hydrogen Bond

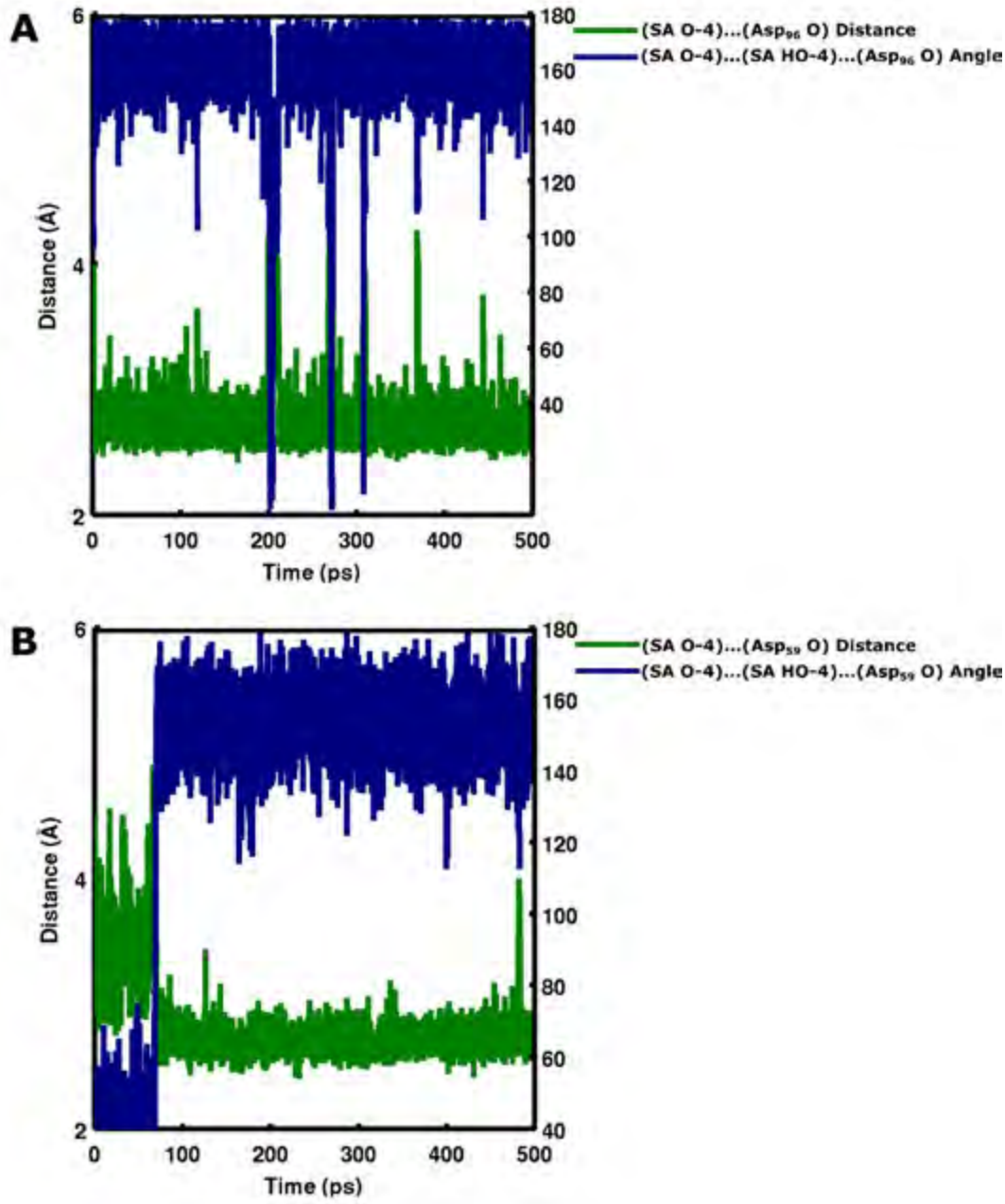


Figure D2 Time series plots are displayed for 500 ps SCC-DFTB/MM equilibration in a water sphere to show the alternative hydrogen bonding interactions of the SA C-4 hydroxyl group with (A) Asp_{96} in the Michaelis complex and (B) Asp_{59} in the covalent intermediate.

References

(1) Leach, A. R. *Molecular Modelling: Principles and Applications*; Addison Wesley Longman Limited: Harlow, 1996.

(2) Berkowitz, M.; McCammon, J. A. *Chem. Phys. Lett.* **1982**, *90*, 215.

(3) Jensen, F. *Introduction to Computational Chemistry*; 2nd ed.; John Wiley & Sons, Ltd, 2007.

(4) Ewald, P. P. *Ann. Phys.* **1921**, *369*, 253.

(5) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H. et al. *J. Chem. Phys.* **1995**, *103*, 8577.

(6) van Gunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem. Int. Ed.* **1990**, *29*, 992.

(7) Steinbach, P. J.; Brooks, B. R. *J. Comp. Chem.* **1994**, *15*, 667.

(8) Verlet, L. *Phys. Rev.* **1967**, *159*, 98.

(9) Hockney, R. W. *J. Comput. Phys.* **1970**, *9*, 136.

(10) Swope, W. C. *Journal of Chemical Physics* **1982**, *76*, 637.

(11) Guthrie, R. D.; Jencks, W. P. *Acc. Chem. Res.* **1989**, *22*, 343.

(12) *MATLAB*; R2015a ed.; The Mathworks Inc.: Massachusetts, 2015.

(13) Corchado, J. C.; Chuang, Y.-Y.; Fast, P. L.; J. Villa; Hu, W.-P. et al. In *POLYRATE*; Version 9.0 ed.; University of Minnesota: Minneapolis.

(14) Garcia-Viloca, M.; Alhambra, C.; Corchado, J. C.; Sanchez, M. L.; Villa, J. et al. In *CHARMMRATE*; Version 2.0 ed.; University of Minnesota: Minneapolis, 2002.

(15) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.