PREDICTING AND CHARACTERISING

CHEMICAL AND BIOCHEMICAL PROCESSES

COMPUTATIONAL APPROACHES TO PREDICTING AND CHARACTERISING CHEMICAL AND BIOCHEMICAL PROCESSES

By

YULI LIU

B.Eng. (Fine Chemicals and Engineering)

M.Sc. (Applied Chemistry)

A Thesis

Submitted to the School of Graduate Studies

In Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

McMaster University

DOCTOR OF PHILOSOPHY (2011) McMaster University

(Chemistry & Chemical Biology) Hamilton, Ontario, Canada

TITLE: Computational approaches to predicting and characterising chemical and biochemical processes

AUTHOR: Yuli Liu B.Eng. – Fine Chemicals & Engineering
(Nanjing University of Science & Technology)
M.Sc. – Applied Chemistry
(Nanjing University of Science & Technology)

SUPERVISOR: Professor Paul W. Ayers

NUMBER OF PAGES: xvi, 212

ABSTRACT

The prediction and characterisation of chemical and biochemical processes are fundamental tasks in computational chemistry. Small chemical systems can be characterised by the stationary points on potential energy surface and reaction paths linking them. For large biological systems, statistical sampling is required to characterising their average properties.

This thesis presents my Ph.D. work on developing new methods to predict and characterise chemical and biological processes. Two path-finding methods for finding the minimum energy reaction path and alternative reaction paths for small gas-phase reactions have been elucidated with examples, and molecular dynamic simulations have been used to characterise the binding affinity of protein-ligand complex and the free energy of protonation processes in a protein.

Specifically, the fast marching method (FMM) has been used to find the minimum energy path (MEP) on the potential energy surface (PES) for small gas-phase reactions. In this thesis, FMM is shown to be one of the most general and reliable surface-walking algorithms for finding the MEP. However, it is an expensive method. Some improvements have been illustrated in chapter 2 and chapter 3.

I also proposed a new method (called QSM-NT) for finding all stationary points, accordingly all alternative reaction paths on the PES. Unlike other path-finding methods,

QSM-NT overcomes the need of an initial guess of the path, and it can find all stationary points on the PES. QSM-NT has been proven to be efficient and reliable through applications on analytical PES and real chemical reaction. The difficulties and pitfalls associated with QSM-NT have been elucidated with examples.

Molecular dynamic (MD) simulation and associated postprocessing procedures have been used to study the binding properties of caffeine-A_{2A} complex. The binding affinities of different binding modes have been calculated using MM/PBSA method. The binding pocket has been characterised with MM/GBSA energy decomposition. Our computational work provides significant insight to the targeted drug design of the adenosine A_{2A} receptor.

The pH-dependent properties of a protein play important roles in the fundamental biological processes. The protonation states, namely, the pK_a values of ionisable residues, especially active-site residues are the prerequisites to understanding of the mechanisms of many biological processes. In this thesis, acetoacetate decarboxylase (AADase) is used as a test case for studying different types of pK_a prediction methods. Our computational results have shown that the site-site interactions from other ionisable residues are crucial to the pK_a prediction of the target residue.

This thesis covers the range from small gas phase reaction prediction to large complex biological systems characterisation using quantum mechanical and molecular mechanical methods.

To my husband, Jian (Jeffrey) Li,

my sons Ethan Chufeng Li and Nathan Chuyang Li

my parents, Guanggan Liu, Kailian Weng,

and my brother, Zhiqian Liu.

For their endless love, support and encouragement

ACKNOWLEDGEMENTS

First I would like to thank McMaster University and the department of chemistry and chemical biology for giving me the opportunity to pursuing my degree. Particularly I would like to thank Carol Dada and Tammy Feher for their guide and help when I first came to Canada and started graduate study in this country.

I would like to thank my advisor, Professor Paul W. Ayers for accepting me into his group, for his encouragement in times of difficulty, his guidance and inspiration along my academic journey. Paul is always there when we need him. He is a role model for his students. There is an old Chinese saying: Once a teacher, forever a mentor. Paul is the best advisor one could have, and he is my lifetime mentor.

I would like to thank Dr. Dumont and Dr. Bain who served on my Ph.D. committee. I am grateful for their guide, help and constructive suggestions on my research projects.

I would like to thank the Ayers' group. I would like to thank the postdoctoral fellows Dr. Bijoy Dey, Dr. David Thompson, Dr. Utpal Sarkar, Dr. Carlos Cardenas, Dr. Lourdes Romero, Dr. Alfredo Guevara, Dr. Peter Limacher and Dr. Steven Burger. Particularly I would like to thank Dr. Bijoy Dey and Dr. Steven Burger for their help on my research projects. I have learned so much from them. I would like to thank the graduate students Dr. Juan Rodriguez, Dr. James Anderson, Ivan Vinogradov, Rogelio Cuevas-Saavedra,

Debajit Chakraborty, Sandra Rabi, Paul Johnson, Ahmed Kamel, Farnaz Heidarzadeh, and Pavel Kulikov. Thank all of them for being such good friends and a warm family.

I would like to thank my parents and my brother for their unfailing faith in me and for always being there for me. Their continuous support and love help me succeed each step of the way.

Last but not the least, I would like to express my gratitude to my husband, Jian (Jeffrey) Li for his love, support, and encouragement.

Financial supports from Natural Sciences and Engineering Research Council (NSERC) of Canada, Canada Research Chairs, Ontario government and McMaster University are acknowledged.

Table of Content

ABSTRACT ii
THESIS DEDICATION..... iv
ACKNOWLEDGEMENT..... v
LIST OF FIGURES..... xi
LIST OF TABLES..... xii
PREFACE..... xiv

Chapter 1: Background

1.1 Introduction 2
1.2 The Potential Energy Surface and the Born-Oppenheimer Approximation 3
1.3 Minimum Energy Path 8
1.3.1 The Fast Marching Method 9
1.3.2 The QSM-NT Method 10
1.4 Molecular Mechanics (MM) 12
1.4.1 Potential Energy Functionals and Force Fields 13
1.4.2 From Microscopic to Macroscopic 15
1.4.3 Molecular Dynamics (MD) Simulation 16
1.5 Free Energy Calculation 19
1.5.1 Thermodynamic Integration 21
1.5.2 pK_a Calculation 22
1.5.3 MM/PBSA and MM/GBSA Binding Energy Calculation 22
1.6 Summary of Ph.D. Work 23
Reference List 27

Chapter 2: The Fast Marching Method for Determining Chemical Reaction Mechanisms in Complex Systems

2.1 Statement of the Problem 32
2.2 Motivation 32
2.3 Background 33
2.3.1 The Minimum Energy Path 33
2.3.2 Two End Methods 34
2.3.3 Surface Walking Algorithms 36
2.3.4 Metadynamics Methods 38
2.3.5 The Fast Marching Method 39
2.4 The Fast Marching Method 39
2.4.1 Introduction to FMM 40

2.4.2 Upwind Difference Approximation 41
2.4.3 Heapsort Technique 42
2.4.4 Shepard Interpolation 44
2.4.5 Interpolating Moving Least Square Method 47
2.4.6 FMM program 49
2.4.7 Application 52
2.5 Quantum Mechanics/Molecular Mechanics (QM/MM) methods applied to Enzyme-catalyzed reactions 63
2.5.1 QM/MM Methods 63
2.5.2 Incorporating the QM/MM-MFEP Methods with FMM 67
2.5.3 Application of the incorporated FMM and QM/MM-MFEP method to enzyme-catalyzed reactions 69
2.6 Summary 71
Reference List 72

Chapter 3: Finding Minimum Energy Reaction Paths on Ab Initio Potential Energy Surfaces Using the Fast Marching Method

3.1 Statement of the Problem 76
3.2 Introduction 76
3.3 The Fast-Marching Method 78
3.4 Applications 83
3.4.1 The S_N2 reaction 83
3.4.2 The isomerization of HSCN to HNCS 88
3.4.3 The dissociation of ionized O-methylhydroxylamine 91
3.5 Conclusion 95
Reference List: 97

Chapter 4: Newton Trajectories for Finding Stationary Points on Molecular Potential Energy Surfaces

4.1 Statement of the Problem 101
4.2 Introduction 101
4.3 Background 102
4.3.1 Methods for Finding the Minimum Energy Path (MEP) 102
4.3.2 The Newton Trajectory (NT) 104
4.4 Mathematical Definitions and Algorithms 106
4.4.1 Quadratic String Method (QSM) 106
4.4.2 Newton Trajectories (NTs) 109
4.5 Applications 110
4.5.1 Müller-Brown PES 110
4.5.2 The 4-well PES 113
4.5.3 The S_N2 reaction 115

4.6 Difficulties **118**
4.6.1 Discontinuous trajectories 118
4.6.2 Multiple minima of $\|\mathbf{g}_{\perp}\|$ on the hypersurface 121
4.7 Conclusion **123**
Appendix 1. The Müller-Brown Potential **125**
Appendix 2. The 4-well Potential **126**
Reference List **127**

Chapter 5: Computational study of the binding modes of caffeine to the adenosine A_{2A} receptor

5.1 Statement of Problem **131**
5.2 Introduction **132**
5.3 Computational Methods **135**
5.3.1 Docking 135
5.3.2 Molecular Dynamics 137
5.3.3 MM/PBSA Binding Energy Calculation 140
5.3.4 Residue-wise MM/GBSA energy decomposition 142
5.4 Results and Discussion **142**
5.4.1 Binding Modes 142
5.4.2 Relative Binding Energy 146
5.4.3 MM/GBSA energy decomposition results 152
5.4.4 Comparison with site-directed mutagenesis studies 155
5.5 Conclusion **162**
Supporting Information **163**
Reference List **165**

Chapter 6: pK_a calculation of Lys115 in Acetoacetate Decarboxylase

6.1 Statement of Problem **173**
6.2 Introduction **175**
6.3 Methods **179**
6.3.1 Molecular dynamics/thermodynamic integration (MD/TI) for calculating pK_a shifts 179
6.3.1.1 The calculation of pK_a shifts and free energy differences 179
6.3.1.2 Thermodynamic integration (TI) for calculating the deprotonation free energies 182
6.3.1.3 Thermodynamic integration with Glu76 unprotonated 185
6.3.1.4 Thermodynamic integration with Glu76 protonated 186
6.3.2 pK_a calculation using MCCE 187
6.3.3 pK_a calculation using PROPKA 2.0 189
6.4 Results and Discussion **189**

6.4.1 Results from MD/TI pK_a calculations 190
6.4.1.1 With Glu76 unprotonated 190
6.4.1.2 With Glu76 protonated 190
6.4.2 Results from MCCE calculations..... 191
6.4.3 Results from PROPKA calculations 192
6.5 Comparison and discussion of results 194
6.6 Conclusion 198
Reference List 200

Chapter 7: Summary and Future Work

7.1 Summary 205
7.2 Future Work 207

Appendix

List of Abbreviations 210

List of Figures

Chapter 1: Background

Figure 1. 1: The 4-well analytical potential energy surface..... 7

Chapter 2: The Fast Marching Method for Determining Chemical Reaction Mechanisms in Complex Systems

Figure 2. 1: A binary Min-heap. 43
Figure 2. 2: The MEP on the 4-well PES..... 55
Figure 2. 3: The MEP on the energy-cost surface transformed from the 4-well PES. 56
Figure 2. 4: The PES of the S_N2 reaction..... 58
Figure 2. 5: The MEP on the energy-cost surface of the S_N2 reaction. 59
Figure 2. 6: The energy profile of the S_N2 reaction. 60
Figure 2. 7: The 3-D equipotential surface. 62
Figure 2. 8: The energy profile of the dissociation of the ionized O-methylhydroxylamine. 63
Figure 2. 9: The QM subsystem..... 65

Chapter 3: Finding Minimum Energy Reaction Paths on Ab Initio Potential Energy Surfaces Using the Fast Marching Method

Figure 3. 1: The PES of the S_N2 reaction..... 85
Figure 3. 2: The energy-cost surface of the S_N2 reaction. 86
Figure 3. 3: The energy profile of the S_N2 reaction 87
Figure 3. 4: The MEP on the PES of the isomerization of HSCN to HNCS. 89
Figure 3. 5: The energy profile for the isomerization of HSCN to HNCS. 90
Figure 3. 6: The 3-D equipotential surface 93
Figure 3. 7: The energy profile of the dissociation of the ionized O-methylhydroxylamine. 94

Chapter 4: Newton Trajectories for Finding Stationary Points on Molecular Potential Energy Surfaces

Figure 4. 1: Newton trajectories on the Müller–Brown potential. 112
Figure 4. 2: Newton trajectories on the 4-well potential. 114
Figure 4. 3: Newton trajectories for the S_N2 reaction. 117

Figure 4. 4: Discontinuous Newton trajectories on the 4-well potential. 120
Figure 4. 5: Multiple minima on the hypersufaces. 122

Chapter 5: Computational study of the binding modes of caffeine to the adenosine A_{2A} receptor

Figure 5. 1: The molecular structure and atom numbering of caffeine. 136
Figure 5. 2: The molecular structure and atom numbering of ZM241385. 136
Figure 5. 3: RMSD of the backbond atoms of the adenosine A_{2A} receptor. 143
Figure 5. 4: Energy contribution of pocket residues. 153
Figure 5. 5: A superimposition of the 4 low-energy binding modes of caffeine and the dominant binding modes of ZM241385. 154
Figure 5. 6: Caffeine binding cavity, side view. 159
Figure 5. 7: Caffeine binding cavity, extracellular view. 160
Figure 5. 8: Interactions between caffeine and the pocket residues. 161

Chapter 6: pK_a calculation of Lys115 in Acetoacetate Decarboxylase

Figure 6. 1: The thermodynamic cycle. 181
Figure 6. 2: The model compound. 181
Figure 6. 3: Mutation of charges on the lysine group during thermodynamic integration. 184

List of Tables

Chapter 2: The Fast Marching Method for Determining Chemical Reaction Mechanisms in Complex Systems

Table 2. 1: Parameters for the 4-well analytical PES. 53

Chapter 4: Newton Trajectories for Finding Stationary Points on Molecular Potential Energy Surfaces

Table 4. 1: Parameters for the Müller–Brown potential. 125
Table 4. 2: Parameters for the 4-well potential. 126

Chapter 5: Computational study of the binding modes of caffeine to the adenosine A_{2A} receptor

Table 5. 1: MM/PBSA relative bindnig energy calculation results. 147
Table 5. 2: The overall binding energy of caffeine. 147
Table 5. 3: Comparison of pocket residues in different binding modes. 151
Table 5. 4: Comparison of MM/GBSA results with site-directed mutagenesis results .. 157
Table 5. 5: PCA calculation results..... 164

Chapter 6: pK_a calculation of Lys115 in Acetoacetate Decarboxylase

Table 6. 1: MD/TI results with Glu76 unprotonated. 190
Table 6. 2: MD/TI results with Glu76 protonated. 191
Table 6. 3: MCCE calculation results. 192
Table 6. 4: PROPKA calculation results..... 193
Table 6. 5: Comparison of results from different methods. 196

PREFACE

This thesis contains some published and unpublished contents. The co-authors of each chapter have been listed in the footnotes. In this section I will specifically clarify my contributions to each chapter.

My thesis consists of an introduction, five journal articles and a summary at the end. The introduction, chapter 1, provides background information and motivation for the projects discussed in the following chapters. Chapter 1 also provides an overview of the thesis, with emphasis on the perspective that motivated this research. Chapter 2-6 of my thesis are reprints of the published articles or manuscripts under review or in preparation. The first section of each chapter is a statement of the problem, which explains the purpose, motivation and results of the research in this chapter and how it fits into the context of this thesis. Chapter 7 summarizes the thesis and suggests directions for future research.

Chapter 2 is a reprint of the book chapter "The Fast Marching Method for Determining Chemical Reaction Mechanisms in Complex Systems" published in Quantum Biochemistry. (Yuli Liu; Steven K.Burger; Bijoy K.Dey; Utpal Sarkar; Marek R.Janicki; Paul W.Ayers; In Quantum Biochemistry, Cherif F.Matta, Ed.; Wiley-VCH: 2010; pp 171-195.) I am the first author of this book chapter. I did the majority of the programming work of the FMM program and performed all the computations. Prof. Paul

W. Ayers did the derivation of Shepard interpolation. Dr. Utpal Sarkar did part of the programming work for Shepard interpolation. Dr. Steven K. Burger did most of the programming work for Shepard interpolation. I wrote the first draft of this book chapter and Prof. Paul W. Ayers revised the draft. And then we discussed the revisions until we both agreed on this final version.

Chapter 3 is a reprint of the article "Finding Minimum Energy Reaction Paths on Ab Initio Potential Energy Surfaces Using the Fast Marching Method", accepted for publication by the Journal of Mathematical Chemistry. I am the first author and Prof. Paul W. Ayers is the co-author. I did all the computational work and wrote the first draft. Prof. Paul W. Ayers modify the draft to the final version.

Chapter 4 is a reprint of the article "Newton Trajectories for Finding Stationary Points on Molecular Potential Energy Surfaces", accepted for publication in the Journal of Mathematical Chemistry. I am the first author and Dr. Steven K. Burger and Prof. Paul W. Ayers are the co-authors. Dr. Steven K. Burger wrote the QSM program, and I revised it into the QSM-NT program. I did all the computational work for this article. I wrote the first draft and Prof. Paul W. Ayers modify it to the final version.

Chapter 5 is a reprint of the article "Computational study of the binding modes of caffeine to the adenosine A_{2A} receptor", submitted to J. Phys. Chem. B., under review at the moment. I am the first author, and Dr. Steven K. Burger, Dr. Esteban Vöhringer-Martinez, and Prof. Paul W. Ayers are the co-authors. I did the majority of the computational work, except that the PCA entropy calculations were performed by Dr. Esteban Vöhringer-Martinez. I wrote the first draft and Dr. Steven K. Burger revised the

draft. Then Prof. Paul W. Ayers and Dr. Estaban Vöhringer-Martinez refined the draft until we all agreed on the final version.

Chapter 6 is a reprint of the manuscript "pK_a calculation of Lys115 in Acetoacetate Decarboxylase", which is still in preparation. I did all the computational work. Dr. Steven K. Burger did auxiliary calculations with MEAD (not included in this chapter). I wrote the first draft of the chapter and Prof. Paul W. Ayers proofread the draft.

The author of this thesis did the majority of the programming, computational and writing work of all content in this thesis, with help from co-authors, such as discussion, proofreading, some calculations, etc.. Chapter 2-4 were principally guided by Prof. Paul W. Ayers, and chapter 5-6 were jointly guided by Prof. Paul W. Ayers and Dr. Steven K. Burger.

Chapter 1:

Background

1.1 Introduction

The ultimate goal of computational chemistry is to model chemical reactions using computers. Suppose that we are given a set of molecules: instead of mixing them in beakers, we could load the information into a computer and, through computer simulation, predict what would happen and explain how it would happen. While computational chemists have made great progress in this quest, it is still far from being fully realized. This thesis features theoretical developments and computational studies that advance our ability to simulate chemical reactions.

There are many ways to predict what happens in a chemical reaction. Some methods focus on a certain property of the reactants or a certain type of reactions; we call these "specific methods." For example, reactivity indicators provide a straightforward way to predict which site of a molecule will be attacked by a specific type of reagent,[1,2] or predict a molecule's susceptibility to a specific type of reaction (e.g., by predicting the quality of a leaving group).[3] General-purpose methods are designed to work for all possible types of chemical processes and typically use the potential energy surface or free energy surface of the molecular system. For gas phase reactions, the potential energy surface (PES) provides important information about the reaction: the energy minima represent the reactant, the product and potential reactive intermediates; the 1st-order saddle points represent the transition states linking these stable structures. Finding the stationary points[4] on the PES gives us detailed information about the chemical reaction

mechanism(s). In most chemical systems, one reaction mechanism is dominant, and knowing the unique minimum energy reaction path linking the reactant(s) and product(s)[5] provides sufficient information to characterize the thermodynamics and kinetics of the chemical system. For gas-phase reactions at sufficiently high temperatures, condensed-phase reactions, and the reactions of complex biological systems, the potential energy surface cannot represent the system's behaviour because the molecule fluctuates and statistically samples a range of different structures. Statistical sampling is required to achieve the thermodynamic properties of a macroscopic system, i.e., the free energy difference between two states. The same principles apply here, but the mechanism should be characterized using free energy surfaces instead of potential energy surfaces.

In this thesis I will present my work on methods for finding minimum energy reaction paths on PES for gas phase reactions and determining free energy differences in complex biological systems using molecular dynamics. The remaining sections of this chapter describe the quantum mechanics (QM) and molecular mechanics (MM) tools and concepts used in subsequent chapters. The motivation, significance, and fundamental ideas behind the various facets of my thesis project are also discussed.

1.2 The Potential Energy Surface and the Born-Oppenheimer Approximation

The molecular potential energy surface (PES) is fundamental to reaction mechanism studies. It represents the electronic energy of a molecular system as a function of all the

relevant atomic positions. The Born-Oppenheimer approximation, or another similar adiabatic approximation, is a precondition for defining the PES.[6]

The Schrödinger equation for a molecule with n electrons, and N nuclei, is

$$\left[-\sum_{\alpha=1}^{N} \frac{\hbar^2}{2M_\alpha} \nabla_\alpha^2 - \sum_{i=1}^{n} \frac{\hbar^2}{2m_e} \nabla_i^2 + \frac{e^2}{4\pi\varepsilon_0} \left(-\sum_{\text{all } i,\alpha} \frac{Z_\alpha}{r_{i\alpha}} + \sum_{i<j} \frac{1}{r_{ij}} + \sum_{\alpha<\beta} \frac{Z_\alpha Z_\beta}{r_{\alpha\beta}} \right) \right] \psi(\mathbf{x},\mathbf{X}) = E_{total} \psi(\mathbf{x},\mathbf{X}) \tag{1.1}$$

Here **x** and **X** represent the electronic and nuclear coordinates; M_α and Z_α are the mass and the atomic number (nuclear charge) of the α^{th} nucleus; m_e is the mass of an electron; e is the charge on a proton and the magnitude of the charge on an electron; $r_{i\alpha}$, r_{ij} and $r_{\alpha\beta}$ represent electron-nucleus, electron-electron and nucleus-nucleus distances.

Since nuclei are thousands of times heavier than electrons (and the rest mass of an electron is approximately 1836 times smaller than that of the proton), they move much more slowly than electrons. Born and Oppenheimer proposed that electrons can be pictured as moving in the field of fixed nuclei. When the nuclei move, the electron density should adjust almost instantaneously. Thus the electronic motion (described by the electron wave function $\psi_e(\mathbf{x};\mathbf{X})$) can be separated from the nuclear motion (described by the nuclear wave function $\psi_n(\mathbf{X})$),

$$\psi(\mathbf{x},\mathbf{X}) = \psi_e(\mathbf{x};\mathbf{X})\psi_n(\mathbf{X}), \tag{1.2}$$

where $\psi_e(\mathbf{x};\mathbf{X})$ is a solution of the Schrödinger equation involving the electronic Hamiltonian or Hamiltonian describing the motion of n electrons in the field of N fixed nuclei (N point charges),

$$\hat{H}_e\psi_e(\mathbf{x};\mathbf{X}) = V_e(\mathbf{X})\psi_e(\mathbf{x};\mathbf{X}). \tag{1.3}$$

$\psi_e(\mathbf{x};\mathbf{X})$ is a function of the electronic coordinates $\mathbf{x}$ and only depends on the nuclear coordinates $\mathbf{X}$ parametrically, since it is solved for a particular choice of nuclear positions.

Since nuclei move much slower than electrons, the nuclear kinetic energy term $\left(-\sum_{\alpha=1}^{N}\frac{\mathsf{h}^2}{2M_\alpha}\nabla_\alpha^2\right)$ in equation (1.1) is often neglected. The nuclear potential energy term $\left(\frac{e^2}{4\pi\varepsilon_0}\sum_{\alpha<\beta}\frac{Z_\alpha Z_\beta}{r_{\alpha\beta}}\right)$ does not depend on the electronic positions. Therefore the electronic Hamiltonian is defined as,

$$\hat{\mathrm{H}}_e = -\sum_{i=1}^{n}\frac{\mathsf{h}^2}{2M_e}\nabla_i^2 + \frac{e^2}{4\pi\varepsilon_0}\left(-\sum_{\text{all } i,\alpha}\frac{Z_\alpha}{r_{i\alpha}} + \sum_{i<j}\frac{1}{r_{ij}}\right). \tag{1.4}$$

The potential energy $V_e(\mathbf{X})$ of a particular nuclear configuration $\mathbf{X}$ is determined by the total electronic energy associated with that nuclear configuration and can be obtained by solving the Schrödinger equation (1.3) involving the electronic Hamiltonian.

Adding the nuclear-nuclear repulsion term to $V_e(\mathbf{X})$ defines the potential energy surface, which is usually denoted as $V(\mathbf{X})$.

The potential energy surface $V(\mathbf{X})$ of a molecular system contains important information on its geometries and the relative energies of its locally stable structures, as well as the most favourable reaction pathways between these structures. Figure 1.1 shows the 3-D surface plot and contour plot of a 4-well 2-D potential energy surface. The bottoms (green, yellow to orange) of the 4 wells represent 4 energy minima (reactant, product or intermediates), and the first order saddle points represent the transition states between each pair of minima. This surface is used throughout this thesis as a useful test-case for algorithms (See section 2.4.7 and 4.5.2 for description of 4-well PES).

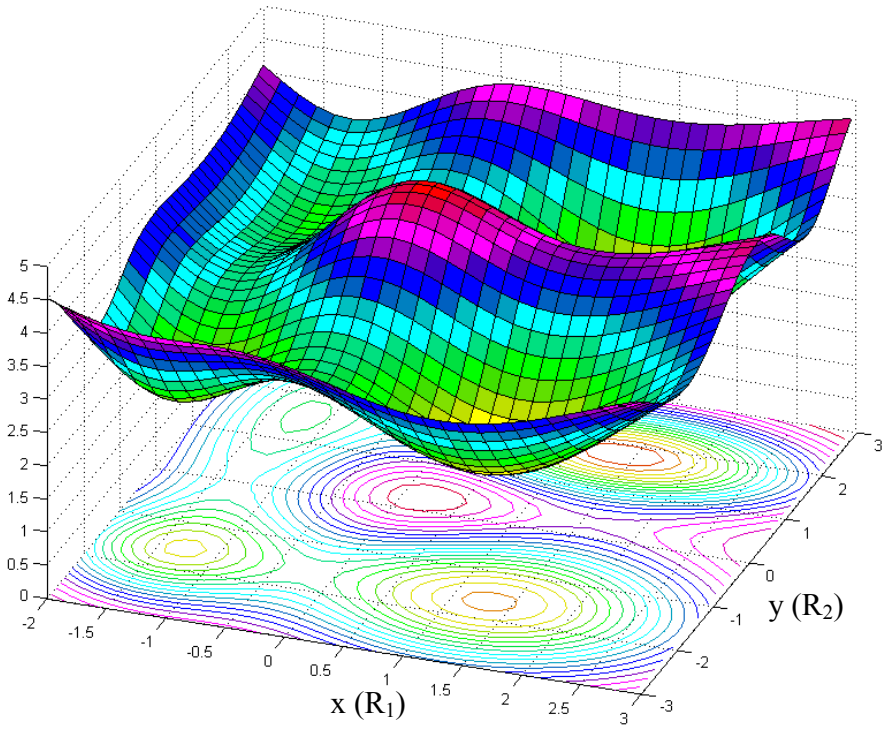


Figure 1. 1: The surface plot and contour plot of a 4-well 2-dimensional potential energy surface.

1.3 Minimum Energy Path

The most widely accepted definition of minimum energy path (MEP) is the intrinsic reaction coordinate (IRC) proposed by Fukui;[7] the IRC is the steepest descent path (SDP) from the first order saddle point down to the adjacent minima on the PES. In this thesis, the MEP is defined as SDP. For multi-step reactions, the overall minimum energy path would be composed by the linking the MEPs of the individual steps, in sequence. When multiple mechanisms (alternative reaction paths) exist, the MEP with the lowest overall energy barrier is the global MEP, others are local MEPs. The MEP provides critical information about chemical reactions, including information about the mechanism, the reaction rate, etc.. Thermodynamic properties like the heat of reaction and the equilibrium constant[8] can also be derived from the MEP. Unsurprisingly, theoretical chemists have exerted great effort towards finding MEPs.

There are two families of algorithms for finding the MEP: the surface walking[9-11] algorithms (an "initial value" formulation) and the two end algorithms[12-17] (a "boundary value" formulation). The two end methods usually require a good guess for the path linking the reactant and product. Only the local MEP can be found using two end methods. By contrast, surface walking methods only need the reactant configuration, and then explore the PES to predict the products and the mechanism along the way. Unfortunately, surface walking algorithms usually are either very expensive or, if a

heuristic is used to simplify the computation, they tend to be unreliable[18] for complicated systems.

One of the main contributions of this thesis is the development of a new surface-walking algorithm, the fast marching method (FMM), for finding the global minimum energy reaction path. We also developed a new two-end algorithm, the QSM-NT method, that locates all the stationary points on a potential energy surface, which allows us to find several alternative minimum energy paths (and helps reduce the need for good guesses of the reaction path). Chapter 2 is a comprehensive review of path-finding methods.

1.3.1 The Fast Marching Method

Fast marching methods are numerical schemes that solve the eikonal equation. The Fast Marching Method (FMM) for determining MEP transforms a multiple-well PES ($V(\mathbf{R})$) to a single-well energy cost surface ($U(\mathbf{R})$) by solving the eikonal equation that defines the cost of traveling from the initial configuration ($\mathbf{R}_0$, the reactant) to another ($\mathbf{R}$) on the PES,[5,11,11]

$$\left|\nabla U_n(\mathbf{R})\right| = \left\{\sqrt{2(E - V(\mathbf{R}))}\right\}^n, \tag{1.5}$$

with the boundary condition of $U(\mathbf{R}_0) = 0$. n is an integer.

This eikonal equation describes wavefront propagation with the local speed function $\frac{1}{\left\{\sqrt{2(E-V(\mathbf{R}))}\right\}^{n}}$. Pictorially, we imagine flooding the PES, starting from the reactant "valley". The "water" level rises until it breaches the lowest-energy "mountain pass" (the transition state) and then races to the bottom of the next "valley" (the intermediate) along the steepest descent path. The "valley flooding" process continues until the product is found. The contour lines that show what portion of the surface was underwater at a given point in time define a new single-well energy cost surface. The steepest descent path from the product to the reactant is the MEP; it is constructed by a process called backtracing.

Since the "water" level will always go to the next "valley" through the lowest energy "mountain pass", FMM can assure that the minima (bottom of valleys) are linked by the lowest energy transition states. Therefore the global minimum energy path is found.

1.3.2 The QSM-NT Method

The string method is a very popular path-finding method.[19] It is a two-end method. The string method divides the initial path into several nodes, which are connected by strings to define the path. The nodes are driven to the steepest descent path (SDP) by a normal force orthogonal to the tangent of the path. The tangent to the path is updated at

each iteration and the nodes are redistributed (to maintain equal spacing between nodes) until a good approximation to the SDP is found.

The string method algorithm can be conceptually described as dropping an elastic pearl necklace on the PES, with the two ends of the necklace fixed on the reactant and the product. The pearls roll down from their initial position until the necklace settles into a local MEP.

The quadratic string method (QSM)[20] uses the same algorithmic structure as the string method, except that a local quadratic approximation of the PES is used; this reduces the number of energy and gradient calculation.

The Newton trajectory (NT) is an alternative reaction path that has been proposed by Quapp and his coworkers.[21,22] A Newton trajectory is a curve on which all gradients are pointing in the same (or opposite) direction, called the searching direction of the NT. Since the magnitude of the gradient at the stationary point is zero, its direction is arbitrary. Therefore a Newton trajectory passes all stationary points on the PES. If carefully chosen, a continuous NT without any turning points or higher-order saddle points can be found; Quapp proposes that this is a good model for the reaction path. The problem is that an NT can contain spurious turning points (non-stationary point), 2^{nd}-order or higher-order saddle points (stationary points), energy maxima (stationary points). All these points could be maxima on the energy profile of the NT and appear to be transition states, which might give a misleading reaction path. Without prior knowledge of the PES, it seems

difficult to find a searching direction that defines a NT in which all turning points are minima or 1st-order saddle points.

To avoid the "turning point problem" associated with using a single NT as the reaction path, we proposed a new method for finding the stationary points on the PES by locating the intersections of two or more NTs. Since a NT passes all stationary points on the PES, the NTs intersect at stationary points; after finding the stationary points we can determine the possible reaction pathways. We adapted the QSM algorithm to find NTs. This new method is called QSM-NT; it is discussed in chapter 4.

1.4 Molecular Mechanics (MM)

Macroscopic systems contain an enormous number of interacting particles. For gas phase reactions, molecules are so dilute and far apart that computational chemists usually consider only one set of reactant molecules for an *ab initio* potential energy calculation. For condensed phase systems, such as reaction in solution or the reactions of macromolecules, the interactions between molecules is too strong to be ignored, and thousands, or even millions, of atoms must be modelled. It is difficult to describe the evolution of such large and complex system in a deterministic way. Instead, statistical sampling is used to study the systems' average behaviour. Just as in path-finding methods for gas-phase reactions, the potential energy is the basic input, but now it is used for statistical sampling. In molecular mechanics, the potential energy of the molecular system

is modelled by classical mechanics, wherein, the atoms and bonds are considered as charged (and perhaps polarisable) balls and springs, respectively. The energy function depends on force constants (to describe the springs' strength) and the displacement from equilibrium.

1.4.1 Potential Energy Functionals and Force Fields

In molecular mechanics the potential energy of a system is calculated using force fields, which include the form of the potential energy function and the values of its associated parameters. The forms of the potential energy functions differ between different force fields, but the general form can be described using the following formulae,[23-25]

$$\begin{aligned} V &= V_{bonded} + V_{non-bonded} \\ V_{bonded} &= V_{bond} + V_{angle} + V_{dihedral} \\ V_{non-bonded} &= V_{van\ der\ Waals} + V_{electrostatic} \end{aligned} \quad . \tag{1.6}$$

The bond and angle terms are usually modelled as harmonic oscillators. The torsion is periodic, so the dihedral or torsional terms are modelled by periodic functions, e.g., a Fourier series. The van der Waals terms are typically modelled using a 6-12 Lennard-Jones potential. The electrostatic terms are modelled using the Coulomb interaction, with set atomic charges for different types of atoms in the molecule. For example, the potential energy functional for the AMBER force field[26] is in the following form,

$$
\begin{aligned}
V &= V_{bond} + V_{angle} + V_{dihedral} + V_{van\ der\ Waals} + V_{electrostatic} \\
&= \sum_{bonds} \frac{1}{2} k_b \left(b - b^0\right)^2 + \sum_{angles} \frac{1}{2} k_\theta \left(\theta - \theta^0\right)^2 + \sum_{torsions} \frac{1}{2} V_n \left[1 + \cos\left(n\omega - \gamma\right)\right] \\
&\quad + \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \varepsilon_{i,j} \left[\left(\frac{r_{ij}^0}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^0}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right\}
\end{aligned}
\tag{1.7}
$$

Here k_b is the stretching force constant and b^0 is the equilibrium bond length; k_θ is the bending force constant and θ^0 is the equilibrium bond angle; V_n is the height of the torsional barrier, n is its periodicity (the number of maxima per full revolution), ω is the torsional angle value, and γ is the phase angle (which is usually 0 or π depending on the periodicity: $\gamma = 0$ if n is odd, and $\gamma = \pi$ if n is even); $\varepsilon_{i,j}$ is the depth of the Lennard-Jones potential well and r_{ij}^0 is the distance between atoms i and j when the potential reaches its minimum $-\varepsilon_{i,j}$; q_i and q_j are partial charges assigned to atoms i and j, respectively, and ε_0 is the electric constant. These parameters are defined for each type of atoms, bonded or non-bonded atom pairs, and bonded triplets (angles) or quadruplets (torsions). For macromolecules, their parameters are usually chosen to reproduce experimental measurements and/or reproduce quantum-mechanical calculations on small molecules.

1.4.2 From Microscopic to Macroscopic

Molecular mechanics is around a million times faster than quantum mechanics at computing potential energy surfaces. This means that molecular mechanics can be applied to much larger systems than quantum mechanics. Although the minimum energy structure on the PES represents the most stable structure of the molecule, molecules are not static; they do not stay at the minimum-energy structure, but fluctuate around it. According to Einstein and Stern's expression of zero-point energy in 1913, a molecule preserves an residual vibrational energy of $\frac{1}{2}h\nu$ at absolute zero temperature (T= 0 K).[27] All quantum mechanical systems undergo structural fluctuations, even in their ground state. In real life the molecule statistically samples a range of different structures. If the molecule is quite small and rigid in structure, and if the temperature is low, then the fluctuations are usually tightly clustered around the minimum energy structure. In this circumstance it is reasonable to use the minimum energy structure on the PES to model the molecule's structure. For molecules that are large and/or floppy, the idea of a unique molecular structure is inadequate, and the molecule should be modeled as a statistical distribution of the structures on the PES. The properties of such systems are no longer determined by a single state, but by averaging over all possible microstates that satisfy the given constraints that define the thermodynamic system; the molecule is represented by the statistical ensemble comprising these microstates. The macroscopic "thermodynamic" properties of the system are the average properties of the ensemble.[28]

The ensemble average can be obtained by statistical sampling methods such as Monte Carlo simulation and molecular dynamics (MD) simulation. In this thesis we only use MD simulation.

1.4.3 Molecular Dynamics (MD) Simulation

According to the ergodic hypothesis, all accessible microstates are equiprobable over a long period of time. Therefore we can assume that the averaging over the statistical ensemble is equivalent to averaging over the time-evolution of the system. Molecular dynamics uses a force field (either from molecular mechanics (MM/MD) or quantum mechanics (QM/MD)) to determine the physical movements of the atoms in a chemical system in time. The basic idea of molecular dynamics simulation is to predict the evolution of a system over a long period of time and then use the time average to calculate the ensemble average.

Some experimental techniques can measure macromolecular systems at atomic resolution. For example, a scanning electron microscope (SEM) can probe a molecular surface and reveal details to less than 1 nm; the most advanced transmission electron microscope (TEM) can even achieve resolution below 0.5 Å; X-ray crystallography can take “snapshots” of crystal structures; nuclear magnetic resonance (NMR) spectroscopy can probe certain features of molecular motions. But the applications of existing experimental techniques are greatly restrained by sample preparation and sample strength.

Moreover, it is still challenging to access both the static and dynamic structures with atomic resolution in the laboratory. MD simulation provides us the opportunity to peer at the motion of the individual atoms in a complex chemical system in a way that is not yet experimentally feasible.

Although MD simulation provides us a way to "visualize" atomic motions, there are some inherent problems and errors associated with MD simulation.

The first problem is related to the ergodic hypothesis. Although it is very hard to prove ergodicity, it is believed that almost all many-body systems are ergodic. However, complex chemical and biological systems might show "nonergodic behaviour" during MD simulations, meaning that the systems do not properly explore phase space. The causes of "nonergodic behaviour" include: 1) Large systems diffuse so slowly that the volume in phase space explored during the computer simulation is insufficient to estimate ensemble average by the time average. In other words, the simulation time is not long enough to apply ergodic hypothesis. 2) Different volumes of phase space are separated by such high energy barriers that the transitions between these volumes become rare events that occur so infrequently that proper sampling of the phase space can not be achieved. 3) Different volumes of phase space are connected by very narrow regions (so-called "entropy bottlenecks"), hence the transitions between them are rarely sampled.[29] When a system appears "nonergodic" during MD simulation, more advanced techniques like stratification (also called multistage sampling) or importance sampling are required to ensure better exploration of the phase space.

The errors of a MD simulation usually come from the following sources: 1) Discretization of time. In MD simulation of a complex system, the classical equation of motion for the atomic nuclei will be solved numerically. This is performed by discretizing the time with a finite timestep Δt, which is the time length between evaluations of the potential. The force on each atom is held constant during the time-span Δt. The timestep has to be smaller than the fastest vibrational frequency in the system. Usually it is set to be $1fs$ ($10^{-15}s$) or less. By using algorithms such as SHAKE, which constrain the vibrations of the fastest atoms, the timestep can be increased. 2) Errors in force fields. In classical MD (or MM/MD), force fields are used to calculate the potential energy of the system. Force fields are usually derived from experimental studies and quantum mechanical computations on small molecules, so they are not exact. They can be very inaccurate if the chemical character (e.g., bonding pattern) of the molecule changes. 3) Non-bonded cutoffs. In an MD simulation, the most time-consuming part is the evaluation of the potential energy as a function of the atomic position. The non-bonded terms are the most expensive part of an MM force field because there is an energy term from every pair of nonbonded atoms. Therefore in most MD simulations, non-bonded cutoffs are applied to reduce the computational cost. Neglecting the Coulomb and Lennard-Jones terms between atoms separated by more than the cutoff distance increases the efficiency of the calculation, but also introduce errors.

1.5 Free Energy Calculation

As we discussed in Section 1.4.2, the thermodynamic properties of a macroscopic system cannot be represented by a single state and require, instead, approximating the ensemble average. For small-molecule gas-phase reactions, the minimum energy reaction path on the PES can provide a good approximation to the reaction mechanism. But for condensed-phase reactions and large complex systems, there is an entire family of reaction pathways, determined by the free energy behaviour of the system. Since it is often impossible to select a single reaction coordinate that captures the full range of possible reaction mechanisms, in these cases we usually focus not on reaction pathways; instead we settle for computing free-energy differences between key chemical species.

The free energy is usually expressed as the Helmhotz free energy, A, or the Gibbs free energy, G. The Helmholtz free energy is the thermodynamic potential of the canonical (NVT: number of particles, volume, temperature) ensemble. The Gibbs free energy is the thermodynamic potential of the isothermal-isobaric (NPT: number of particles, pressure, temperature) ensemble. They can be expressed in terms of the partition function. For example, the Helmholtz free energy can be expressed as,

$$A(N,V,T) = -k_B T \ln Z(N,V,T), \tag{1.8}$$

where k_B is the Boltzmann constant, the canonical partition function can be written as,

$$Z\left(N,V,T\right) = \sum_{\upsilon} e^{-\frac{E_\upsilon}{k_B T}}, \tag{1.9}$$

if all points υ in the phase space are visited; or

$$Z(N,V,T) = \iint d\mathbf{p}^N d\mathbf{r}^N \exp\left(\frac{-E(\mathbf{p}^N,\mathbf{r}^N)}{k_B T}\right), \tag{1.10}$$

if integrating over classical phase space defined by position $\mathbf{r}^N$ and momentum $\mathbf{p}^N$. The Helmhotz free energy can be expressed as,

$$A = k_B T \ln\frac{1}{Z(N,V,T)} = k_B T \ln\left(\iint d\mathbf{p}^N d\mathbf{r}^N \exp\left(\frac{E(\mathbf{p}^N,\mathbf{r}^N)}{k_B T}\right)\rho\left(\mathbf{p}^N,\mathbf{r}^N\right)\right), \tag{1.11}$$

where $\rho\left(\mathbf{p}^N,\mathbf{r}^N\right) = \dfrac{\exp\left(\dfrac{-E(\mathbf{p}^N,\mathbf{r}^N)}{k_B T}\right)}{Z(N,V,T)}$ is the probability of the state with energy $E\left(\mathbf{p}^N,\mathbf{r}^N\right)$.

If it were possible to visit all points in the phase space, then the partition function could be calculated using equation (1.9). In general an accurate estimation of the partition function is impossible due to inadequate sampling during finite simulation time.[30] Calculating the free energy from direct MD or Monte Carlo simulation is very difficult because MD or Monte Carlo simulations preferentially generate states of low energy and, according to equation (1.11), high energy states can also make significant contribution to the free energy. Fortunately in most cases we are only interested in the free energy differences between two systems or two states of a system. Since free energy is a state function and it does not depend on the path, but only depends on the initial and final states, the calculation of free energy difference can be carried out through a series of mixed states using free energy perturbation theory or thermodynamic integration.

1.5.1 Thermodynamic Integration

In thermodynamic integration, a order parameter λ is defined so that the free energy is a continuous function of λ,[31]

$$\Delta A = \int_0^1 \frac{\partial A(\lambda)}{\partial \lambda} d\lambda = \int_0^1 \left\langle \frac{\partial H(\mathbf{p}^N, \mathbf{r}^N, \lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda . \tag{1.12}$$

When using empirical molecular mechanics force fields, the kinetic-energy portion of the Hamiltonian does not depend on λ so the only contribution is from the potential function $V(\mathbf{r}^N, \lambda)$. In practice, the numerical integration, such as Gaussian quadrature formula, $\Delta A = \sum_i \omega_i \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_i$ is used to approximate the integral in equation (1.12). For each discrete value, λ_i, MD or Monte Carlo simulation is performed to estimate the value of $\left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_i$. The potential function for each λ value can be expressed as a weighted average of the initial ($\lambda = 0$) and final perturbed state ($\lambda = 1$),

$$V(\mathbf{r}^N, \lambda) = (1-\lambda)^k V\left(\mathbf{r}^N, 0\right) + \left[1-(1-\lambda)^k\right] V\left(\mathbf{r}^N, 1\right), \tag{1.13}$$

where $k = 1$ represents linear mixing when no atom appears or disappears in the perturbed state, and $k > 1$ is typically used when dummy atoms are used in the perturbed state.

1.5.2 pK_a Calculation

The pK_a value is proportional to the free energy of deprotonating an ionisable group,

$$pK_a = \frac{\Delta A}{2.3026RT}. \tag{1.14}$$

Therefore, the essence of pK_a calculations is directly related to the calculation of free energy differences. Usually we are interested in the pK_a difference of the free amino acid in solvent and in protein environment, in which case, the pK_a shift (ΔpK_a) is calculated through the difference of the deprotonation free energy between the model compound (e.g., the amino acid in a dipeptide chain) and the amino acid in the protein.

1.5.3 MM/PBSA and MM/GBSA Binding Energy Calculation

The molecular mechanics/Poisson Boltzmann surface area (MM/PBSA) and its complementary molecular mechanics/General Born surface area (MM/GBSA) approach are postprocessing methods to calculate the binding free energy based on the sets of structures from an MD trajectory or a Monte Carlo simulation. The MM/PBSA method combines the molecular mechanical energies with continuum solvent so that the binding free energy between two species (ligand-protein, protein-protein, etc.) can be expressed as,

$$\begin{aligned}\Delta G_{bind} &= \left\langle G_{complex} \right\rangle - \left(\left\langle G_{protein} \right\rangle + \left\langle G_{ligand} \right\rangle \right) \\ &= \Delta E_{MM} + \Delta G_{solv} - T\Delta S\end{aligned}. \tag{1.15}$$

where ΔE_{MM} is the molecular mechanical energy difference between the bound state (complex) and unbound state (receptor and ligand), which can be evaluated based on the structures (called snapshots) taken from an MD trajectory. ΔG_{solv} is the solvation free energy difference between the bound and unbound state. The solvation free energy includes two components: the electrostatic energy for transferring the solute from the vacuum to the solvent, and the non-electrostatic contribution that combines the free energy required to form the cavity and van der Waals terms. In MM/PBSA the electrostatic component of the solvation free energy is calculated by solving the Poisson-Boltzmann equation in the vacuum and the solvent, $\Delta G_{PB} = \frac{1}{2}\sum_i q_i\left(\phi_i^{80} - \phi_i^{1}\right)$.[31] The non-polar contribution is calculated using an empirical solvent accessible surface area formula, $\Delta G_{nonpolar} = \gamma \text{SASA}, \gamma = 0.0072 \text{kcal}\,\text{Å}^{-2}$. ΔS is the entropy change upon binding. The MM/GBSA method is the same as MM/PBSA except that in MM/GBSA the electrostatic component of solvation free energy is calculated using the generalized-Born continuum solvent model.[31]

1.6 Summary of Ph.D. Work

Chapter 2, chapter 3 and chapter 4 of the thesis focus on the development of new methods for finding reaction paths on the PES. Chapter 2 gives a detailed review on the popular path-finding methods, followed by the introduction of the fast marching method

(FMM). In this chapter, the mathematical backgrounds and numerical algorithms used in FMM are discussed in detail. Applications of FMM to both analytical PES and real chemical reactions are also included in this chapter. At the end, the possibility of interfacing FMM with QM/MM methods for finding reaction path of complex chemical reactions is discussed.

Chapter 3 presents the extension of the FMM program for finding MEP so that it can use *Gaussian 03* to calculate the potential energy surface. A conceptual description of FMM and its applications to 2-D and 3-D reduced PESs are given in this chapter. The methods presented in chapter 2 and chapter 3 are examples of the surface-walking path-finding methods.

Chapter 4 presents a new method for finding all the stationary points on the PES: the QSM-NT method. It is a combination of the quadratic string method (QSM) and Newton trajectory (NT), and therefore falls into the category of the two end methods. The idea and numerical structure of this method are presented. The applications of QSM-NT to analytical PESs and real chemical reactions are presented. The advantages and pitfalls of this method are elucidated with examples as well.

In addition to chapter 2, chapter 3 and chapter 4, I also co-authored a paper on numerical algorithms used to improve the efficiency of FMM[32] that is not included in this thesis. In this paper, the moving-least-squares enhanced Shepard interpolation is used to cut down the number of potential energy and gradient evaluations, and consequently reducing the computational cost of FMM.

Except for minimum energy reaction paths, there are other, more specific, ways to predict the qualitative characteristics of a chemical reaction. One example is the development of reactivity indicators for predicting the quality of leaving groups in organic molecules. In another non-thesis publication,[33] we studied 66 different reactivity indicators for predicting the quality of organic leaving groups and tested them against experimental data.

Chapter 5 and chapter 6 focus on MM/MD simulations on complex biological systems. Chapter 5 presents a computational study on the binding modes of caffeine to the adenosine A_{2A} receptor. Using the engineered crystal structure of the adenosine A_{2A} receptor, we docked caffeine to the receptor. 5 ns MD simulations were performed on each selected docking pose in the approximate physiological environment. Then the relative binding energy of each binding mode was determined using MM/PBSA. Finally, the critical residues in the binding pocket were identified using the MM/GBSA energy decomposition. The results of our computational studies bring new insight to targeted drug design for the adenosine A_{2A} receptor.

Chapter 6 presents our attempts to calculate the pK_a shift of Lys115 in acetoacetate decarboxylase (AADase) using different kinds of computational methods: the molecular dynamics/thermodynamic integration (MD/TI) method, a Poisson-Boltzmann equation based method (MCCE), and an empirical method (PROPKA). According to the recent crystal structure of AADase, the large pK_a shift of Lys115 is mainly due to its location in a solvent-inaccessible hydrophobic environment, that is, the desolvation effect. Using the

natural protonation patterns of other ionisable residues, none of the above-mentioned computational methods predict the right protonation state of Lys115. Since these methods do not explicitly sample the protonation patterns of other ionisable residues, we postulate that site-site interaction from other ionisable residues play important roles in the pK_a shift of Lys115. Inspired by the results from site-directed mutagenesis studies, we use protonated Glu76 (neutral) for MD/TI calculation, and the pK_a of Lys115 is calculated as 5.3, which agrees well with the experimental value of 5.9.

The classical statistical models that are used to describe complex biological systems confined in a box of water molecules can be generalized to quantum mechanics, where they can be used to model electrons in atoms, molecules, and electronic devices. In yet another non-thesis publication, we used a statistical method to model the electrons in a cubic quantum dot.[34]

In summary, my Ph.D. work focused on the development of new methods for predicting and characterizing the chemical reaction paths of gas-phase chemical reactions and MM/MD computations of free energy differences in biochemical processes in proteins.

Reference List

1. Anderson, J. S. M.; Melin, J.; Ayers, P. W. Conceptual density-functional theory for general chemical reactions, including those that are neither charge- nor frontier-orbital-controlled. 2. Application to molecules where frontier molecular orbital theory fails. *Journal of Chemical Theory and Computation* **2007,** *3* (2), 375-389.

2. Anderson, J. S. M.; Melin, J.; Ayers, P. W. Conceptual density-functional theory for general chemical reactions, including those that are neither charge- nor frontier-orbital-controlled. 1. Theory and derivation of a general-purpose reactivity indicator. *Journal of Chemical Theory and Computation* **2007,** *3* (2), 358-374.

3. Anderson, J. S. M.; Liu, Y. L.; Thomson, J. W.; Ayers, P. W. Predicting the quality of leaving groups in organic chemistry: Tests against experimental data *Journal of Molecular Structure-Theochem* **2010,** *943* (1-3), 168-177.

4. Yuli Liu; Steven K.Burger; Paul W.Ayers Newton Trajectories for Finding Stationary Points on Molecular Potential Energy Surfaces. *(*accepted by *Journal of Mathematical Chemistry.)* **2011**.

5. Yuli Liu; Paul W.Ayers Finding Minimum Energy Reaction Paths on Ab Initio Potential Energy Surfaces Using the Fast marching Method. *Journal of Mathematical Chemistry* **2011**, 49(7), 1291-1301.

6. David J.Wales The Born-Oppenheimer Approximation and Normal Modes. In *Energy Landscapes*, Cambridge University Press: 2003; pp 119-160.

7. Fukui K. A Formulation of the Reaction Coordinate. *Journal of Physical Chemistry* **1970,** *74*, 4161-4163.

8. Dey, B. K.; Janicki, M. R.; Ayers, P. W. Hamilton-Jacobi equation for the least-action/least-time dynamical path based on fast marching method. *J. Chem. Phys.* **2004,** *121* (14), 6667-6679.

9. Dey, B. K.; Ayers, P. W. A Hamilton-Jacobi type equation for computing minimum potential energy paths. *Molecular Physics* **2006,** *104* (4), 541-558.

10. Irikura, K. K.; Johnson, R. D. Predicting unexpected chemical reactions by isopotential searching. *Journal of Physical Chemistry A* **2000,** *104* (11), 2191-2194.

11. Dey, B. K.; Ayers, P. W. A Hamilton-Jacobi type equation for computing minimum potential energy paths. *Molecular Physics* **2006,** *104* (4), 541-558.

12. Chu, J.-W.; Trout, B. L.; Brooks, B. R. A super-linear minimization scheme for the nudged elastic band method. A super-linear minimization scheme for the nudged elastic band method. **2003**, 119(24), 12708-12717.

13. E Weinan; Ren Weiqing Finite temperature string method for the study of rare events. *J. Chem. Phys.* **2005,** *109* (14), 6688-6693.

14. E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. String method for the study of rare events. *Physical Review B* **2002,** *66* (5), 052301.

15. Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004,** *120* (17), 7877-7886.

16. Trygubenko, S. A.; Wales, D. J. A doubly nudged elastic band method for finding transition states. *J. Chem. Phys.* **2004,** *120* (5), 2082-2094.

17. Xie, L.; Liu, H. Y.; Yang, W. T. Adapting the nudged elastic band method for determining minimum-energy paths of chemical reactions in enzymes. *J. Chem. Phys.* **2004,** *120* (17), 8039-8052.

18. Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. A growing string method for determining transition states: Comparison to the nudged elastic band and string methods. *J. Chem. Phys.* **2004,** *120* (17), 7877-7886.

19. E, W. N.; Ren, W. Q.; Vanden Eijnden, E. String method for the study of rare events. *Physical Review B* **2002,** *66* (5).

20. Burger, S. K.; Yang, W. T. Quadratic string method for determining the minimum-energy path based on multiobjective optimization. *J. Chem. Phys.* **2006,** *124* (5), 054109.

21. Quapp, W. Newton trajectories in the curvilinear metric of internal coordinates. *Journal of Mathematical Chemistry* **2004,** *36* (4), 365-379.

22. Quapp, W.; Hirsch, M.; Heidrich, D. An approach to reaction path branching using valley-ridge inflection points of potential-energy surfaces. *Theoretical Chemistry Accounts* **2004,** *112* (1), 40-51.

23. Andrew R.Leach Empirical Force Field Models: Molecular Mechanics. In *Molecular Modelling: Principles and Applications*, Longman: 1999; pp 131-210.

24. Christopher J.Cramer Molecular Mechanics. In *Essentials of Computational Chemistry: Theories and Models*, Wiley: 2004; pp 17-68.

25. Tamar Schlick Theoretical and Computational Approaches to Biomolecular Structure. In *Molecular Modeling and Simulation: An Interdisciplinary Guide*, Springer: New York, 2002; pp 199-224.

26. Case D.A.; Darden T.A.; Cheatham T.E.; Simmerling C.L.; Wang J.; Duke R.E.; Luo R.; Crowley M.; Walker R.C.; Zhang W.; Merz K.M.; Wang B.; Hayik S.Roitberg A.; Seabra G.; Kolossvary I.; Wong K.F.; Paesani F.; Vanicek J.; Wu X.; Brozell S.R.; Steinbrecher T.; Gohlke H.; Yang L.; Tan C.; Mongan J.; Hornak V.; Cui G.; Mathews D.H.; Seetin M.G.; Sagui C.; Babin V.; Kollman P.A. 2008, *AMBER 10*, University of California, San Francisco.

27. Einstein, A.; Stern, O. Einige Argumente fur die Annahme einer molekularen Agitation beim absoluten Nullpunkt. A*nn. Phys.* 1**913,** 3*45* (3), 551-560.

28. Christopher J.Cramer Simulation of Molecular Ensembles. In E*ssentials of Computational Chemsitry: Theories and Models,* Wiley: 2004; pp 69-104.

29. Christohe Chipot; M.Scott Shell; Andrew Pohorille Introduction. In F*ree Energy Calculations,* Springer: 2007; pp 1-32.

30. Andrew R.Leach Computer Simulation Methods. In M*olecular Modelling: Principles and Applications,* Longman: 1999; pp 261-312.

31. Andrew R.Leach Three Challenges in Molecular Modelling: Free Energies, Solvation and Simulating Reactions. In M*olecular Modelling: Principles and Applications,* Longman: 1996; pp 481-542.

32. Burger, S. K.; Liu, Y.; Sarkar, U.; Ayers, P. W. Moving Least-Squares Enhanced Shepard Interpolation for the Fast Marching and String Methods. J. *Chem. Phys. (accepted)* 2**009,** 1*30* (2), 024103.

33. Anderson, J. S. M.; Liu, Y. L.; Thomson, J. W.; Ayers, P. W. Predicting the quality of leaving groups in organic chemistry: Tests against experimental data J*ournal of Molecular Structure-Theochem* 2**010,** 9*43* (1-3), 168-177.

34. Thompson, D. C.; Liu, Y. L.; Ayers, P. W. A confined noninteracting many electron system: Accurate corrections to a statistical model. P*hys. Lett. A* 2**006,** 3*51* (6), 439-445.

Chapter 2:

The Fast Marching Method for Determining Chemical Reaction Mechanisms in Complex Systems*

* The content of this chapter has been published: Yuli Liu; Steven K.Burger; Bijoy K.Dey; Utpal Sarkar; Marek R.Janicki; Paul W.Ayers The Fast Marching Method for Determining Chemical Reaction Mechanisms in Complex Systems. In *Quantum Biochemistry*, Cherif F.Matta, Ed.; Wiley-VCH: 2010; pp 171-195.

2.1 Statement of the Problem

This chapter is a review of the popular path-finding methods and a conceptual introduction to the fast marching method (FMM). FMM is a reliable but yet expensive path-finding method. The moving-least-square enhanced Shepard interpolation has been applied to reduce the computational cost of FMM. The mathematics and numerical schemes used in FMM have been described in detail in this chapter. The Shepard interpolation improved FMM has been applied to analytical PES, 2-dimensional and 3-dimensional reduced PES of real chemical reactions. The possibility of interfacing FMM with QM/MM program and applying it to complex system is discussed.

2.2 Motivation

Suppose that one is given a set of molecules (the reagents) and their reaction conditions (solvent or gas phase, temperature, etc.). Does a chemical reaction occur? What kind of reaction? What is(are) the product(s)? How and why does the reaction happen? These are the fundamental problems of chemistry.

The theoretical solution to these problems requires finding the chemical reaction pathway. For example, given the minimum energy path (MEP), one can determine molecular structures and energies of the reactants, products, and transition states. The difference in energy between the reactants and products is the reaction energy; the difference in energy between the reactants and the transition state structure is the

activation energy, which is related to the rate of reaction. The MEP provides key information about reaction thermodynamics and kinetics. In addition, tracing the MEP from the reactant, through reactive intermediates and the transition state, to the products gives us the chemical reaction mechanism. The reaction mechanism is the key to understanding how and why a reaction occurs, and it is important for optimizing reaction conditions and designing catalysts.

This chapter will review our recent work on computational algorithms for finding the MEP, with particular emphasis on the fast marching method (FMM). In the second section, we will present more information about MEPs and review alternatives to FMM. Section 3 provides algorithmic details about FMM and some applications to small systems. Section 4 reviews the development of the quantum mechanics/molecular mechanics (QM/MM) methods for studying enzyme-catalyzed reactions and presents the idea and some preliminary work on incorporating FMM with QM/MM methods. Section 5 summarizes our results to date and presents our perspective on future research directions.

2.3 Background

2.3.1 The Minimum Energy Path

The reaction path is usually identified with the steepest descent path linking a transition state structure and its adjacent minima (such as reactant, product and reactive intermediates). When there is more than one steepest descent path, the one with the

lowest energy barrier is MEP. The steepest descent path defines the intrinsic reaction coordinate for a chemical reaction[1].

There are two main families of algorithms for finding the MEP: two end methods[2] and surface walking methods. The two end methods require a good guess for the path linking the reactant and product; if the mechanism in the initial guess is qualitatively correct (i.e., the path threads its way through the correct "mountain passes" on the potential energy surface (PES)), the right MEP will be located. Surface walking methods do not require an initial guess. They start exploring the PES from the reactant configuration, and eventually predict the products and the mechanism of the chemical reaction. Unfortunately, surface walking algorithms are usually either very expensive or, if a heuristic is used to simplify the calculation, they tend to be unreliable for complicated systems. The two end methods have great advantages from the viewpoint of computational cost and numerical stability.

2.3.2 Two End Methods

A simple example of a two end method is the Nudged Elastic Band (NEB) method[3-8]. In this method an initial guess of the path is given which is divided up into a series of beads, with springs in between each bead. The beads are then propagated down the PES. One of the significant improvements of NEB over previous methods is it decouples the problem, by projecting the spring force parallel to the path and the force from the

potential perpendicular to the path. This prevents corner cutting of the path and ensures that the path will eventually converge to the MEP.

String methods are similar to NEB, but they do not use a fictitious force to ensure that the molecular conformations that define the reaction path are well-spaced. In the following paragraphs, we will discuss the original string method of Ren and Vanden-Eijnden[9-11] and two improved string methods: the growing string method (GSM)[12] and the quadratic string method (QSM)[13].

Ren and Vanden-Eijnden proposed a zero-temperature string method[14,15] for finding the MEP on the PES. Like the NEB method, the string method drives the initial path to the MEP by perpendicular forces on the bead. The continuity of the path is ensured by reparameterizing the approximate path at each iteration so that the nodes are spaced evenly along the path.

GSM has the same algorithmic structure as the string method. The difference is that the string grows from two ends of the reaction path (the reactant and product) toward the transition state along an interpolated pathway until the growing ends meet. However, the growth of the string depends on the interpolated pathway, which is determined by the update to all nodes in the previous iteration. The dependence on the previous iteration makes it very difficult to parallelize this method. Furthermore, the growing two ends will not meet unless the original interpolated pathway is a good guess for the MEP.

QSM uses the local quadratic approximation of the PES[16]. Compared with the string method, it is more accurate and it converges faster. QSM applies an adaptive step-size

Runge-Kutta method and accordingly removes the need for the user to decide the step size[17,18]. Formulated as a multi-objective optimization problem, it can be easily parallelized. QSM is considered one of the most efficient two end methods for large reaction systems.

Unfortunately, even the best two-end methods require that one have enough prior knowledge of the PES to guess an accurate initial path. Guessing an initial path is almost impossible when exploring new chemistry, in which case one could use the surface walking algorithms instead.

2.3.3 Surface Walking Algorithms

Surface walking algorithms usually start from a stationary point and search for energy minima and transition state (TS) by walking on the PES. Some popular surface walking methods are the eigenvector following (EF) method[19], the gradient extremal following (GEF) method[20], the reduced gradient following (RGF) method[21], the scaled hypersphere search (SHS) method[22-25], and the fast marching method (FMM)[26-33]. Since walking uphill is much more difficult than downhill, most surface walking algorithms focus on the uphill walking algorithm, and aim at global mapping of the PES. The fundamental problem with walking uphill is deciding which walking direction leads from the minimum to the TS.

The eigenvector following method can locate local minima and first-order saddle points by walking through the PES. Starting from an arbitrary point on the PES, the eigenvector following method locates the stationary points by walking along an eigenvector of the Hessian (second-derivative) matrix. By walking along all $2(3N-6)$ eigenvector directions, the eigenvector following method can potentially find all local minima and saddle points in an N-atom molecular system.

The gradient extremal following method walks uphill and downhill by following the extreme absolute values of the gradient along the potential contours[19]. Gradient extremals are curves that intersect the potential energy isosurfaces, $V(\mathbf{R}) = k$, where the curvature of these contour surfaces is an extremum. Since the curvature of an isosurface at a stationary point is infinity, the gradient extremal curves (e.g., the gradient maximum and gradient minimum) are supposed to cross at the stationary points. So finding the crossing points of gradient extremals will give the stationary points. One problem with the gradient extremal following method is that sometimes gradient extremals also intersect at points other than the stationary points[25].

The idea of the reduced gradient following method comes from the zero gradient criterion for stationary points[21]. Starting from a minimum, RGF finds the set of points whose potential gradients are all aligned to the direction of a chosen coordinate. RGF curves connect stationary points differing in their index by 1 and they intersect at the stationary points. The index of a stationary point is the number of negative eigenvalues of

the Hessian matrix at this point[34]. Examples of searching for saddle points using the crossing points of RGF curves are shown in reference 21. RGF curves have been extended to a more general concept: the Newton Trajectory (NT)[35,36]. The searching direction of NT is not limited to one of the coordinates. It could be any direction. To avoid constructing trajectories that wander around the high energy regions of the PES, Quapp applied the growing string algorithm to find the NT[37-39]. An NT without a turning point can be used as approximation to the reaction path. Unfortunately, because there are infinitely many searching directions, it is sometimes difficult to locate a NT that approximates the reaction path.

The scaled hypersphere searching method[22,23,25] is based on the chemical intuition that energy-lowering interactions distort the potential surface downwards as one moves towards the TS[23]. SHS can walk toward the TS by following the extreme magnitude of anharmonicity from the second-order surface expanded at the starting minimum. The efficiency of SHS method is claimed to be 2(3N-6) energy minimization calculations on each hypersphere, but expensive calculations of the Hessian matrix are required.

2.3.4 Metadynamics Methods

Other energy minima searching approaches, such as the free-energy minima escaping method proposed by Laio and Parrinello[40], and the conformation flooding approach by Grubmuller[41], are based on self-avoiding molecular dynamics trajectories on

the potential energy surface. These trajectories do not pass precisely through the TS and reactive intermediate structures, so they do not provide a satisfactory representation for the reaction path. However, a reactive trajectory from these methods can be used as an initial guess for a two end method.

2.3.5 The Fast Marching Method

FMM is a wavefront propagation method that solves the nonlinear eikonal equation[29-32]. FMM has been successfully applied to find the MEP on the PES. As previously mentioned, uphill walking on the PES is more troublesome than downhill walking. FMM avoids the uphill walking problem and transforms the multi-well PES into a single-well energy cost surface by solving the eikonal equation. The only well on the energy cost surface is the starting point, where the cost is defined to be zero. Then the MEP from any point on the PES to the starting point can be found by a downhill backtracing from this point to the bottom of the energy cost surface. Unlike the two end methods, FMM does not need an initial guess of the path and it always converges to the MEP. If the ending point of the path is not specified, FMM will eventually evaluate the whole PES. Details of the FMM algorithm will be presented in next section.

2.4 The Fast Marching Method

2.4.1 Introduction to FMM

We define the cost function at **R** as the "minimum cost" required to attain this configuration starting from the reactant configuration $\mathbf{R}_0$ [28,33]:

$$U(\mathbf{R}) = \underbrace{\min}_{\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(s)} \int_0^L \left\{\sqrt{2(E - V(\mathbf{C}(s)))}\right\}^n ds \,. \tag{2.1}$$

Here the minimization is over all paths, $\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(s)$, that start at $\mathbf{R}_0$ and end at $\mathbf{R}$, E is the total energy of the system, $V(\mathbf{R})$ is the potential energy surface, and L is the path length. (The variable s parameterizes the path so that $\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(0) = \mathbf{R}_0$ and $\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(L) = \mathbf{R}$.) The path integral problem (2.1), can be conveniently restated as an eikonal equation, namely,

$$\left|\nabla U(\mathbf{R})\right| = \left\{\sqrt{2(E - V(\mathbf{R}))}\right\}^n . \tag{2.2}$$

The energy cost of the reactant is zero by definition ($U(\mathbf{R}_0) = 0$); this is the boundary condition for the eikonal equation.

This eikonal equation describes wavefront propagation with the local speed function $\dfrac{1}{\left\{\sqrt{2(E - V(\mathbf{R}))}\right\}^n}$. To locate the MEP, we need the cost of molecular configurations with higher potential energy to be infinitely larger than the cost of configurations with lower potential energy. (Equivalently, we need for configurations that are lower in energy to be attained infinitely faster than configurations that are higher in energy.) This can be achieved by letting $n \to -\infty$, which ensures that higher energy

paths in equation (2.1) are cut off from the set of paths ($C_{\mathbf{R}_0,\mathbf{R}}$), giving only the MEP. Of course, in computational implementations we will choose n to be a sizeable (but non-infinite) negative number. In practice, results are usually good when $n < -10$.

Solving this eikonal equation transforms a multi-well potential energy surface, $V(\mathbf{R})$, into a conical energy cost surface $U(\mathbf{R})$. The numerical algorithms for solving the eikonal equation will be discussed in the following sections.

2.4.2 Upwind Difference Approximation

We need to solve the eikonal equation using an "upwind" finite difference approximation that preserves the causality of the solutions. To do this, we discretize the eikonal equation as follows:

$$\left(\frac{(U-a_1)^+}{dR(1)}\right)^2+\left(\frac{(U-a_2)^+}{dR(2)}\right)^2+\cdots+\left(\frac{(U-a_d)^+}{dR(d)}\right)^2=\left[2(E-V(\mathbf{R}))\right]^n \qquad (2.3)$$

Here $dR(i)$ is the i^{th} component of the grid size vector $d\mathbf{R}$. a_i is the smaller cost value of point $\mathbf{R}'s$ two neighbouring points in direction i, $a_i=\min(U_{left},U_{right})$. The upwind finite difference approximation defines $(U-a_i)^+=(U-a_i)$ if $U>a_i$ and $(U-a_i)^+=0$ otherwise. (That is, $(U-a_i)^+=\max(0,U-a_i)$.)

The upwind finite difference enforces the causality condition in the fast marching method, which means the cost can only increase while the wavefront moves outward. In

other words, for the point in question, its unknown cost value U has to be greater than the cost value, a_i, of its known neighbouring point; if $a_i > U$, then the cost value of this neighbouring point must not be known either. We cannot use an unknown point, so we discard it by letting $(U - a_i)^+ = 0$. This is the idea behind the upwind finite difference approximation.

Equation (2.3) can be solved in an iterative way. First, sort the a_i's in increasing order. Second, start from $j = 1$ and solve the truncated equation:

$$\left(\frac{(U - a_1)^+}{dR(1)} \right)^2 = \left[2(E - V(\mathbf{R})) \right]^n \tag{2.4}$$

If the solution $U_1 \le a_2$, then $U_1 \le a_3 \le \cdots \le a_d$ and thus $U = U_1$ is also the solution to equation (2.3). If $U_1 > a_2$, then let $j = j + 1$, and continue to solve the truncated equation with 2 terms on the left side. This process is repeated until we find the j^{th} solution $U_j \le a_{j+1}$, $1 \le j \le d$. $U = U_j$ is the solution to equation (2.3)[42].

2.4.3 Heapsort Technique

As the wavefront propagates outward, energy cost values of grids on the wavefront are computed by solving the discretized eikonal equation. After computing the energy cost values of all points on the wavefront, we need to identify the point with minimum

energy cost value (thus with maximum local speed) because this is the point that the wavefront is going to pass next. The heapsort technique is used to sort these values.

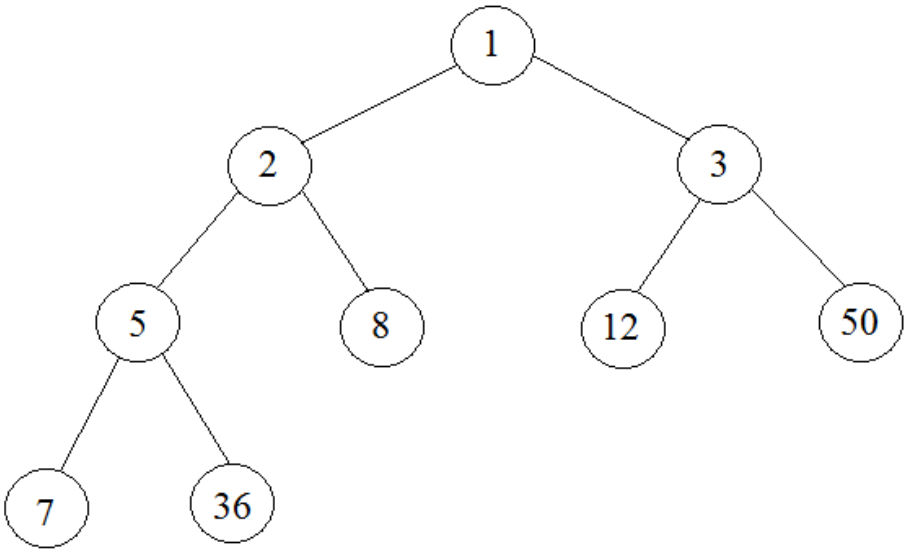


Figure 2. 1: A binary Min-heap

Heapsort is an "in-place" sorting algorithm, requiring no auxiliary storage[43]. It has a runtime of $O(N \log_2 N)$ for the worst case, where N is the number of data. A "sift-up" process is applied to arrange the input data into a binary heap. The sift-up process is analogous to corporate promotion. It can be described as the following two parts.

"add to heap" process: We can imagine the first data added to the heap as the first employee. Once we "hire" another one, he will temporarily be the subordinate first (add to the same, if there is vacancy, or lower level).

"update heap" process: We compare the newly-hired employee with his supervisor, if he is more capable, swap their positions, and repeat this comparison until we reach the top of the heap; if not, he stays put. This "update heap" process ensures that the most capable employee always stays at the top and that each upper level employee is always more capable than his subordinates. If the capability of the employees is evaluated by numbers, the sift-up process gives us a min-heap like in Figure 2. 1.

2.4.4 Shepard Interpolation

The computational cost of FMM is dominated by the potential energy calculation. One *Gaussian* calculation for a 5-atom reaction system (as shown in Equation (2.19)) takes about 3 minutes using B3LYP/6-311++G**. At a reasonable grid size, a 2-dimensional PES consists of thousands of points, so it might take several weeks to compute the entire potential energy surface. FMM does not need the entire potential energy surface, but only a narrow band along the reaction path. This saves up to 70% of *Gaussian* calculations for 2-dimensional PES and even more for higher dimensional PES.

The number of *Gaussian* calculations can be reduced even further by building the PES using Shepard interpolation. Based on N accurately calculated points (we call them "reference" points), we can approximate the potential energy at another point, $\mathbf{R}$, using the Taylor series[26,44]

$$\left(T^{(i)}(\mathbf{R}) = V(\mathbf{R}^{(i)}) + (\mathbf{R}-\mathbf{R}^{(i)})\cdot\nabla V(\mathbf{R}^{(i)}) + \frac{1}{2}(\mathbf{R}-\mathbf{R}^{(i)})\cdot\nabla\nabla V(\mathbf{R}^{(i)})(\mathbf{R}-\mathbf{R}^{(i)}) + \cdots \right)_{i=1}^{N} . \quad (2.5)$$

Due to different distances of the reference points from $\mathbf{R}$, the Taylor series from each of these points makes a different contribution to $V(\mathbf{R})$. If we model their contribution using a weight function $\omega^{(i)}(\mathbf{R})$, then the interpolated potential for point $\mathbf{R}$ is,

$$\tilde{V}(\mathbf{R}) = \sum_{i=1}^{N} \omega^{(i)}(\mathbf{R}) T^{(i)}(\mathbf{R}), \quad (2.6)$$

where $T^{(i)}(\mathbf{R})$ is the Taylor series expansion given in Eq.(2.5), and the weight function $\omega^{(i)}(\mathbf{R})$ is a non-negative, normalized function. Normalization can be enforced by

$$\omega^{(i)}(\mathbf{R}) = \frac{\upsilon^{(i)}(\mathbf{R})}{\sum_{j=1}^{N} \upsilon^{(j)}(\mathbf{R})} . \quad (2.7)$$

It is well known that the asymptotic form of $\upsilon^{(i)}(\mathbf{R})$ should be $\left\|\mathbf{R}-\mathbf{R}^{(i)}\right\|^{-(n+1)}$ if $T^{(i)}(\mathbf{R})$ is truncated at the n^{th} order term. We use the following form,

$$\upsilon^{(i)}(\mathbf{R}) = \frac{e^{-\frac{1}{2}\sum_{k=1}^{d}\left(\frac{R_k - R_k^{(i)}}{\sigma_k^{(i)}}\right)^2}}{\sum_{k=1}^{d}\left(\frac{R_k - R_k^{(i)}}{\sigma_k^{(i)}}\right)^{n+1}} , \quad (2.8)$$

where $\sigma_k^{(i)}$ is the trust radius of reference point i in the k^{th} dimension. Rather than using the Bettens-Collins isotropic formula[45] to calculate the trust radius,

$$\sigma^{(i)} = \left[\frac{1}{M} \sum_{j=1}^{M} \frac{\left(V(\mathbf{R}^{(j)}) - T(\mathbf{R}^{(i)}) \right)^2}{(\varepsilon_v)^2 \left\| \mathbf{R}^{(j)} - \mathbf{R}^{(i)} \right\|^{2n+2}} \right]^{-\frac{1}{2n+2}}, \tag{2.9}$$

we use the direction dependent formula of the form,

$$\sigma_k^{(i)} = \left[\frac{1}{M} \sum_{j=1}^{M} \frac{\left(\left(\frac{\partial V(\mathbf{R}^{(j)})}{\partial R_k} - \frac{\partial T(\mathbf{R}^{(i)})}{\partial R_k} \right) \left(R_k^{(j)} - R_k^{(i)} \right) \right)^2}{(\varepsilon_v)^2 \left\| \mathbf{R}^{(j)} - \mathbf{R}^{(i)} \right\|^{2n+2}} \right]^{-\frac{1}{2n+2}}. \tag{2.10}$$

Given the interpolated potential value $\tilde{V}(\mathbf{R})$, the error of the Shepard interpolant is estimated by,

$$err = \sqrt{\sum_{i=1}^{N} \omega^{(i)}(\mathbf{R}) \left(\tilde{V}(\mathbf{R}) - T^{(i)}(\mathbf{R}) \right)^2}. \tag{2.11}$$

If the estimated error given by Eq. (2.11) is less than the error threshold, we accept $\tilde{V}(\mathbf{R})$ as the potential for point $\mathbf{R}$. If the error is too large, then we do not use the Shepard interpolant $\tilde{V}(\mathbf{R})$, and instead we calculate the PES at this point using *Gaussian*.

If we truncate the Taylor Series at higher order terms, we expect the accuracy of Shepard interpolation to be improved. Unfortunately, computing the higher-order derivatives is very expensive. Instead we use the interpolating moving least squares

method and the potential and gradient values from *Gaussian* calculation to fit the higher-order derivatives.

2.4.5 Interpolating Moving Least Square Method

For the interpolated moving least square method, the basic equation we need to solve is[26,44]

$$\min_{\mathbf{x}} \left\| \mathbf{A}\mathbf{x} - \mathbf{b} \right\|, \tag{2.12}$$

where $\mathbf{x}$ is the vector of higher order derivatives at the point $X^{(j)}$. For these equations we assume that the energy and the first-order derivatives are available at all calculated points. If we denote the set of M neighbour points for the jth point as $\mathbf{Q}(j)$ then we can write the vector of known data $\mathbf{b}$ in the form,

$$\begin{aligned}
b_1 &= V(\mathbf{X}^{(j)}) - V(\mathbf{X}^{(Q_1(j))}) - (\mathbf{X}^{(Q_1(j))} - \mathbf{X}^{(j)}) \cdot \nabla V(\mathbf{X}^{(Q_1(j))}) \\
&\vdots \\
b_M &= V(\mathbf{X}^{(j)}) - V(\mathbf{X}^{(Q_M(j))}) - (\mathbf{X}^{(Q_M(j))} - \mathbf{X}^{(j)}) \cdot \nabla V(\mathbf{X}^{(Q_M(j))}) \\
b_{M+1} &= \frac{\partial V(\mathbf{X}_j)}{\partial \mathbf{X}_{j,1}} - \frac{\partial V(\mathbf{X}_{Q_1(j)})}{\partial \mathbf{X}_{Q_1(j),1}} \\
&\vdots \\
b_{(d+1)M} &= \frac{\partial V(\mathbf{X}_j)}{\partial \mathbf{X}_{j,d}} - \frac{\partial V(\mathbf{X}_{Q_M(j)})}{\partial \mathbf{X}_{Q_M(j),d}}
\end{aligned} \quad , \tag{2.13}$$

which has $(d+1)M$ elements. The unknown vector $\mathbf{x}$ contains the derivatives of the potential with the redundant elements removed,

$$\mathbf{x} = \begin{bmatrix} \frac{1}{2}\frac{\partial^2 V(X^{(j)})}{\partial X_1^{(j)} \partial X_1^{(j)}} \\ \frac{\partial^2 V(X^{(j)})}{\partial X_1^{(j)} \partial X_2^{(j)}} \\ \vdots \\ \frac{1}{2}\frac{\partial^2 V(X^{(j)})}{\partial X_d^{(j)} \partial X_d^{(j)}} \\ \frac{1}{6}\frac{\partial^3 V(X^{(j)})}{\partial X_1^{(j)} \partial X_1^{(j)} \partial X_1^{(j)}} \\ \frac{1}{2}\frac{\partial^3 V(X^{(j)})}{\partial X_1^{(j)} \partial X_1^{(j)} \partial X_2^{(j)}} \\ \vdots \end{bmatrix} \tag{2.14}$$

which has $\frac{1}{n!}\prod_{i=0}^{n-1}(d+i)$ elements. The matrix $\mathbf{A}$ takes the form,

$$\mathbf{A} = \begin{pmatrix} (X_1^{(j)} - X_1^{Q_1(j)})^2 & \cdots & (X_d^{(j)} - X_d^{Q_1(j)})^n \\ \vdots & \ddots & \vdots \\ (X_1^{(j)} - X_1^{Q_M(j)})^2 & \cdots & (X_d^{(j)} - X_d^{Q_M(j)})^n \\ \frac{1}{2}(X_1^{(j)} - X_1^{Q_1(j)}) & \cdots & \frac{1}{n}(X_1^{(j)} - X_1^{Q_1(j)})^{n-1} \\ \vdots & \ddots & \vdots \\ \frac{1}{2}(X_1^{(j)} - X_1^{Q_M(j)}) & \cdots & \frac{1}{n}(X_d^{(j)} - X_d^{Q_M(j)})^{n-1} \end{pmatrix}. \tag{2.15}$$

At order n each point contributes $d+1$ elements to Eq.(2.13), so there needs to be at least $\frac{1}{(d+1)n!}\prod_{i=0}^{n-1}(d+i)$ points available to solve for $\mathbf{x}$.

2.4.6 FMM program

To apply FMM to real chemical systems, we need to interface FMM with a quantum chemistry package to computer the potential energy. In this section, we will discuss how FMM is interfaced to the *Gaussian* quantum chemistry program.

A. Setup, Definitions, and Notation

i. Define the grid space

Given a chemical reaction, the first step is to determine the dimensionality of the PES that will be used in Eq. (2.3) of the FMM program. To minimize the computational cost, we use a reduced PES by choosing a few key coordinates that are essential for describing the reaction coordinate. The dimension of reduced PES is the number of key coordinates. We denote it as d. We also need to decide the minimum and maximum values of all key coordinates, so that we can limit our calculation to the region of the PES that we are interested in.

ii. Categorize the grid points

The wavefront starts from a point (usually the reactant) and propagates outward. We need to categorize the grid points inside (evaluated) and outside (unevaluated) of the wavefront and points on the wavefront (being evaluated).

"*alive*" points: points inside the wavefront. The energy cost values of the *alive* points have been evaluated and will not change any more.

"*near*" points: points on the wavefront. These points are under evaluation and their energy cost values are temporary. The energy cost of these points will be updated whenever the cost of one of their neighbouring points changes.

"*far*" points: points outside of the wavefront. These points will not be evaluated until the wavefront moves close. The energy cost values of all *far* points are assigned as infinity.

B. Initialize the calculation

i. Tag all points as *far*, and set their energy cost values as infinity.

ii. Call Gaussian to compute the potential energy and gradient of the starting point. Set the energy cost of the starting point to zero and tag it as *alive*.

iii. Tag the $2d$ neighbouring points of this first alive point as *near* and add them to the heap. Call Gaussian to compute the potential energy and potential energy gradient of each *near* point, and calculate the energy cost by solving the discretized eikonal equation, Eq. (2.3). Update the heap according to the updated energy cost values, so that the point with minimum energy cost value is at the top of the heap.

iv. Initialize the Shepard interpolation. We call these *Gaussian* points "reference points" because they will be used to approximate the potential values of nearby points. For each reference point, we need a neighbour list. This neighbour list contains M points that are used to determine the trust radius of Shepard interpolation weights and to calculate higher-order derivatives by using interpolated moving least squares.

Once we have a new *Gaussian* point, we compare its distance to the existing reference points. If it lies within an acceptable distance of a reference point, then we add it to the neighbour list of this reference point.

C. Updating the heap

i. Tag the top point of the heap as *alive*, and tag its *far* neighbouring point(s) as *near*. Add them to the heap.

ii. For each of these new *near* points, call Shepard interpolation to approximate the potential energy. If the estimated error is acceptable, then use the potential energy from the Shepard interpolant. If the estimated error is over the error threshold, call *Gaussian* to compute the potential energy and gradient. Use the potential energy to compute the energy cost and then update the heap.

iii. Repeat the above steps (i) and (ii) until the product is found or another stopping criterion is met.

D. Backtracing from the ending point to the starting point on the energy cost surface

Due to the causality condition of the eikonal equation, as the wavefront moves outwards the energy cost will always increase, which ensures that the energy cost surface is a one-well conical surface. The starting point is at the bottom. So a simple steepest

descent path from the ending point to the starting point on the energy cost surface will give the MEP. In our program this is done with Euler integration[26,46],

$$R_{k+1} = R_k - h\frac{\nabla U(R_k)}{\left\|\nabla U(R_k)\right\|}, \tag{2.16}$$

where, for simplicity, we use a fixed step size $h = \frac{\left\|d\mathbf{R}\right\|}{20}$. To compute the energy cost and gradient at point $\mathbf{R}$, we can use its 2^d nearest neighbour grid points to form a linear set of equations,

$$U(\mathbf{R}_q) = b + \sum_{i=1}^{d} a_i(\mathbf{R}_q - \mathbf{R}), \tag{2.17}$$

where $\nabla U(\mathbf{R}) = \mathbf{a}$, $U(\mathbf{R}) = b$, $q = 1...2^d$ and $\mathbf{R}_q$ are the coordinates of the nearest neighbour grid points. Since the energy cost at the neighbour grid points $U(\mathbf{R}_q)$ are known, the energy cost and gradient values at point $\mathbf{R}$ can be fitted by solving this linear set of equations.

2.4.7 Application

Our FMM program is interfaced with *Gaussian 03*.[31] All *Gaussian* calculations were done using density-functional theory (BhandhLYP/6-311++G**).

A. The 4-well analytical PES

The 4-well PES is defined by the following analytical function[28],

$$V(R_1,R_2) = V_0 + a_0 e^{-(R_1-b_1)^2-(R_2-b_2)^2} - \sum_{i=1}^{4} a_i e^{-p_i(R_1-\alpha_i)^2-q_i(R_2-\beta_i)^2}, \quad (2.18)$$

where all parameters are listed in the following table:

Table 2. 1 Parameters for the 4-well analytical PES

Parameters	Values	Parameters	Values	Parameters	Values
V_0	5.0 kcal/mol	p_1	0.3 Å^{-2}	α_1	1.3 Å
a_0	0.6 kcal/mol	p_2	1.0 Å^{-2}	α_2	-1.5 Å
a_1	3.0 kcal/mol	p_3	0.4 Å^{-2}	α_3	1.4 Å
a_2	1.5 kcal/mol	p_4	1.0 Å^{-2}	α_4	-1.3 Å
a_3	3.2 kcal/mol	q_1	0.4 Å^{-2}	β_1	-1.6 Å
a_4	2.0 kcal/mol	q_2	1.0 Å^{-2}	β_2	-1.7 Å
b_1	0.1 Å	q_3	1.0 Å^{-2}	β_3	1.8 Å
b_2	0.1 Å	q_4	0.1 Å^{-2}	β_4	1.23 Å

The 4-well PES is a standard test system for the FMM. There are four minima on this PES, and four transition states between each pair of minima. If we choose the minimum at the bottom right (R) as the reactant and the one at the top right (P) as the product, then there are two possible pathways: (a) the "direct" 1-step pathway, and (b) the "C-shaped" 3-step pathway passing through by two intermediates (I and II) and three

transition states (TS1, TS2 and TS3) as shown in Figure 2. 2. Starting from the reactant, we imagine the FMM procedure as slowly adding water to the reactant valley[33]; the "water" level can be considered as the propagating wavefront. The water level will keep going up, wetting the contours of the potential energy surface as it does so. Eventually the water level will rise to the level of the lowest-energy TS, which is the lowest "mountain pass" for exiting the reactant valley. At this stage a narrow thread of water will follow the steepest-descent path to the bottom of the next valley. The water keeps flooding mountain valleys in this way until the product is found. In FMM, the "energy cost" contours record which portions of the PES are "flooded" at any given point in time (see Figure 2. 3). Notice that only the "flooded" portion of the surface needs to be computed. This reduces the computational cost significantly.

The lower energy region of the PES in Figure 2. 2 is transformed into an energy cost surface in Figure 2. 3, and higher energy part is cut off. Backtracing from the product to the reactant along the steepest descent path gives the MEP.

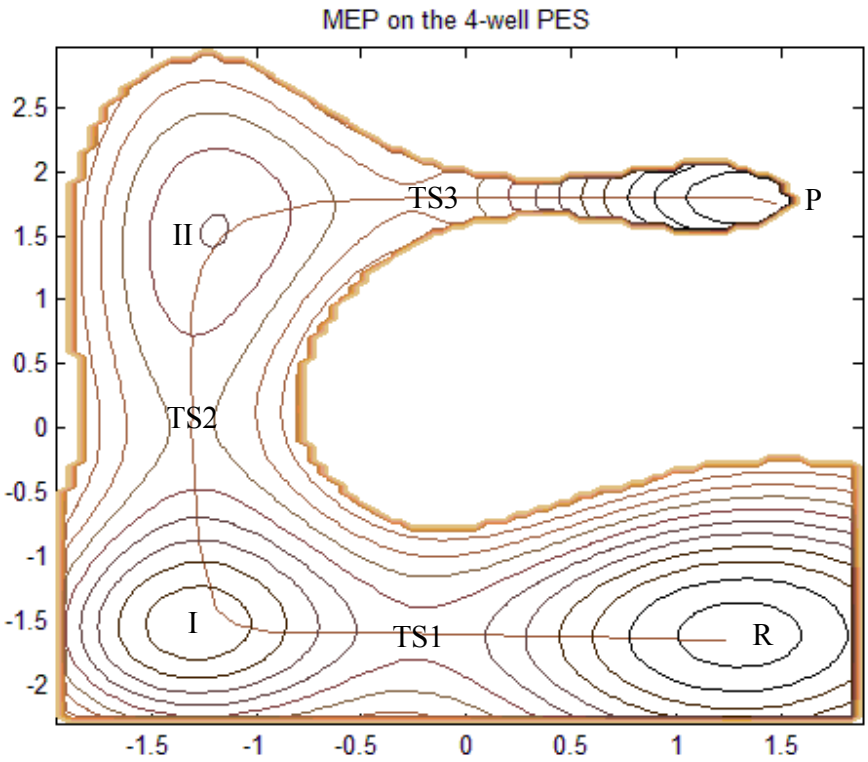


Figure 2. 2: The MEP on the 4-well PES. The grid sizes on both dimensions are $dR = 0.05$**.**

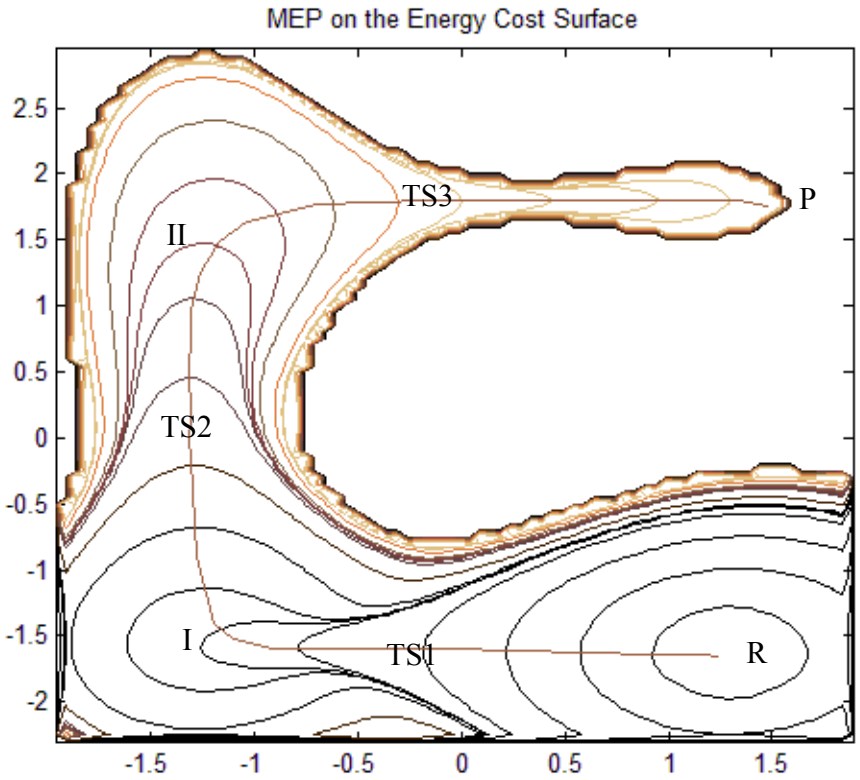
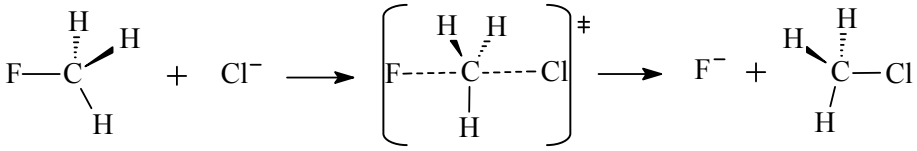


Figure 2. 3: The MEP on the energy cost surface transformed from the 4-well PES by solving the eikonal equation. The MEP is determined by backtracing from the product to the reactant along the steepest descent path on the energy cost surface.

B. The S_N2 Reaction[33]

The mechanism of the S_N2 reaction has been studied intensively by experimental and theoretical methods, so it is a good test for FMM. Equation (2.19) is an example of the S_N2 reaction. This is a one-step reaction, so we expect two minima (the reactant R and product P) and one TS on the PES.



(R) (TS) (P) (2.19)

In this reaction, only C-F and C-Cl bonds are involved in bond-breaking and bond-forming, so the PES of this reaction can be modelled using a 2-dimensional reduced PES based on the C-F and C-Cl coordinates. At each grid point, we will freeze the C-F and C-Cl bond lengths at the given values and minimize the energy with respect to the other coordinates. The 2-dimensional reduced PES and the MEP computed by the FMM program are depicted in Figure 2. 4. About 20% of grid points are in the "flooded" region and are computed by *Gaussian 03*. The energy-cost surface with the reactant (R) as starting point and the MEP found on this surface are shown in Figure 2. 5. Plotting the change in potential energy along the MEP gives the energy profile of the reaction coordinate (Figure 2. 6).

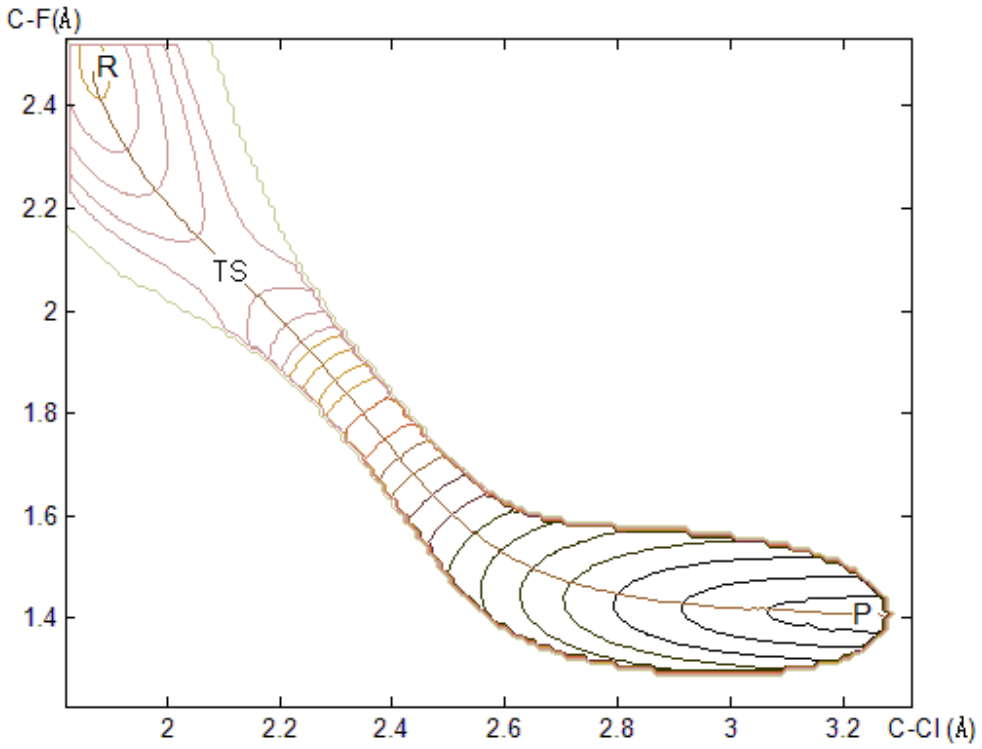


Figure 2. 4: The PES of the S_N2 reaction based on C-F and C-Cl bond lengths. The grid sizes on both dimensions are dR = 0.01 Å. The calculation starts from the reactant (R), fills the reactant valley, breaches the reaction barrier at the transition state (TS), and then "flows" down to the product (P). The FMM program transforms this PES to an energy-cost surface (Figure 2.5).

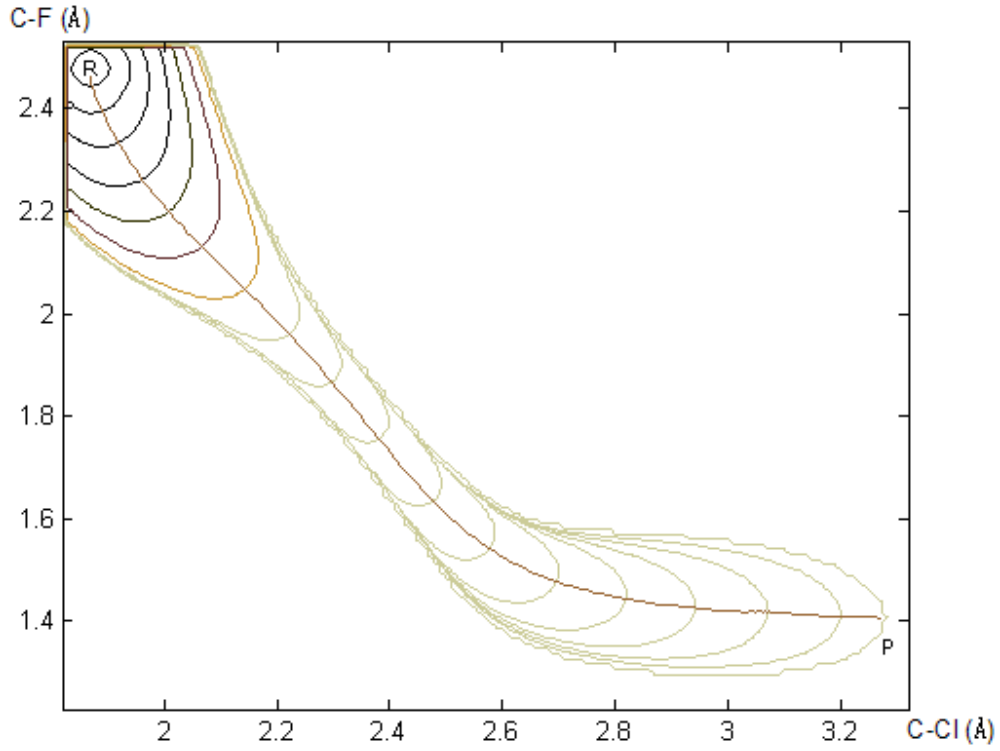


Figure 2. 5: The energy-cost surface transformed from the PES on Figure 2. 1. The MEP is determined by backtracing from the product to the reactant along the steepest descent path on the energy-cost surface.

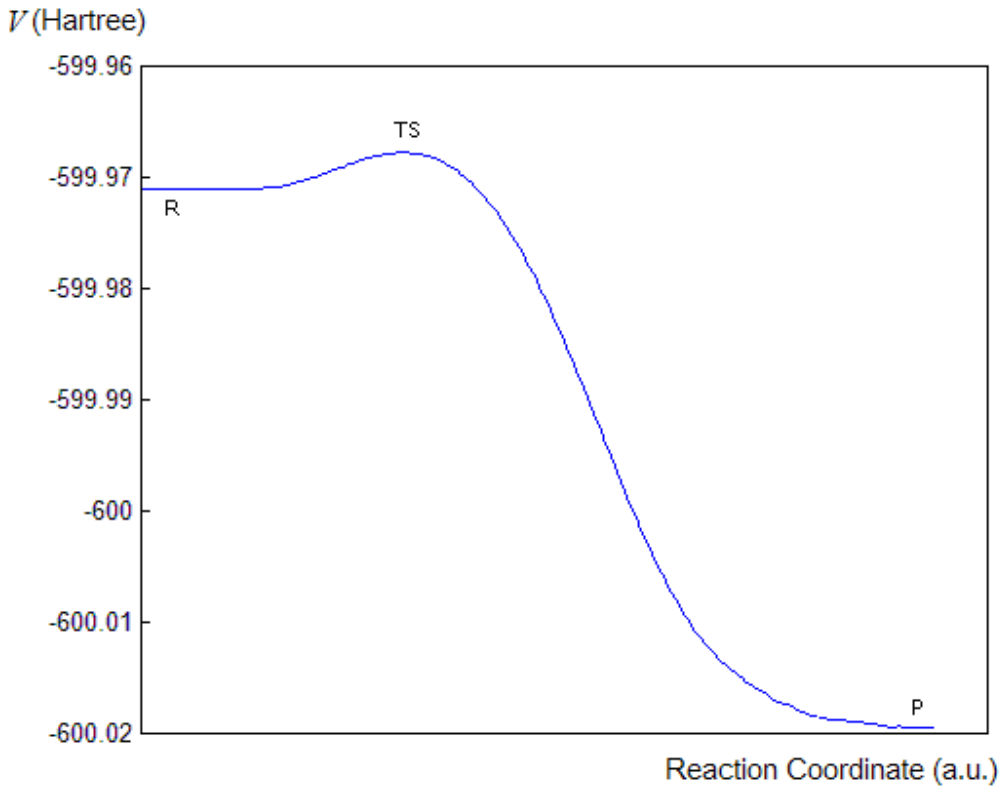


Figure 2. 6: The energy profile of the S_N2 reaction.

C. The dissociation of ionized O-methylhydroxylamine[33]

The PES of $[CH_5NO]^{+\cdot}$ has been studied using mass spectroscopy and computational methods[47]. The following dissociation reaction has been observed,

$$[CH_5NO]^{+\cdot} \longrightarrow [CH_2NH_2]^{+} + OH^{\cdot} \qquad (2.20)$$

Terlouw and coworkers proposed the following mechanism for this dissociation reaction[47],

$$[CH_3{-}O{-}NH_2]^{+\bullet} \longrightarrow [O{-}NH_2{-}CH_3]^{+\bullet} \longrightarrow [HO{-}NH_2{-}CH_2]^{+\bullet} \longrightarrow HO\cdot + NH_2\overset{+}{C}H_2 \qquad (2.21)$$

Using bond lengths C-N, N-O and O-H as key coordinates, FMM finds a reduced 3-dimensional PES. The 3-dimensional equipotential surfaces have onion-like structures. Each layer of the "onion" represents a certain value of the potential energy. Figure 2.7 shows one layer of the "onion" with a potential value of -170.534 Hartrees. The cores of the onions represent minima on the PES. We can see that there are 4 minima on this PES, the reactant (R), two intermediates (I, II), and the product (P). The coordinates of the minima show that the structures of intermediate (I) and (II) coincide with $\dot{O}{-}\overset{+}{N}H_2{-}CH_3$ and $H\dot{O}{-}NH_2{-}\overset{+}{C}H_2$ respectively. The energy profile is shown in Figure 2.8. The FMM calculation confirms that the mechanism in (2.21) is the minimum energy reaction pathway.

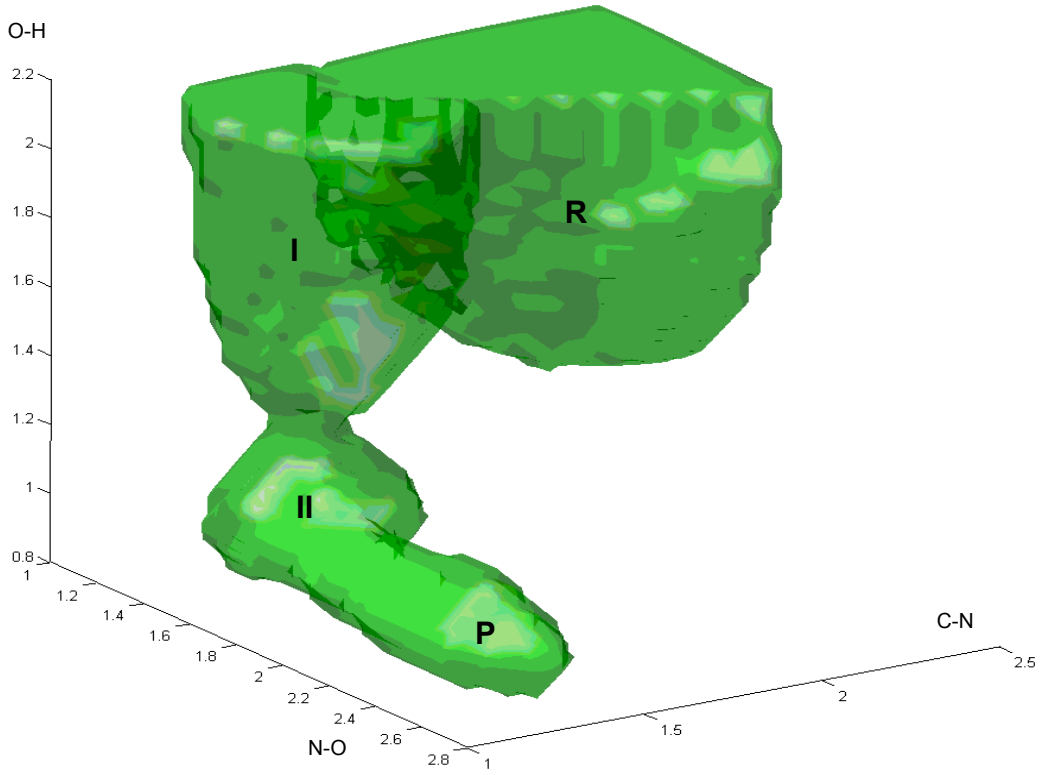


Figure 2. 7: The isosurface with a potential value of -170.534 Hartrees, which is one layer of the reduced 3-dimensional PES for the dissociation reaction of ionized O-methylhydroxylamine. The 3-dimensional equipotential surfaces have an onion-like structure. Each layer of the "onion" represents a certain value of the potential energy. The cores of the "onions" are minima on the PES.

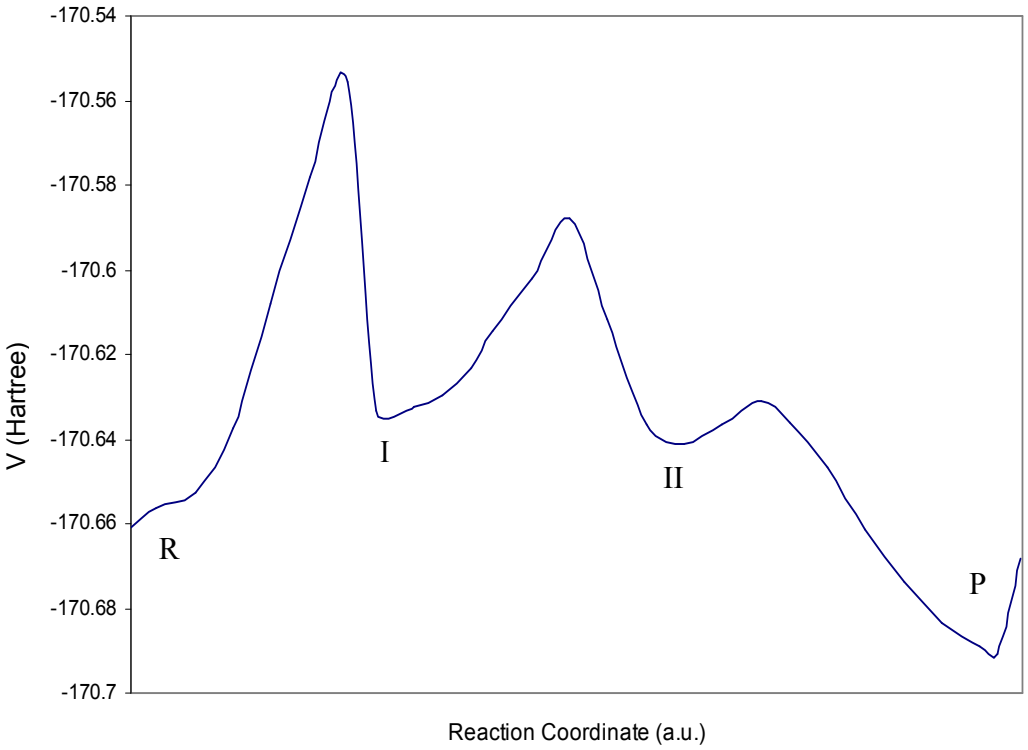


Figure 2. 8: The energy profile of the dissociation reaction of ionized O-methylhydroxylamine.

2.5 Quantum Mechanics/Molecular Mechanics (QM/MM) methods applied to Enzyme-catalyzed reactions

2.5.1 QM/MM Methods

Enzyme-catalyzed reactions are of great importance in the biological sciences and pharmaceutical industry because of their efficiency and specificity. Using computational

tools to study the mechanism of enzyme-catalyzed reactions is one of our ultimate goals. Even with the advances of modern computers and new computational methods, studying the mechanism of enzyme-catalyzed reactions is still a great challenge due to the large size of the enzyme system. QM methods are accurate but expensive, and so are generally limited to systems of less than 100 atoms. For enzyme-catalyzed reactions that involve thousands of atoms, it is impossible to apply QM methods to the entire system. To deal with larger systems molecular mechanics is commonly employed. The accuracy of MM methods can be poor and it is unsuitable for studying bond-breaking and bond-forming processes in chemical reactions.

In a typical enzyme-catalyzed reaction, only a small number of atoms are involved directly in the bond-breaking and bond-forming processes; the primary role of the other atoms is to provide a favourable steric and electrostatic environment. This realization led Warshel and Levitt to propose the hybrid QM/MM approach[48,49]. In QM/MM the enzyme reaction system is divided into two parts: the atoms that are directly involved in the reaction are evaluated quantum mechanically, while the rest of the atoms are treated with MM methods. This approach combines the advantages of the high accuracy of QM methods for the small QM subsystem, and the computational affordability of MM methods for the remainder of the molecules (see Figure 2. 9). After three decades of development, QM/MM methods have been successfully applied in simulations of various enzyme-catalyzed reactions[50-60].

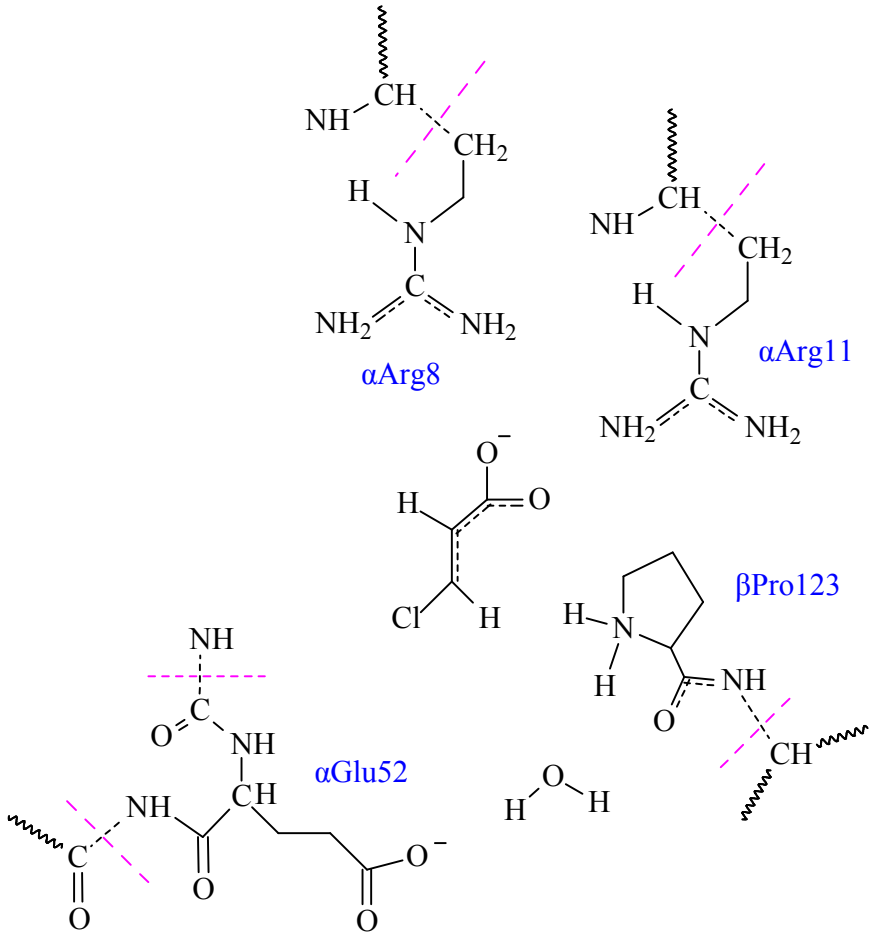


Figure 2. 9: The QM subsystem (the substrates, part of residue αArginine8, αArginine11, αGlutamine52, and βProline123) and MM subsystem (the rest of the system) of the dechlorination of trans-3-chloroacrylate catalyzed by trans-3-ChloroAcrylic Acid Dehalogenase (CAAD).[69]

One important problem associated with QM/MM methods is how to deal with the QM and MM covalent boundary. The link atom approach is one of the most commonly used methods[61-64]. In the link atom approach, link atoms like hydrogen or pseudohalogen atoms are inserted to cover the free valence of the QM subsystem so that the QM

subsystem will still be a closed-shell system. The problem with the link atom approach is that it introduces additional degrees of freedom and some double counting of the interactions into the system, which can be difficult to correct for. Due to the deficiency of the link atom approach, in the following discussion we will focus our attention on the pseudobond QM/MM method developed by Yang's group[57,58,60]. The pseudobond approach does not introduce additional atoms to the system. Instead this approach replaces the MM boundary atom with a seven-valence-electron atom with an effective core potential and forms a pseudobond between this atom and the QM boundary atom[61,62]. The pseudobond approach gives a smooth interface between the QM and MM subsystems and provides a consistent and well-defined *ab initio* QM/MM potential energy surface.

QM/MM methods can be categorized into two types: the semiempirical QM/MM methods and the *ab initio* QM/MM methods depending on the level of QM theory used. Semiempirical QM/MM methods are much faster computationally so that classical statistical sampling can be applied. But semiempirical QM/MM methods are often not sufficiently accurate to give reliable free energies[65]. *Ab initio* QM/MM methods are accurate but expensive, so reaction path ensemble sampling is not feasible. The QM/MM Free-Energy Perturbation (QM/MM-FEP) method developed by Yang's group utilizes the pseudobond approach to form a smooth interface between QM and MM subsystems, then applies an efficient, iterative optimization procedure[59,66] to optimize the QM and MM subsystems of a given conformation independently and iteratively until convergence. Incorporated with a reaction path optimization method[3,8,13,66,67], the reaction path can be

found on the PES. The last step is to perform free-energy perturbation calculations on the reaction path to give the free-energy profile of the reaction. The problem with the QM/MM-FEP method is that the optimization of the reaction path depends on the PES of a single MM conformation[54]. To eliminate this dependence one can instead do a direct path optimization on the free energy surface.

The most recent QM/MM minimum free-energy path (QM/MM-MFEP) method[54,56] is one of the more efficient and reliable *ab initio* QM/MM methods. Unlike other *ab initio* QM/MM methods, the free energy profile obtained in the QM/MM-MFEP method is not built from a previously sampled PES of a random chosen initial conformation of the system, instead it is generated naturally because the reaction path is optimized on the potential of mean force (PMF) surface, which is the free energy expression of the QM subsystem with the MM contributions averaged out. Thus, the problem of finding the reaction path in a complicated phase space with the same number of degrees of freedom of the entire QM/MM system is simplified to a problem of exploring the PMF surface depending on just the QM degrees of freedom[54].

2.5.2 Incorporating the QM/MM-MFEP Methods with FMM

QM/MM-MFEP methods[54,56] have been incorporated with several path optimization methods, such as NEB[3], the Ayala-Schlegel second-order MEP method[67], and QSM[13]. All these methods aim to find the local MEP of the enzyme reaction. To ensure the

convergence to the global MEP, we can implement the QM/MM-MFEP methods with FMM.

To carry out reaction path optimization on the PMF surface, the relative free energies between adjacent QM conformations and free-energy gradients for each individual QM conformation need to be computed[54]. The relative free energies between adjacent QM conformations are computed by the QM/MM-FEP method. The free energy difference is defined as[56],

$$\Delta A = A^{(n)}(r_{QM}) - A_{ref} = -\frac{1}{\beta}\ln[\frac{1}{N}\sum_{\tau=1}^{N}\exp\{-\beta[E(\mathbf{r}_{QM},\mathbf{r}_{MM}^{(n)}(\tau)) - E_{ref}(\mathbf{r}_{MM}^{(n)}(\tau))]\}] \tag{2.22}$$

where an MD simulation is performed on the MM subsystem with the QM conformation frozen. Then FMM is performed within a trust radius using the same MD ensemble. Outside of this trust radius the FMM algorithm can continue only when a new MD simulation is performed with a new QM conformation.

The free-energy gradients of the QM subsystem are computed through molecular dynamics sampling of the MM environment. The free-energy gradient associated with Eq. (2.22) is computed as[56],

$$\frac{\partial A(\mathbf{r}_{QM})}{\partial \mathbf{r}_{QM}} = \frac{\frac{1}{N}\sum_{\tau=1}^{N}\left(\frac{\partial E(\mathbf{r}_{QM},\mathbf{r}_{MM})}{\partial \mathbf{r}_{QM}}\exp\{-\beta[E(\mathbf{r}_{QM},\mathbf{r}_{MM}^{(n)}(\tau)) - E_{ref}(\mathbf{r}_{MM}^{(n)}(\tau))]\}\right)}{\frac{1}{N}\sum_{\tau=1}^{N}\exp\{-\beta[E(\mathbf{r}_{QM},\mathbf{r}_{MM}^{(n)}(\tau)) - E_{ref}(\mathbf{r}_{MM}^{(n)}(\tau))]\}} \tag{2.23}$$

The rest of the FMM algorithm is the same as in Section 3.

2.5.3 Application of the incorporated FMM and QM/MM-MFEP method to enzyme-catalyzed reactions

Using FMM as the path optimization algorithm, the QM/MM-MFEP method can be applied to find the global MEP for solution-phase reactions and enzyme-catalyzed reactions. Below we will present a representative sample of the applications we are currently pursuing using this new methodology.

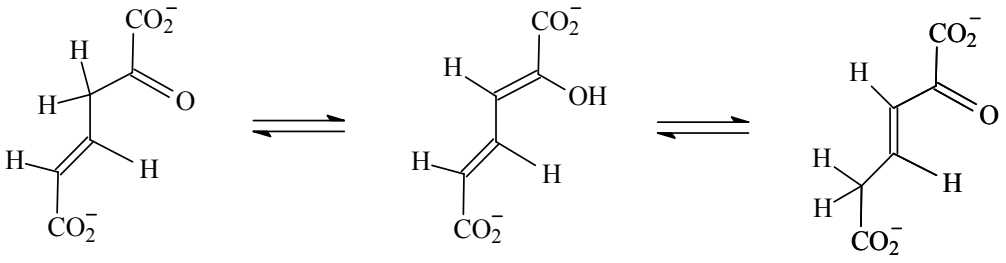
A. The S_N2 Reaction in solvent

The solvent S_N2 Reaction is a good test for the incorporated FMM and QM/MM-MFEP method. This reaction has been studied intensively by experimental and theoretical methods. So there is plenty of data to compare with.

Because of the rapid exchange of solvent molecules, QM/MM methods that depend on the initial conformation of the system cannot give reliable results because the initial conformation does not reflect the rapid change of solvent. Since the QM/MM-MFEP method eliminates this dependence, we expect better results for this solvent reaction.

B. The isomerization reaction catalyzed by 4-oxalocrotonate tautomerase (4-OT)[51,68]

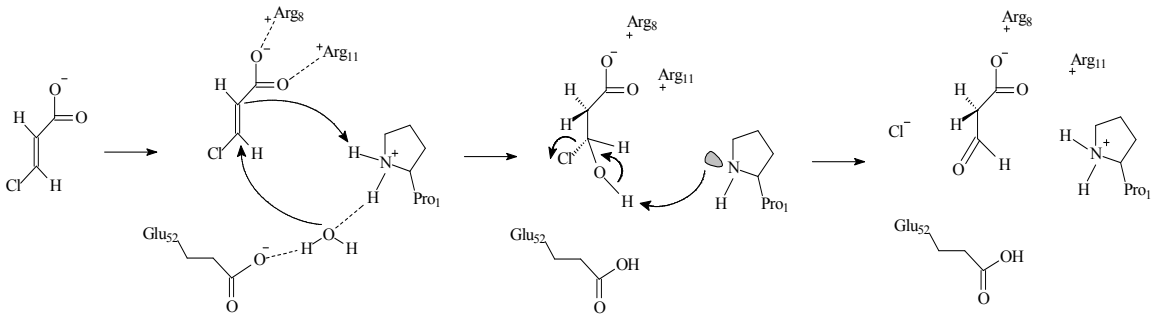
The mechanism of this reaction has been studied using the QM/MM-FEP method. The incorporated FMM and QM/MM-MFEP method can confirm whether the reaction path is a global MEP.



(2.24)

C. The dechlorination reaction catalyzed by trans-3-chloroacrylic acid dehalogenase (CAAD)

3-chloroacrylic acid is an unnatural substance degraded from the active ingredient of the nematocides Shell D-D and Telone II. Its uncatalyzed half life is about 10,000 years[69]. While catalyzed by CAAD, this hydrolytic dechlorination reaction proceeds with a rate enhancement of 2×10^{12} . The X-ray structure of trans-3-chloroacrylic acid dehalogenase gives some hint on the mechanism of the dechlorination reaction of trans-3-chloroacrylic acid[70]. We are planning to apply the incorporated FMM and QM/MM-MFEP method to study the mechanism of this reaction (as shown in Equation (2.25)).



(2.25)

2.6 Summary

In this chapter, we gave a brief review of some numerical methods that locate the MEP on the PES or free energy surface. We focused on the FMM, which is one of the most general and reliable methods for finding the chemical reaction path. Unlike most competing methods, FMM always finds the global MEP. Some proof-of-principle examples of applying FMM to small gas phase reactions were shown in Section 3. Most reactions are more complicated than this. Our ultimate goal is to study the mechanism of more realistic systems such as those solution phase or enzyme-catalyzed reactions. To deal with the effects of the complicated molecular environment, QM/MM methods were introduced. A brief history of the development of QM/MM methods was given in Section 4, followed by the key ideas required to merge FMM with the recent QM/MM potential of mean force-based free energy path finding methods. The combination of QM/MM methods with FMM is a promising approach for determining chemical reaction mechanisms in complex reaction systems.

Reference List

1. Fukui, K. *Accounts of Chemical Research* **1981,** *14* (12), 363-368.
2. Koslover, E. F.; Wales, D. J. *J. Chem. Phys.* **2007,** *127* (13).
3. Jonsson, H.; Mills, G.; Jacobsen, K. W. World Scientific: Singapore, 1998; pp 385-404.
4. Alfonso, D. R.; Jordan, K. D. *J. of Compt. Chem.* **2003,** *24* (8), 990-996.
5. Chu, J.-W.; Trout, B. L.; Brooks, B. R. 2003; pp 12708-12717.
6. Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **2000,** *113* (22), 9978-9985.
7. Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004,** *120* (5), 2082-2094.
8. Xie, L.; Liu, H. Y.; Yang, W. T. *J. Chem. Phys.* **2004,** *120* (17), 8039-8052.
9. E, W. N.; Ren, W. Q.; Vanden Eijnden, E. *Physical Review B* **2002,** *66* (5), 052301.
10. E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. *J. Chem. Phys.* **2007,** *126* (16), 164103.
11. Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. N. *J. Chem. Phys.* **2005,** *123* (13), 134109.
12. Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004,** *120* (17), 7877-7886.
13. Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2006,** *124* (5), 054109.
14. E W. N.; Ren W. Q.; *J. Chem. Phys.* **2005,** *109* (14), 6688-6693.
15. E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. *Physical Review B* **2002,** *66* (5), 052301-.
16. Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2006,** *124* (5), 054109.
17. Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2006,** *125* (24), 244108.
18. Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2006,** *124* (22), 224102.
19. Tsai, C. J.; Jordan, K. D. *J. of Phys. Chem.* **1993,** *97* (43), 11227-11237.
20. Sun, J. Q.; Ruedenberg, K. *J. Chem. Phys.* **1993,** *98* (12), 9707-9714.
21. Quapp, W.; Hirsch, M.; Imig, O.; Heidrich, D. *Journal of Computational Chemistry* **1998,** *19* (9), 1087-1100.
22. Maeda, S.; Watanabe, Y.; Ohno, K. *Chemical Physics Letters* **2005,** *414* (4-6), 265-270.
23. Maeda, S.; Ohno, K. *Journal of Physical Chemistry A* **2005,** *109* (25), 5742-5753.
24. Maeda, S.; Ohno, K. *Chemical Physics Letters* **2005,** *404* (1-3), 95-99.
25. Ohno, K.; Maeda, S. *Chemical Physics Letters* **2004,** *384* (4-6), 277-282.
26. Burger, S. K.; Liu, Y.; Sarkar, U.; Ayers, P. W. *J. Chem. Phys.* **2009**, 130, 024103.

27. Dey, B. K.; Janicki, M. R.; Ayers, P. W. *J. Chem. Phys.* **2004,** *121* (14), 6667-6679.
28. Dey, B. K.; Ayers, P. W. *Molecular Physics* **2006,** *104* (4), 541-558.
29. Sethian, J. A. *Proceedings of the National Academy of Sciences of the United States of America* **1996,** *93* (4), 1591-1595.
30. Sethian, J. A.; Adalsteinsson, D. *IEEE Transactions on Semiconductor Manufacturing* **1997,** *10* (1), 167-184.
31. Sethian, J. A. *Siam Review* **1999,** *41* (2), 199-235.
32. Sethian, J. A.; Vladimirsky, A. *Proceedings of the National Academy of Sciences of the United States of America* **2000,** *97* (11), 5699-5703.
33. Yuli Liu; Paul W.Ayers *J. Math. Chem.***2011**,49(7), 1291-1301.
34. Hirsch, M.; Quapp, W. *Journal of Molecular Structure-Theochem* **2004,** *683* (1-3), 1-13.
35. Quapp, W. *Journal of Mathematical Chemistry* **2004,** *36* (4), 365-379.
36. Quapp, W.; Hirsch, M.; Heidrich, D. *Theoretical Chemistry Accounts* **2004,** *112* (1), 40-51.
37. Quapp, W. *Journal of Computational Chemistry* **2004,** *25* (10), 1277-1285.
38. Quapp, W. *J. Chem. Phys.* **2005,** *122* (17), 174106.
39. Quapp, W. *Journal of Computational Chemistry* **2007,** *28* (11), 1834-1847.
40. Laio, A.; Parrinello, M. *Proceedings of the National Academy of Sciences of the United States of America* **2002,** *99* (20), 12562-12566.
41. Grubmuller, H. *Physical Review E.* **1995,** *52* (3), 2893-2906.
42. Zhao, H. K. *Mathematics of Computation* **2005,** *74* (250), 603-627.
43. Press W.H.; Teukolsky, S. A.; Vetterling W.T.; Flannery, B. P. Numerical Recipes. 2008.
44. Collins, M. A. *Theoretical Chemistry Accounts* **2002,** *108* (6), 313-324.
45. Bettens, R. P. A.; Collins, M. A. *J. Chem. Phys.* **1999,** *111* (3), 816-826.
46. Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2006,** *125* (24), 244108.
47. Burgers, P. C.; Lifshitz, C.; Ruttink, P. J. A.; Schaftenaar, G.; Terlouw, J. K. *Organic Mass Spectrometry* **1989,** *24* (8), 579-590.
48. Warshel, A.; Levitt, M. *Journal of Molecular Biology* **1976,** *103* (2), 227-249.
49. Warshel, A.; Hwang, J. K.; Aqvist, J. *Faraday Discussions* **1992,** (93), 225-238.
50. Bentzien, J.; Muller, R. P.; Florian, J.; Warshel, A. *Journal of Physical Chemistry B* **1998,** *102* (12), 2293-2301.
51. Cisneros, G. A.; Liu, H. Y.; Zhang, Y. K.; Yang, W. T. *Journal of the American Chemical Society* **2003,** *125* (34), 10384-10393.
52. Cisneros, G. A.; Wang, M.; Silinski, P.; Fitzgerald, M. C.; Yang, W. T. *Biochemistry* **2004,** *43* (22), 6885-6892.
53. Cisneros, G. A.; Wang, M.; Silinski, P.; Fitzgerald, M. C.; Yang, W. T. *Journal of Physical Chemistry A* **2006,** *110* (2), 700-708.

54. Hu, H.; Lu, Z. Y.; Yang, W. T. *Journal of Chemical Theory and Computation* **2007,** *3* (2), 390-406.
55. Hu, H.; Yang, W. T. *Annual Review of Physical Chemistry* **2008,** *59*, 573-601.
56. Hu, H.; Lu, Z. Y.; Parks, J. M.; Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2008,** *128* (3), 034105.
57. Zhang, Y. K.; Yang, W. T. *Abstracts of Papers of the American Chemical Society* **1999,** *218*, U528.
58. Zhang, Y. K.; Lee, T. S.; Yang, W. T. *J. Chem. Phys.* **1999,** *110* (1), 46-54.
59. Zhang, Y. K.; Liu, H. Y.; Yang, W. T. *J. Chem. Phys.* **2000,** *112* (8), 3483-3492.
60. Zhang, Y. K. *J. Chem. Phys.* **2005,** *122* (2), 024114.
61. Gao, J. L.; Amara, P.; Alhambra, C.; Field, M. J. *Journal of Physical Chemistry A* **1998,** *102* (24), 4714-4721.
62. Eurenius, K. P.; Chatfield, D. C.; Brooks, B. R.; Hodoscek, M. *International Journal of Quantum Chemistry* **1996,** *60* (6), 1189-1200.
63. Das, D.; Eurenius, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodoscek, M.; Brooks, B. R. *J. Chem. Phys.* **2002,** *117* (23), 10534-10547.
64. Amara, P.; Field, M. J. *Theoretical Chemistry Accounts* **2003,** *109* (1), 43-52.
65. vi-Kesavan, L. S.; Garcia-Viloca, M.; Gao, J. *Theoretical Chemistry Accounts* **2003,** *109* (3), 133-139.
66. Liu, H. Y.; Lu, Z. Y.; Cisneros, G. A.; Yang, W. T. *J. Chem. Phys.* **2004,** *121* (2), 697-706.
67. Ayala, P. Y.; Schlegel, H. B. *J. Chem. Phys.* **1997,** *107* (2), 375-384.
68. Wang, S. C.; Johnson, W. H.; Whitman, C. P. *Journal of the American Chemical Society* **2003,** *125* (47), 14282-14283.
69. Horvat, C. M.; Wolfenden, R. V. *Proceedings of the National Academy of Sciences of the United States of America* **2005,** *102* (45), 16199-16202.
70. de Jong, R. M.; Brugman, W.; Poelarends, G. J.; Whitman, C. P.; Dijkstra, B. W. *Journal of Biological Chemistry* **2004,** *279* (12), 11546-11552.

Chapter 3:

Finding Minimum Energy Reaction Paths on *Ab Initio* Potential Energy Surfaces Using the Fast Marching Method*

* This chapter is accepted for publication by J. Math. Chem.: Yuli Liu; Paul W.Ayers Finding Minimum Energy Reaction Paths on Ab Initio Potential Energy Surfaces Using the Fast marching Method.

3.1 Statement of the Problem

This chapter presents the fast marching method (FMM) for determining minimum-cost paths. FMM has been extended to compute the minimum-energy reaction coordinates in chemical reactions. This was accomplished by building an interface between FMM and the *Gaussian* program. We demonstrate the new method using an S_N2 reaction, the isomerization of HSCN to HNCS, and a gas-phase rearrangement reaction of relevance in mass spectrometry. Some of these example reactions have also been used in chapter 2.

Both chapter 2 and chapter 3 present the fast marching method. Chapter 2 focuses on the improvement and numerical details of FMM and its advantages compared to other path-finding methods. Chapter 3 focuses on the basic idea and concept of FMM and the extension of its application from analytical PES to real chemical reactions.

3.2 Introduction

The multidimensional potential energy surface (PES) of a molecular system contains important information about the geometries and relative energies of its locally stable structures, as well as the reaction paths between them. When the reactants and products are known, the chemical reaction coordinate is often associated with the minimum energy path (MEP) connecting the reactant to the product. Once the MEP has been found, one knows the reaction mechanism, which is ordinarily characterized in terms of the local

energy minima (reactive intermediates) and energy maxima (transition states) along the reaction path. One can also use the MEP to estimate the rate of reaction (using transition state theory).

Unfortunately, finding the minimum-energy reaction path is generally difficult. For simple reactions or reactions in which abundant experimental information is available beforehand, one can often make a "good guess" for the reaction path and then refine it using a method like the nudged elastic band[1-4] or the quadratic string method[5-8]. In other cases, one can profitably use methods like coordinate driving[9], eigenvector-following techniques[10-13], or synchronous transit-guided Quasi-Newton[14,15] to locate the transition state; the ordinary intrinsic-reaction coordinate (IRC) algorithm[16-19] then suffices to find the minimum energy path. This approach, however, does not work for multi-step reaction mechanisms. While two-point methods like the quadratic string method are perfectly valid even for complex multi-step reaction mechanisms, finding a "good guess" for the reaction path becomes exponentially more difficult as the complexity of the reaction mechanism increases. We would like to have a more robust way to determine reaction paths for complex reaction mechanisms.

Recently, we have proposed adapting the fast-marching method (FMM) to determine minimum-energy reaction paths in complex multi-step reactions[20-25]. The fast-marching method is a numerical scheme for solving the nonlinear eikonal equation and related static Hamilton-Jacobi equations[26-29]. By defining an energy-cost function between the reactant and any other point, the fast marching method can transform the

potential energy surface (PES) or free energy surface to an energy-cost surface. Unlike the multi-well PES, the energy-cost surface is a conical surface with a single minimum (the starting point, which is ordinarily either the reactant or the product of the chemical reaction). Moreover, given any molecular configuration, the minimum energy path is simply located by finding the steepest-descent path on the energy-cost surface. The process of finding this steepest-descent path is called backtracing.

In previous work, we applied the FMM method to some analytical two-dimensional potential energy surfaces.[22] Those results were quite favorable, so we elected to extend the method to higher-dimensional PES and interface the method with an electronic structure theory program, so that we could explore reactions for which analytical PES are unavailable. This paper reports our efforts in these directions. We have also worked on developing interpolation methods to reduce the number of quantum chemistry calculations that are needed to model the potential energy surface; those results have been reported separately.[24,30]

In the next section of this paper, we will briefly review the FMM methodology. Then we will present applications to three chemical reactions, of increasing complexity. We conclude with some comments on the extensions and improvements that we are currently pursuing.

3.3 The Fast-Marching Method

We define the cost function at **R** as the "minimum cost" required to attain this configuration starting from configuration $\mathbf{R}_0$:[22]

$$U_n(\mathbf{R}) = \underbrace{\min}_{\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(s)} \int_0^L \left\{\sqrt{2(E - V(\mathbf{C}(s)))}\right\}^n ds \,. \tag{3.1}$$

Here the minimization is over all paths, $\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(s)$, that start at $\mathbf{R}_0$ and end at $\mathbf{R}$, E is the total energy of the system, $V(\mathbf{R})$ is the potential energy surface, and L is the path length. (So $\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(0) = \mathbf{R}_0$ and $\mathbf{C}_{\mathbf{R}_0,\mathbf{R}}(L) = \mathbf{R}$.) The path integral problem, (2.1), can be conveniently restated as an eikonal equation, namely,

$$\left|\nabla U_n(\mathbf{R})\right| = \left\{\sqrt{2(E - V(\mathbf{R}))}\right\}^n \tag{3.2}$$

The energy-cost of the reactant is zero by definition ($U(\mathbf{R}_0) = 0$); this is the boundary condition for the eikonal equation.

This eikonal equation describes wavefront propagation with the local speed function $\dfrac{1}{\left\{\sqrt{2(E - V(\mathbf{R}))}\right\}^n}$. To locate the MEP, we need the cost of molecular configurations with higher potential energy to be infinitely larger than the cost of configurations with lower potential energy. (Equivalently, we need for configurations that are lower in energy to be attained infinitely faster than configurations that are higher in energy.) This can be achieved by letting $n \to -\infty$, which ensures that higher energy paths in equation (3.1) are cut off from the set of paths ($C_{\mathbf{R}_0,\mathbf{R}}$), giving only the MEP. Of

course, in computational implementations we will choose n to be a sizeable (but non-infinite) negative number. In practice, it seems that results are usually good as soon as n is less than minus ten.

Solving this eikonal equation transforms a multi-well potential energy surface, $V(\mathbf{R})$, into a conical energy-cost surface $U(\mathbf{R})$. Information about the fast-marching scheme for the 2-dimensional eikonal equation can be found in references [22,23]. Our previous implementation required a complete PES scan in advance, which is very expensive and tends to fail at some regions (e.g., maxima on the PES). In order to apply the fast marching method to arbitrary dimensional PES of chemical reactions, we need two key innovations: first we need an improved upwind derivative formula so that we can solve the d-dimensional eikonal equation. Second we need to interface the fast-marching program with a quantum chemistry package (we are using *Gaussian 03*) so that the potential energy can be computed at molecular configurations of interest. The new FMM program doesn't compute the complete PES; instead it will push the advancing wavefront outward along the equipotential contours and call *Gaussian 03* to calculate the potentials of the points on the front. The energy-cost values are then computed by solving the eikonal equation (3.2).

We need to solve the eikonal equation using an upwind finite difference approximation, which will preserve the causality of the solutions. To do this, we discretize the eikonal equation as follows:

$$\left(\frac{(U-a_1)^+}{dR(1)}\right)^2+\left(\frac{(U-a_2)^+}{dR(2)}\right)^2+\cdots+\left(\frac{(U-a_d)^+}{dR(d)}\right)^2=\left[2(E-V(\mathbf{R}))\right]^n \qquad (3.3)$$

Here $dR(i)$ is the i^{th} component of the grid size vector $d\mathbf{R}$; a_i is the smaller cost value of point $\mathbf{R}'s$ two neighbouring points in dimension i, $a_i=\min(U_{left},U_{right})$; $(U-a_i)^+=(U-a_i)$ if $U>a_i$ and $(U-a_i)^+=0$ otherwise. (That is, $(U-a_i)^+=\max(0,U-a_i)$.) The "upwind finite difference," $(U-a_i)^+$, enforces the causality condition in the fast marching method, which means the cost can only increase while the interface moves outward. In other words, for the point in question, its unknown cost value U has to be greater than the cost value, a_i, of its known neighbouring point; if $a_i>U$, then the cost value of this neighbouring point must not be known either. We can not use an unknown point, so we discard it by letting $(U-a_i)^+=0$. This is the idea behind the upwind finite difference approximation.

Equation (3.3) can be solved in an iterative way. First the $a_i's$ are sorted in increasing order. Start from $j=1$ and solve the truncated equation:

$$\left(\frac{(U-a_1)^+}{dR(1)}\right)^2=\left[2(E-V(\mathbf{R}))\right]^n \qquad (3.4)$$

If the solution $U_1\le a_2$, then $U_1\le a_3\le\cdots\le a_d$ and $U=U_1$ is also the solution to equation (3.3). If $U_1>a_2$, then let $j = j + 1$, and continue to solve the truncated equation with 2

terms on the left side. This process is repeated until we find the j^{th} solution $U_j \le a_{j+1}$, $1 \le j \le d$. $U = U_j$ is the solution to equation (3.3).[29]

After computing the energy-cost values of all points on the interface, the heapsort technique is used to sort these values to find the point with minimum cost (thus the maximum local speed). The FMM program then accepts this point and calls *Gaussian 03* to compute the potentials at any neighboring points where the potential is unknown.

The FMM loop is then repeated, and the set of points for which the energy cost, U_j, is known systematically expands until the product is found.

Conceptually, we imagine slowly adding water to the reactant valley; the "water" level will keep going up, wetting the contours of the potential energy surface as it does so. Eventually the water level will rise to the level of the lowest-energy transition state, which is the lowest "mountain pass" for exiting the reactant valley. At this stage a narrow thread of water will follow the steepest-descent path to the bottom of the next valley. This process is mimicked by the FMM procedure. In FMM, the "energy cost" contours to record which portions of the potential energy surface are "flooded" at any given point in time.

Notice that only the "flooded" portion of the surface needs to be computed. This reduces the computational cost significantly. Moreover, the method is amenable to parallel computation: since the "beach" where the water meets the land expands in many

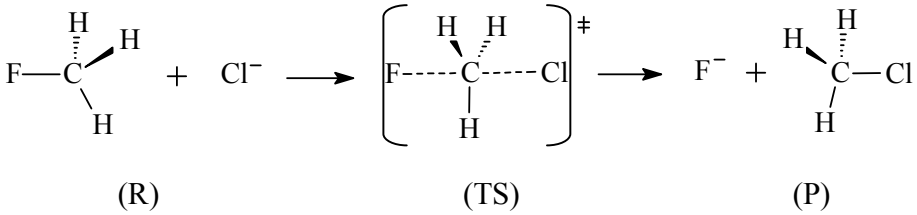
directions at once, many different potential energy calculations can be performed simultaneously.

3.4 Applications

To demonstrate our revised and extended FMM program, we will give some examples using a reduced 2-dimensional PES. Our FMM program is interfaced with *Gaussian 03*.[31] All *Gaussian* calculations were done using density-functional theory (BhandhLYP/6-311++G**).[32-36] The "half and half" hybrid functional has less self-interaction error than the more popular hybrid functionals (like B3LYP). This is believed to be important for modeling the transition state structures where the exchange-correlation hole is delocalized.[37-40] (The S_N2 reaction is a classic example of such a transition state.)

3.4.1 The S_N2 reaction

The mechanism of the S_N2 reaction has been studied intensively by experimental and theoretical methods, so it is a good test for our method. This is a one-step reaction, so we expect two minima (the reactant R and product P) and one transition state (TS) on the PES.



(R) (TS) (P) (3.5)

In this reaction, only C-F and C-Cl bonds are involved in bond-breaking and bond-forming, so the PES of this reaction can be modelled using a 2-dimensional reduced potential energy surface based on the C-F and C-Cl coordinates.[41] At each grid point, we will freeze the C-F and C-Cl bond length at the given values and minimize the energy with respect to the other coordinates. The 2-dimensional reduced PES and the MEP computed by the FMM program are depicted in Figure 3. 1. About 20% of grid points are in the "flooded" region and are computed by *Gaussian 03*. The energy-cost surface with the reactant (R) as starting point and the MEP found on this surface are shown in Figure 3. 2. Plotting the change in potential energy along the MEP gives the energy profile of the reaction coordinate (Figure 3. 3).

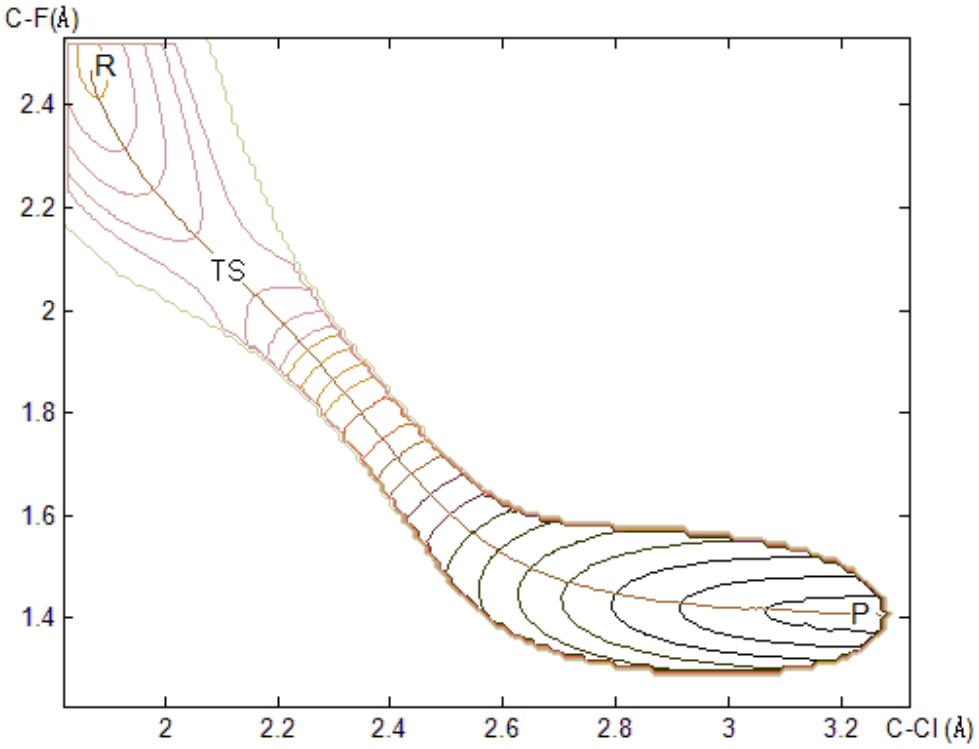


Figure 3. 1: The Potential Energy Surface of the S_N2 reaction based on C-F and C-Cl bond lengths. The grid sizes on both dimensions are dR = 0.01 Å. The calculation starts from the reactant (R), fills the reactant valley, breaches the reaction barrier at the transition state (TS), and then "flows" down to the product (P). The FMM program transforms this PES to an energy-cost surface (Figure 3. 2).

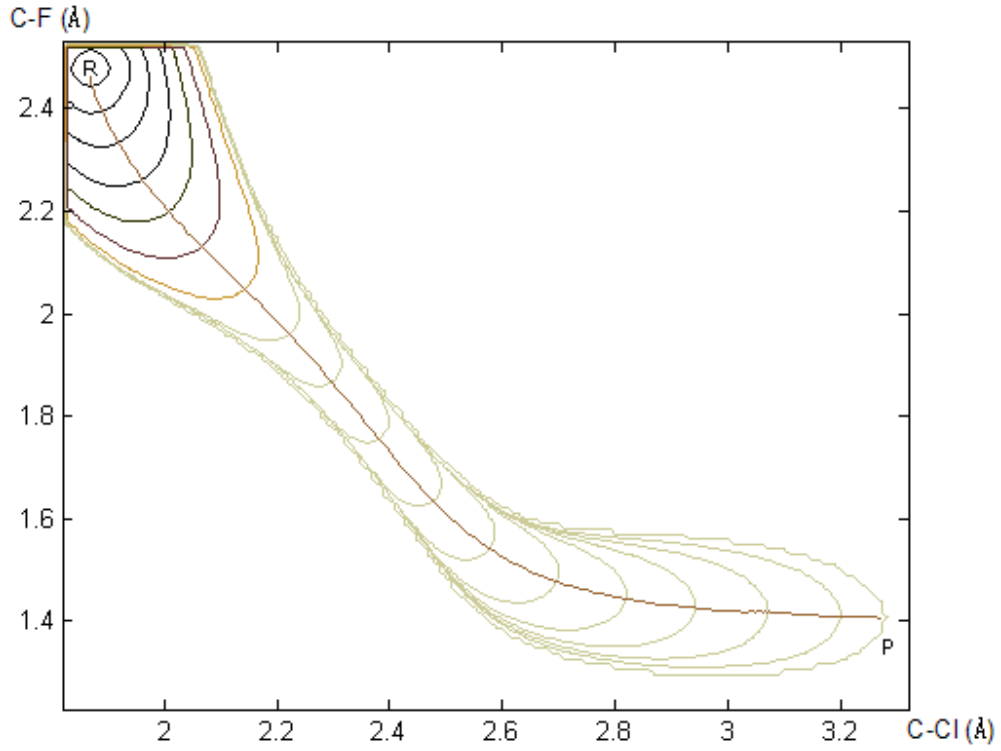


Figure 3. 2: The energy-cost surface transformed from the PES on Figure 3. 1. The MEP is determined by backtracing from the product to the reactant along the steepest descent path on the energy-cost surface.

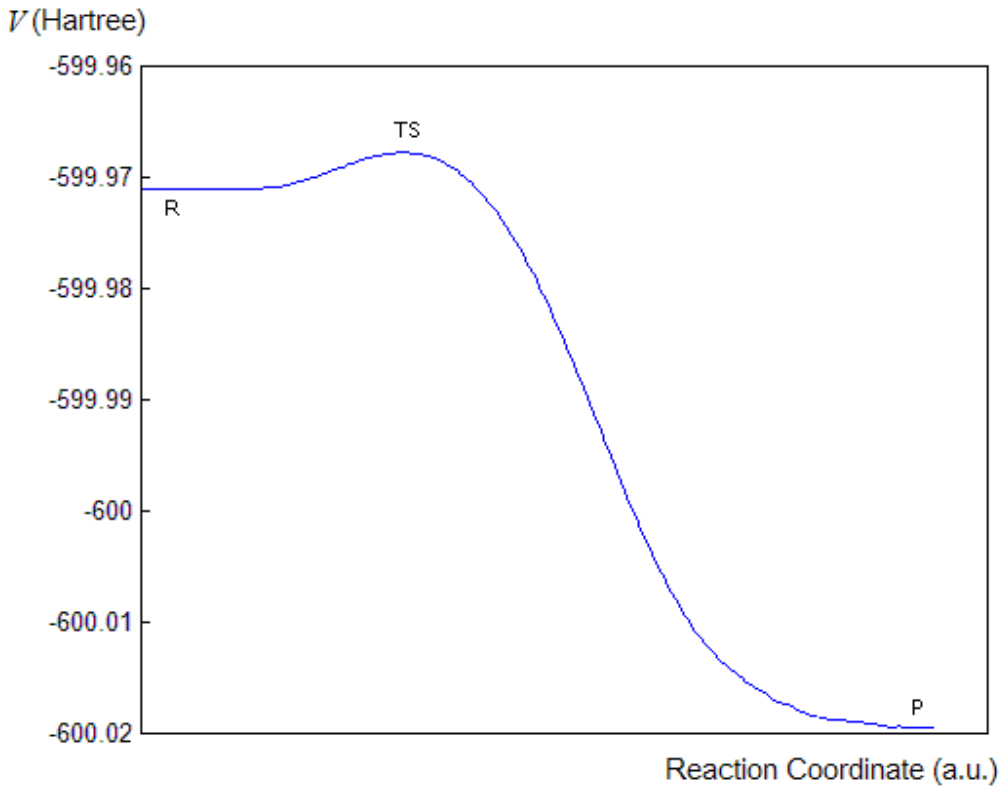
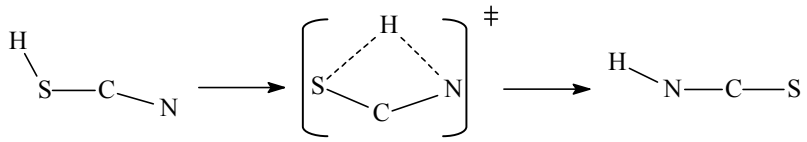


Figure 3. 3: The energy profile of the S_N2 reaction in Figure 3.1.

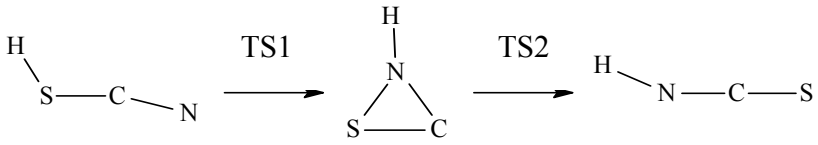
3.4.2 The isomerization of HSCN to HNCS

Wierzejewska and Moc reported 9 isomers of HSCN.[42] For the isomerization reaction of HSCN to HNCS, they proposed two competitive mechanisms. The first one is a one-step mechanism,



(3.6)

The second mechanism is a two-step mechanism with a ring structure intermediate,



(3.7)

According to Wierzejewska and Moc,[42] the first mechanism has the lower energy barrier. We will test this conclusion using the FMM program. The H-S and H-N bond lengths are chosen as the key coordinates. Figure 3. 4 shows the reduced 2-dimensional PES based on these two key coordinates and the MEP found on the energy-cost surface. The results are similar to those from previous studies, but the FMM program only computes about 30% of the grid points.

The energy profile of the reaction coordinate is shown in Figure 3. 5.

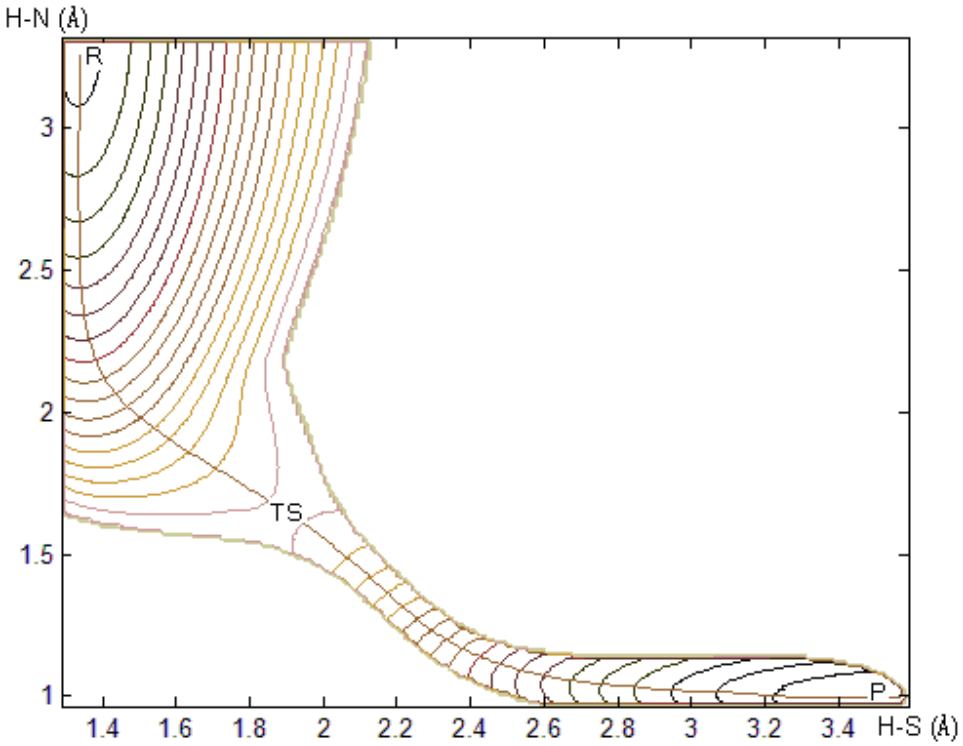


Figure 3. 4: The reduced 2-dimensional PES for the isomerization of HSCN to HNCS, based on H-S and H-N bond lengths.

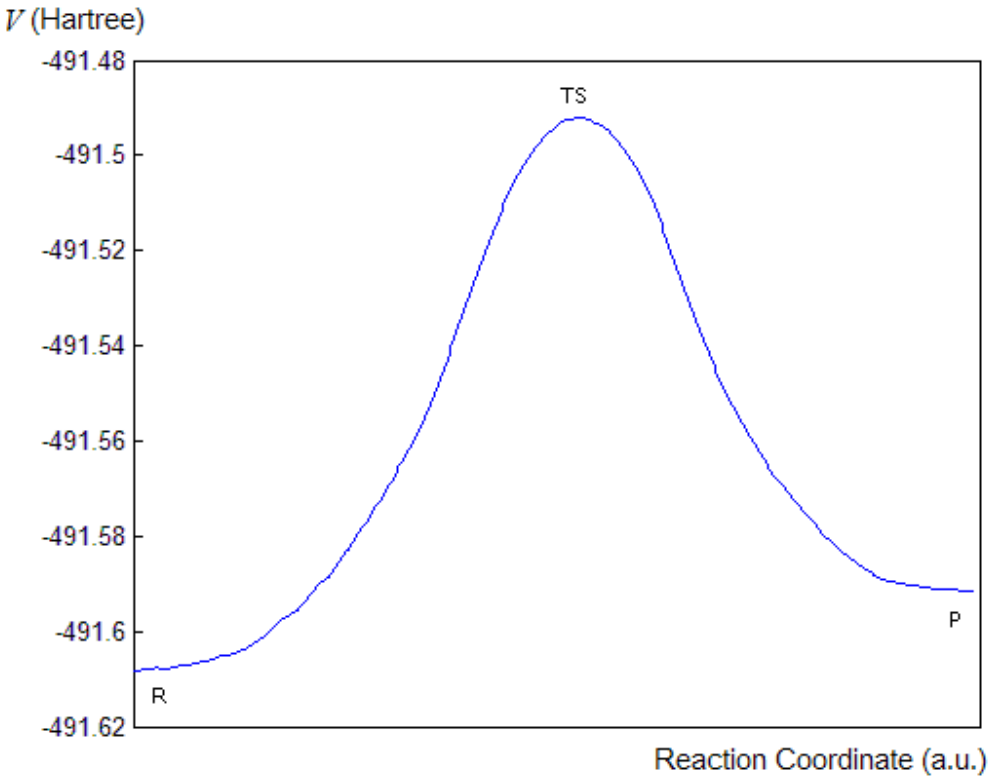


Figure 3. 5: The energy profile for the isomerization of HSCN to HNCS.

3.4.3 The dissociation of ionized O-methylhydroxylamine

One strength of the FMM approach is that it can be applied to any number of dimensions. The cost, of course, will grow exponentially with increasing dimensionality because of the increasing number of stationary points on the potential energy surface, and the results become increasing difficult to visualize as dimensionality increases.[43] As a simple example that is neither too expensive nor too difficult to visualize, we consider the dissociation of $[CH_5NO]^{+\bullet}$.

The PES of $[CH_5NO]^{+\bullet}$ has been studied using mass spectrometry and theoretical methods.[44] The following dissociation reaction has been observed,

$$[CH_5NO]^{+\bullet} \rightarrow [CH_2NH_2]^{+} + OH^{\bullet} \tag{3.8}$$

and the mechanism proposed for this dissociation reaction is [44]

$$CH_3\text{–}O\text{–}NH_2^{+\bullet} \rightarrow \dot{O}\text{–}\overset{+}{N}H_2\text{–}CH_3 \rightarrow H\dot{O}\text{–}NH_2\text{–}\overset{+}{C}H_2 \rightarrow H\dot{O} + NH_2\overset{+}{C}H_2 \tag{3.9}$$

We selected the bond lengths C-N, N-O and O-H as the key coordinates, then performed FMM on the 3-dimensional reduced PES defined by these points. The 3-dimensional equi-potential surfaces have onion-like structures. Each layer of the "onion" represents a certain value of the potential energy. Figure 3. 6 shows one layer of the "onion" with a potential value of -170.534 Hartrees. The cores of onions represent minima on the PES. We can see that there are 4 minima on this PES, the reactant (R), two intermediates (I, II), and the product (P). By examining the structures obtained, we

deduce that intermediate (I) and (II) coincide with $\dot{O}–\overset{+}{N}H_2–CH_3$ and $H\dot{O}–NH_2–\overset{+}{C}H_2$, respectively. The FMM calculation confirms that the mechanism in (3.9) is the minimum energy reaction pathway.

Figure 3. 7 shows the energy profile along the dissociation pathway in Rxn (3.9). This sort of several-step reaction is very difficult for most reaction-path methods, but FMM works just as well for this reaction as it does for the much simpler isomerization of HSCN.

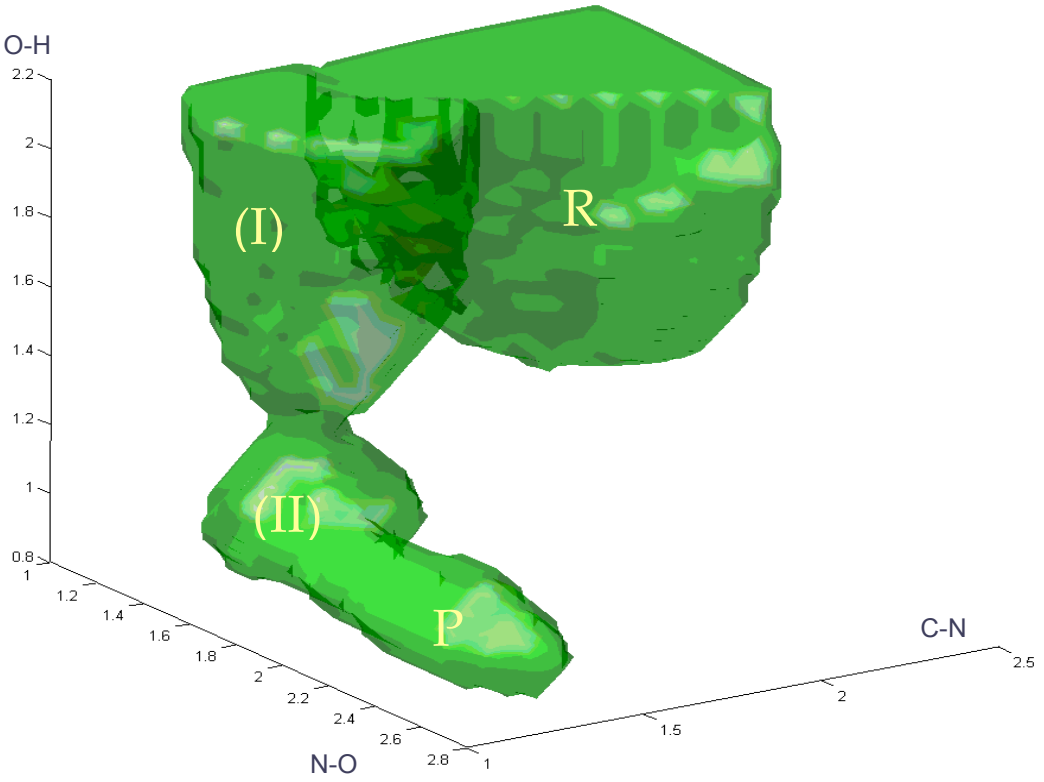


Figure 3. 6: The isosurface with a potential value of -170.534 Hartrees, which is one layer of the reduced 3-dimensional PES for the dissociation reaction of ionized O-methylhydroxylamine. The 3-dimensional equi-potential surfaces have an onion-like structure. Each layer of the "onion" represents a certain value of the potential energy. The cores of the "onions" are minima on the PES.

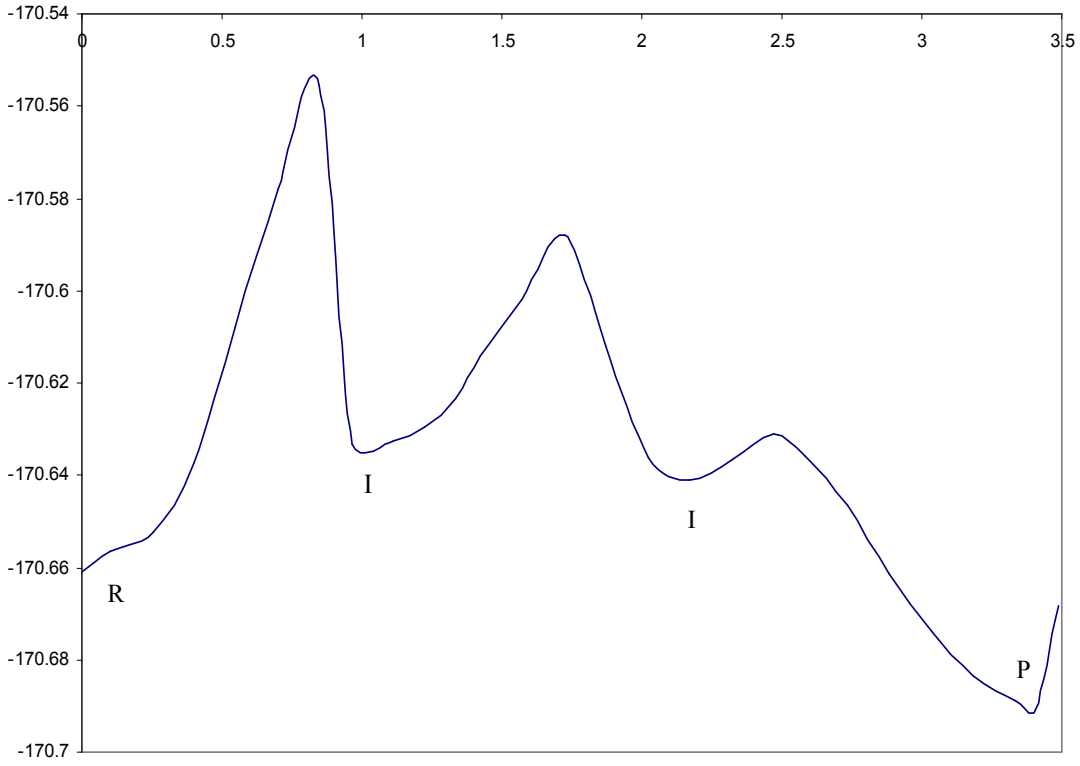


Figure 3. 7: The energy profile of the dissociation reaction of ionized O-methylhydroxylamine.

3.5 Conclusion

The fast marching method (FMM) is a very general method for finding the minimum energy path (MEP). Without any information about the mechanism in advance, it can find the minimum energy reaction path linking the reactant and any other point on the potential energy surface (PES). Due to its reliable and unbiased nature, the fast marching method can find the minimum energy path for any kind of chemical reactions. However, the computation cost of FMM is relatively high because a significant portion of the potential energy surface has to be computed in order to rigorously determine the MEP. This is especially true for 3 or higher dimensional PES.

The computation of points on the potential energy surface is several orders of magnitude more costly than solving the eikonal equation or performing the heap sort. Thus, before we make routine applications to higher-dimensional PES and more complicated molecules, we need to reduce the number of potential-energy computations that are required. Using a combination of moving-least-squares[45-49] and Shepard interpolation[50,51] reduces the number of potential energy computations that are required.[24] We can also reduce the cost of this method by performing potential energy computations all along the expanding front concurrently, with each point on the expanding front assigned to a different processor, and/or by parallelizing the constrained geometry optimizations that are needed to construct the reduced potential energy surfaces.[52] It is also useful to run the fast-marching method on a relatively course grid,

and then refine the estimates of the geometries and energies of the transition states using conventional methods. (One such method, recently developed in our group, exploits the same "reduced potential energy" structure as the underlying FMM approach.[53]) Finally, in cases where the full reaction path is not needed, and it suffices to only characterize the preferred mechanism and the transition state of the rate-limiting step, dual grid methods like the boundary-low-path method may be preferred.[30]

Reference List:

1. G. Mills and H. Jonsson, Phys. Rev. Lett. **72**, 1124 (1994).
2. G. Henkelman and H. Jonsson, J. Chem. Phys. **113**, 9978 (2000).
3. G. Henkelman, B. P. Uberuaga, and H. Jonsson, J. Chem. Phys. **113**, 9901 (2000).
4. B. Peters, A. Heyden, A. T. Bell, and A. Chakraborty, J. Chem. Phys. **120**, 7877 (2004).
5. S. K. Burger and W. T. Yang, J. Chem. Phys. **124**, 224102 (2006).
6. S. K. Burger and W. T. Yang, J. Chem. Phys. **124**, 054109 (2006).
7. W. N. E, W. Q. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
8. W. N. E, W. Q. Ren, and E. Vanden-Eijnden, J. Chem. Phys. **126**, 164103 (2007).
9. I. Berente and G. Naray-Szabo, J. Phys. Chem. A **110**, 772 (2006).
10. J. Nichols, H. Taylor, P. Schmidt, and J. Simons, J. Chem. Phys. **92**, 340 (1990).
11. J. Simons, P. Jorgensen, H. Taylor, and J. Ozment, J. Phys. Chem. **87**, 2745 (1983).
12. A. Banerjee, N. Adams, J. Simons, and R. Shepard, J. Phys. Chem. **89**, 52 (1985).
13. G. Henkelman and H. Jonsson, J. Chem. Phys. **111** (15), 7010 (1999).
14. C. Y. Peng and H. B. Schlegel, Isr. J. Chem. **33**, 449 (1993).
15. C. Y. Peng, P. Y. Ayala, H. B. Schlegel, and M. J. Frisch, J. Comput. Chem. **17**, 49 (1996).
16. K. Fukui, Acc. Chem. Res. **14** (12), 363 (1981).
17. C. Gonzalez and H. B. Schlegel, J. Chem. Phys. **90**, 2154 (1989).
18. C. Gonzalez and H. B. Schlegel, J. Phys. Chem. **94**, 5523 (1990).
19. S. K. Burger and W. T. Yang, J. Chem. Phys. **125**, 244108 (2006).
20. B. K. Dey, S. Bothwell, and P. W. Ayers, J. Math. Chem. **41**, 1 (2007).
21. B. K. Dey and P. W. Ayers, Mol. Phys. **105**, 71 (2007).
22. B. K. Dey and P. W. Ayers, Mol. Phys. **104**, 541 (2006).
23. B. K. Dey, M. R. Janicki, and P. W. Ayers, J. Chem. Phys. **121**, 6667 (2004).
24. S. K. Burger, Y. L. Liu, U. Sarkar, and P. W. Ayers, J. Chem. Phys. **130**, 024103 (2009).
25. Y. L. Liu, S. K. Burger, B. K. Dey, U. Sarkar, M. Janicki, and P. W. Ayers, in *Quantum Biochemistry*, edited by C. F. Matta (Wiley-VCH, Boston, 2010).
26. J. A. Sethian and A. Vladimirsky, Siam Journal on Numerical Analysis **41** (1), 325 (2003).
27. J. A. Sethian, SIAM Rev. **41** (2), 199 (1999).
28. J. A. Sethian, Proc. Natl. Acad. Sci. **93** (4), 1591 (1996).
29. H. K. Zhao, Mathematics of Computation **74** (250), 603 (2005).
30. S. K. Burger and P. W. Ayers, Journal of Chemical Theory and Computation **6**, 1490 (2010).

31. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Peersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, O. Salvetti, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, *Gaussian03, Revision C.02*. (Gaussian Inc., Wallingford, CT, 2004).
32. A. D. Becke, Phys. Rev. A **38**, 3098 (1988).
33. A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
34. A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
35. B. Miehlich, A. Savin, H. Stoll, and H. Preuss, Chem. Phys. Lett. **157** (3), 200 (1989).
36. C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
37. P. W. Ayers and W. Yang, in *Computational Medicinal Chemistry for Drug Discovery*, edited by P. Bultinck, H. de Winter, W. Langenaeker, and J. P. Tollenaere (Dekker, New York, 2003), pp. 571.
38. O. V. Gritsenko, B. Ensing, P. R. T. Schipper, and E. J. Baerends, J. Phys. Chem. A **104**, 8558 (2000).
39. Y. Zhang and W. Yang, J. Chem. Phys. **109**, 2604 (1998).
40. A. J. Cohen, P. Mori-Sanchez, and W. T. Yang, Science **321**, 792 (2008).
41. P. W. Ayers and R. G. Parr, J. Am. Chem. Soc. **123**, 2007 (2001).
42. M. Wierzejewska and J. Moc, J. Phys. Chem. A **107**, 11209 (2003).
43. F. H. Stillinger and T. A. Weber, Science **225** (4666), 983 (1984).
44. P. C. Burgers, C. Lifshitz, P. J. A. Ruttink, G. Schaftenaar, and J. K. Terlouw, Org. Mass Spectrom. **24** (8), 579 (1989).
45. T. Ishida and G. C. Schatz, J. Comput. Chem. **24** (9), 1077 (2003).
46. T. Ishida and G. C. Schatz, Chem. Phys. Lett. **314** (3-4), 369 (1999).
47. G. G. Maisuradze and D. L. Thompson, J. Phys. Chem. A **107**, 7118 (2003).
48. G. G. Maisuradze, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **119**, 10002 (2003).
49. G. G. Maisuradze, A. Kawano, D. L. Thompson, A. F. Wagner, and M. Minkoff, J. Chem. Phys. **121**, 10329 (2004).
50. D. L. Crittenden and M. J. T. Jordan, J. Chem. Phys. **122** (4) (2005).

51. M. A. Collins, Theor. Chem. Acc. **108**, 313 (2002).
52. S. K. Burger and P. W. Ayers, J. Chem. Phys. **133**, 034116 (2010).
53. S. K. Burger and P. W. Ayers, J. Chem. Phys. **132**, 234110 (2010).

Chapter 4:

Newton Trajectories for Finding Stationary Points on Molecular Potential Energy Surfaces*

* The content of this chapter is accepted for publication by J. Math. Chem: Yuli Liu; Steven K. Burger; Paul W. Ayers; Newton Trajectories for Finding Stationary Points on Molecular Potential Energy Surfaces.

4.1 Statement of the Problem

This chapter presents a new algorithm for computing Newton trajectories based on the Quadratic String Method (QSM) and explains how this can be used to find key stationary points on the molecular potential energy surface (PES). This method starts by using the intersections of Newton trajectories to locate stationary points on the PES. These points could then be used to determine the minimum energy path. The new method, called QSM-NT, is shown to be efficient and reliable for both analytical potential energy surfaces and potential energy surfaces computed from quantum chemistry calculations. The advantages and pitfalls of this method for exploring PES are discussed. In particular, the problem of discontinuous Newton trajectories is elucidated.

4.2 Introduction

Finding a reaction path that connects the reactant to the product via transition states and key intermediates on the molecular potential energy surface (PES) is the key to obtaining both quantitative data and qualitative understanding of chemical reaction mechanisms. From the molecular structures along the path, the qualitative features (e.g., the sequence in which bonds fracture and form) of the chemical reaction are clear. From the energy of the transition state and the relative energies of the reactants and products, key quantitative information about chemical kinetics and equilibria may be computed. It

is unsurprising, then, that computational chemists continue to devise novel algorithms for finding chemical reaction paths.[1] This paper is yet another contribution along these lines.

The distinguishing feature of this work is the focus on stationary points of the PES. Most of these methods focus on low-energy regions of the PES or directions on the PES with small force constants (large compliance). Chemically, however, the most important portions of PESs are the places where the gradient is zero: such stationary points are associated with reactants, products, intermediates, and transition states. Newton trajectories intersect at these locations, with the "density" of Newton trajectories being highest where the gradient of the potential energy is the smallest.[2-4] Using Newton trajectories to explore PESs, then, biases the search towards the most chemically relevant portions of the surface. This motivates the approach in this paper, where the quadratic string method[5] is used to compute Newton trajectories and the intersection of Newton trajectories is used to identify stationary points.

4.3 Background

4.3.1 Methods for Finding the Minimum Energy Path (MEP)

Most work on finding chemical reaction paths has focussed on the minimum energy path (MEP). The MEP is typically obtained by taking the union of the steepest-descent

paths from the transition state structure(s) to the reactant and product (and intermediate) structures. With rare exceptions, the "intrinsic reaction coordinate" so defined is equal to the MEP.[6]

There are generally two families of algorithms for finding the MEP: the surface-walking algorithms ("initial value" formulations; [7-15]) and the two-end algorithms ("boundary value" formulations; [5,16-29]). The two end methods require a good guess for the path linking the reactant and the product; otherwise a local MEP instead of the global MEP will be found. The surface-walking methods only need the reactant configuration, and then predict the products and the mechanism. Some surface-walking methods, like the fast marching method [8,9,12,30-34], guarantee that the global MEP will be found. Unfortunately, surface-walking algorithms usually are either very expensive or, if a heuristic is used to simplify the construction, they tend to be unreliable[24] for complicated systems. The two-end methods have great advantages in terms of computation costs and numerical stability.

The most efficient two-end methods include the string method (SM; [17,23]) and some improved string methods, such as the quadratic string method (QSM; [5]) and the growing string method (GSM; [24,35,36]). The string method divides the initial path into a certain number of nodes, then drives the nodes towards the MEP by a normal force along the corresponding hypersurface orthogonal to the path tangent at the node, reparameterizing the approximate path so that the nodes will be spaced evenly along the

path. The quadratic string method uses the local quadratic approximation for the hypersurfaces to reduce the number of calculations on the potential and gradient, which makes it more efficient and affordable for larger systems. Conceptually, string methods can be imagined as what would occur if one draped a pearl necklace over the PES; the pearls (nodes) would slip towards lower potential energy, but remain evenly spaced between reactant and product.

4.3.2 The Newton Trajectory (NT)

While the MEP is almost universally used to represent the reaction path, an alternative reaction path called the Newton trajectory (NT) has been proposed and explored by Quapp and coworkers.[2-4,35-47] They argue that the reaction path can be considered to be any curve that connects reactant to the product through the saddle point as long as the highest-energy point on the curve corresponds to the saddle point. The basic idea is that while the molecular structures of stationary points on the PES have chemical significance, intermediate structures are of superficial significance. (This contrasts somewhat with the common assertion that the MEP serves as a "leading line" about which reactive molecular dynamics trajectories are centered when the energy barrier is much higher than kT.[48])

A Newton trajectory is defined as a curve on which all gradients are pointing in the same (or opposite) direction. Each NT passes through all stationary points on the PES because when the magnitude of the gradient is zero, its direction is arbitrary. By the preceding argument, any NT that connects the reactant to the product through the transition state can be chosen as a reaction path. Notice that some NTs that can mimic these reaction paths, but have maxima that correspond to turning points in the trajectory, rather than stationary points on the PES.[49] The growing string method (GSM) is commonly used to find an appropriate NT.[35,36,44] Because only NTs without spurious turning points are acceptable reaction paths, it can be difficult to choose a gradient direction that defines an appropriate NT. Without prior knowledge of the PES, the fact that candidate NT-based reaction paths might have turning points means that each local maximum on a NT must be assessed to see if it is actually a saddle point.

In this paper we propose a new way to find the stationary points on the PES using NTs that avoids the "turning point problem" associated with the GSM-NT method. Specifically, since a complete NT passes all stationary points on the PES, the intersections of two NTs locate all the stationary points on a PES. We also provide a variant of QSM to find NTs; this reduces the computational cost of this approach significantly. To test our methods, we consider analytical PESs. With their validity established, we interfaced our approach to the *Gaussian* program and characterized an S_N2 reaction.

The primary advantage of this approach over more conventional approaches is that no prior knowledge of the PES is required and the computational effort is concentrated in regions of the PES where the NTs are close together—that is, the chemically important regions associated with stationary points. Unlike the fast marching method and most other "surface walking" approaches, all possible transition states, and therefore all possible reaction pathways, are found. This is particularly important where there are several competing reaction mechanisms.

4.4 Mathematical Definitions and Algorithms

4.4.1 Quadratic String Method (QSM)

The Minimum Energy Path (MEP) is defined as the steepest-descent path (SDP) from the transition states (1st order saddle points) to their adjacent minima. The SDP, $\mathbf{x}(t)$, can be obtained by solving the following differential equation,

$$\frac{d\mathbf{x}(t)}{dt} = -\mathbf{g} = -\nabla V\left(\mathbf{x}(t)\right), \tag{4.1}$$

which indicates that the tangent of the SDP is always directed against the gradient of the potential $\mathbf{g} \equiv \nabla V\left(\mathbf{x}(t)\right)$.

To simplify the equations, we introduce the projection operators proposed by Quapp.[39] For a unit vector $\hat{\mathbf{u}}$, the dyadic product, $\mathbf{D}_{\hat{\mathbf{u}}} = \hat{\mathbf{u}}\hat{\mathbf{u}}^T$ projects a vector in the

direction defined by $\hat{\mathbf{u}}$. The projection operator $\mathbf{P}_{\hat{\mathbf{u}}} = \mathbf{I} - \mathbf{D}_{\hat{\mathbf{u}}}$, projects a vector into the hyperplane perpendicular to $\hat{\mathbf{u}}$.

Parameterize the path $\mathbf{x}(t)$ by the arc length s. Then $d\mathbf{x}(t)/ds$ is the normalized tangent of the path. We denote the normalized tangent of the path as $\hat{\mathbf{o}}(\mathbf{x})$. Then, from Eq. (4.1),

$$\hat{\mathbf{o}}(\mathbf{x}) = -\frac{\mathbf{g}}{\|\mathbf{g}\|}. \tag{4.2}$$

For a point on the steepest descent path, the gradient and $\hat{\mathbf{o}}$ are both tangent to the SDP, so $\mathbf{P}_{\hat{\mathbf{o}}}\mathbf{g} = \mathbf{0}$. Away from the optimal path, however, there is an component of the gradient in the hyperplane orthogonal to the SDP,

$$\mathbf{g}_{\perp} = \mathbf{P}_{\hat{\mathbf{o}}}\mathbf{g}. \tag{4.3}$$

Minimizing $\|\mathbf{g}_{\perp}\|$ leads to the SDP; this transforms the differential equation in Eq. (4.1) into a minimization problem. String methods work by solving the minimization problem.

The basic string-method algorithm is:

1. An initial guess of the reaction path is given.
2. The initial path is discretized as several nodes.
3. The energies and gradients of all nodes are evaluated.
4. For each node, *i*, on the path,

a. The tangent to the guessed reaction path, $\hat{\mathbf{o}}_i$ is calculated.

b. The orthogonal projection is performed, giving $\mathbf{g}_{\perp,i} = \mathbf{P}_{\hat{\mathbf{o}}_i}\mathbf{g}_i$.

c. Search in the direction $\mathbf{d}_i = -\mathbf{g}_{\perp,i}$ to minimize the value of $\left\|\mathbf{g}_{\perp,i}\right\|$ on the hyperplane orthogonal to the guessed path. This defines the updated position for node *i*.

5. If necessary, the nodes are redistributed so that they are relatively evenly spaced along the path.

6. If $\left\|\mathbf{g}_{\perp,i}\right\|$ is sufficiently small at each node, then a sufficiently accurate approximation to the SDP has been found. Otherwise, return to step 3.

The quadratic string method (QSM) uses the same algorithmic structure as the string method.[5] The primary difference is that instead of minimizing the projection of the gradient on the hyperplane perpendicular to the path (step 4c), the minimization occurs on a quadratic hypersurface that is tangent to the hyperplane,

$$\mathbf{d}_i(\mathbf{x}) = -\mathbf{P}_{\hat{\mathbf{o}}_i}(\mathbf{g}_i^0 + \mathbf{H}_i(\mathbf{x} - \mathbf{x}_i^0)) \tag{4.4}$$

The hypersurface is obtained using a quasi-Newton Hessian constructed from a variable step-size Runge-Kutta method. For further details, see ref. [5]. The QSM converges superlinearly and requires fewer energy/gradient calculations than simple string methods.

4.4.2 Newton Trajectories (NTs)

A NT follows a curve on which the gradients point to the same (or opposite) direction. This direction is called the searching direction of the NT. For a given searching direction $\hat{\mathbf{u}}$, the corresponding NT can be formulated as a minimization problem: minimize $\|\mathbf{g}_\perp\|$, where

$$\mathbf{g}_\perp = \mathbf{P}_{\hat{\mathbf{u}}}\mathbf{g}. \tag{4.5}$$

This problem is mathematically similar to finding the SDP; the only change is that now we minimize on a hyperplane (or hypersurface) that is orthogonal to the searching direction $\hat{\mathbf{u}}$, rather than to the tangent vector of the guessed SDP, $\hat{\mathbf{o}}_i$. Notice that $\hat{\mathbf{u}}$ does not change along the reaction path; in this sense finding NTs is even simpler than finding the SDP.

It is easy to modify the QSM algorithm to find NTs; one merely fixes the searching direction, replacing $\mathbf{P}_{\hat{\mathbf{o}}_i}$ with $\mathbf{P}_{\hat{\mathbf{u}}}$ in step 4 and in Eq. (4.4). We call this algorithm QSM-NT.

The NTs from all possible search directions intersect at each stationary point on the PES. The next step in the procedure is to choose two different search directions and then compute the associated NTs. These NTs intersect at stationary points (minima, maxima, and saddle points) on the PES. An NT is said to be complete if it passes through every stationary point on the PES. Unfortunately, not all NTs are complete; some NTs are

discontinuous. Most path-finding methods, including QSM, only find part of the NT in the discontinuous case. In such cases, not every stationary point is found. As an initial guess for the path in QSM, we usually use the straight line linking the reactant and product. This setup does not guarantee the exploration of all stationary points, but it maximizes the possibility of finding all the stationary points in the "interesting" area of the PES that is positioned between the reactant and the product.

4.5 Applications

The QSM-NT method has been applied to the analytical PESs: the 4-well potential, the Müller–Brown potential; and a simple 1-step chemical reaction: the S_N2 reaction. For each PES, two NTs with different searching directions were found and plotted on the PES. We can clearly see that the intersections of the two NTs coincide with the reactive intermediate(s) and transition states (Figure 4.1, Figure 4.2, and Figure 4. 3).

4.5.1 Müller-Brown PES

To apply the QSM-NT method, we need to use two searching directions. The first searching direction, shown as the black path in Figure 4.1 is the vector linking the two ends of the path: [-0.8, 1.0]. The gradient of the potential at each node in the QSM algorithm points either towards [-0.8, 1.0] or in the opposite direction [0.8, -1.0]. For the

second searching direction, shown as the red path in Figure 4.1, we chose [-0.2, 1.0]. These two NTs cross at 5 points: the reactant (R), transition state 1 (TS1), intermediate (Int), transition state 2 (TS2) and the product (P).

The form of the Müller-Brown PES is given in Appendix 1.[50]

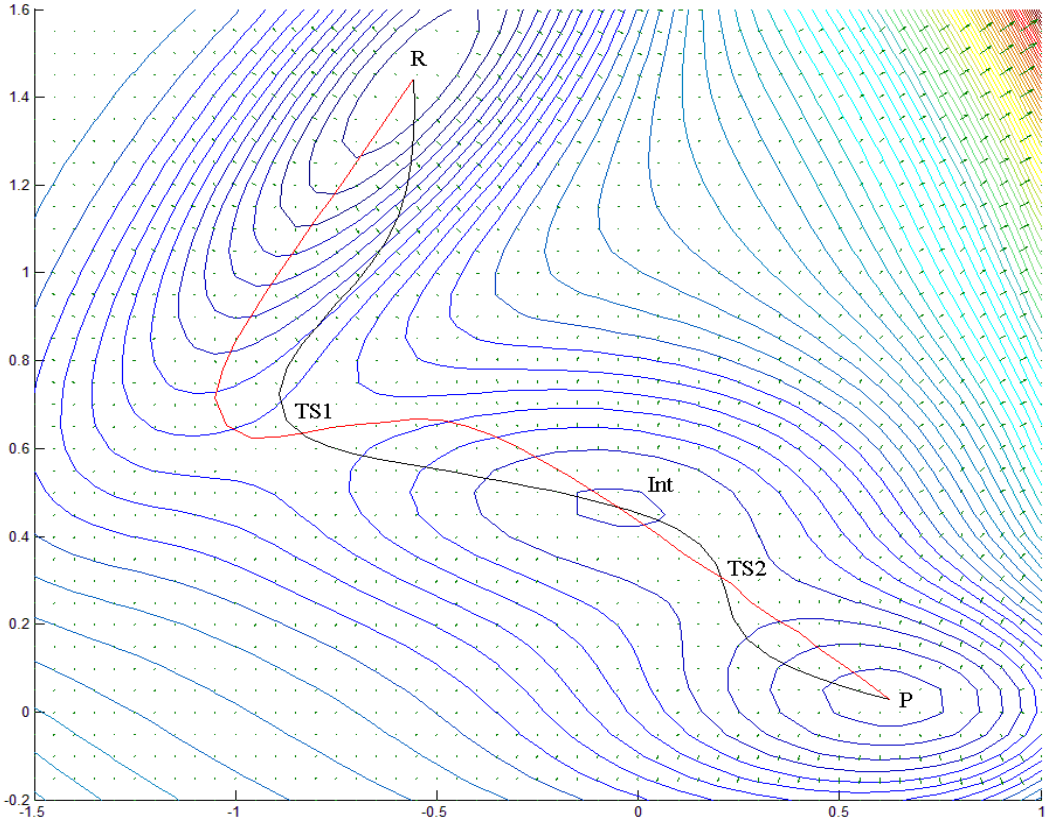


Figure 4. 1: **Newton trajectories on the Müller–Brown potential. The gradient directions of the two Newton trajectories are: black [-0.8, 1.0] and red [-0.2, 1.0], respectively. The arrows represent the directions of the gradients on the grids of the PES, which are orthogonal to the contours. The intersections of these two NTs accurately show the stationary points (the intermediate and the transition states) linking the reactant and the product.**

4.5.2 The 4-well PES

The 4-well PES is used to explore the case where there are two low-energy pathways between the reactant and the product structure; results are shown in Figure 4.2. The searching for the black path is the vector linking the two ends of the path: [1.1, 1.0]. The searching direction of the red path is [1.9, 1.0]. These two NTs cross at 5 points: the reactant (R), transition state 1 (TS1), intermediate (Int), transition state 2 (TS2) and the product (P).

The form of the 4-well PES is given in Appendix 2.[30]

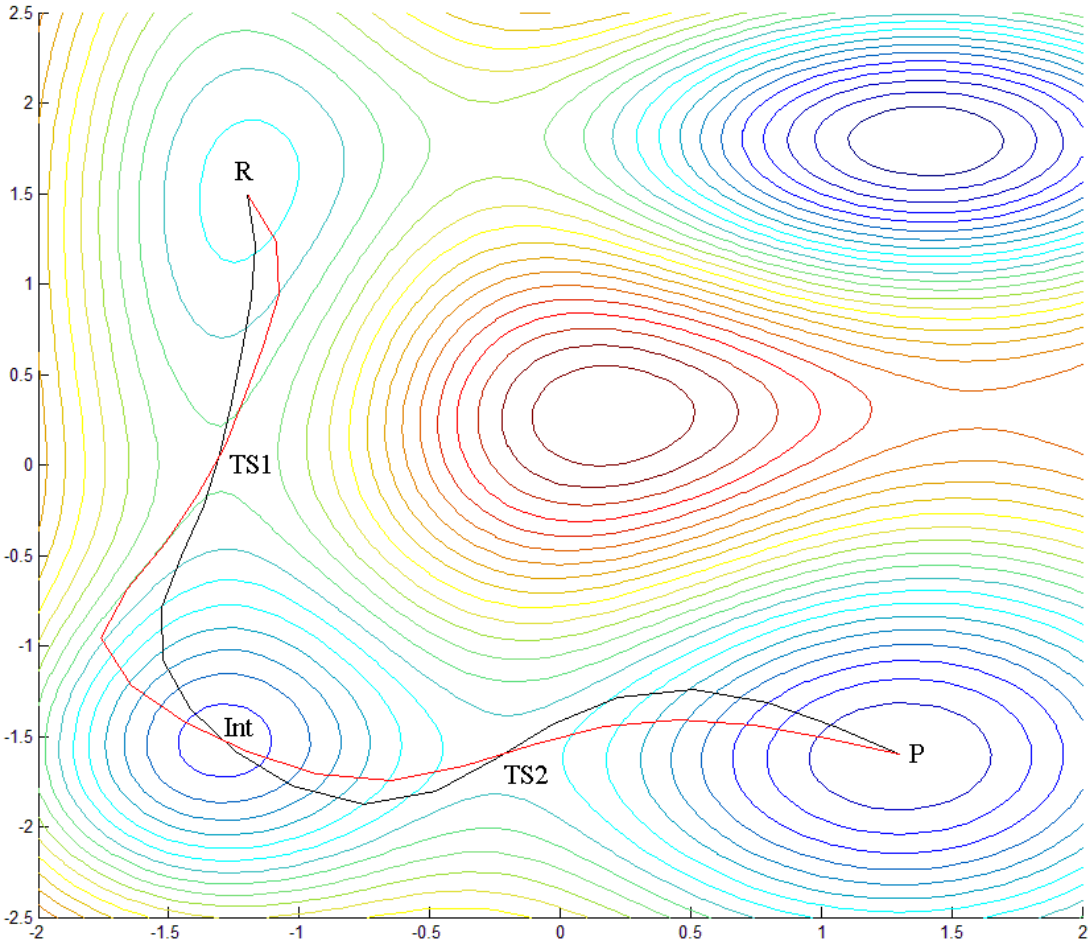
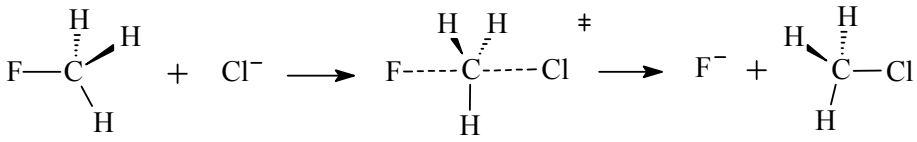


Figure 4. 2: **Newton trajectories on the 4-well potential. The gradient directions of the two NTs are: black: [1.1, 1.0]; red: [1.9, 1.0]. The intersections of the two NTs clearly show the stationary points (transition states and intermediate) linking the reactant and the product structures.**

4.5.3 The S_N2 reaction

Encouraged by the favourable results for the aforementioned analytical PES, we built an interface between the QSM-NT program and *Gaussian 03*.[51] We then tested the method on the following gas-phase S_N2 reaction,



(4.6)

All potential energy and derivative calculations are done using the BhandhLYP/6-311++G** level of theory.[52,53] For transition states, the larger amount of exact exchange in the Bhandh exchange functional, compared to the conventional hybrid functionals, is usually preferred.

We focused on the lengths of the C–F and C–Cl bonds that are formed/broken in the reaction. For each choice of these bond lengths, all of the other internal coordinates are minimized, so the only nonzero components of the gradient are in these directions. This corresponds to choosing searching directions of the form [$R_{C–F}$, $R_{C–Cl}$, 0, 0, 0, …0]. For simplicity, the zero-gradient directions will not be shown in what follows.

The searching direction of the first NT (the black path in Figure 4.3) is [1.3, 1.0], the vector linking the reactant and the product. The searching direction of the red NT is [0.4, 1.0]. (The second direction was chosen to resemble the first one, because search directions that are less similar can lead problems with discontinuous trajectories.) As

expected, the two NTs cross at the transition state (Figure 4. 3). Even without any previous knowledge about the PES of the S_N2 reaction, the transition state can still be located by finding the crossing point of two NTs.

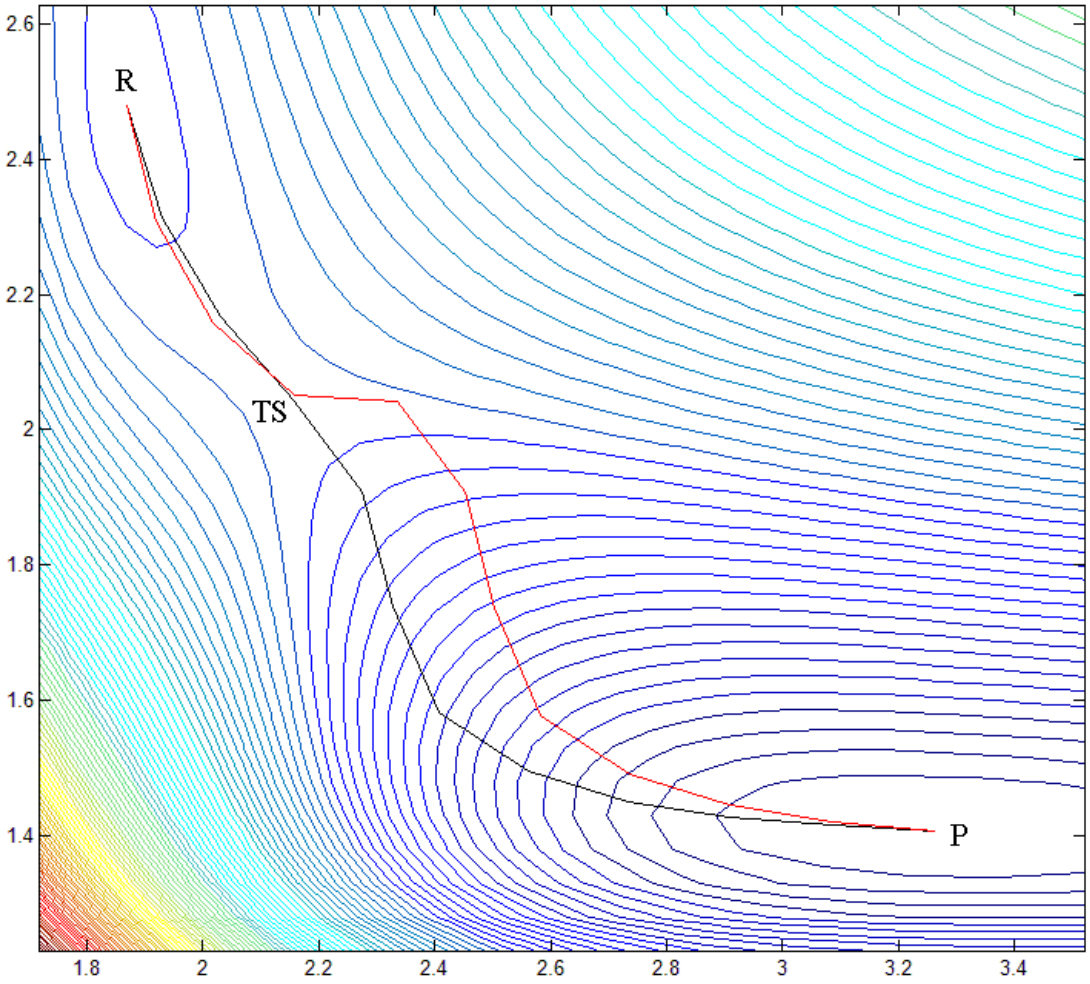


Figure 4. 3: **Newton trajectories for the S_N2 reaction. The gradients of the two NTs are: black: [-1.3, 1.0]; red: [-0.4, 1.0]. The NTs cross at the transition state (TS) linking the reactant and the product.**

4.6 Difficulties

In favourable cases, the QSM-NT method converges to a NT that passes through all the stationary points of interest. The intersection points of two such NTs locates all the stationary points on the PES, and from that point more conventional tools for analyzing PES suffice. In some cases, however, the QSM-NT method fails.

4.6.1 Discontinuous trajectories

For some searching direction, the NT is composed of two or more branches; such NTs are said to be discontinuous. For example the NT in Figure 4. 4 (searching direction [-1.75, 1.0]) is composed of two branches: branch 1 crosses three minima (M1, M2, M3) and two saddle points (S1, S2); branch 2 is a closed curve that crosses one minimum (M4), the maximum (Max), and two saddle points (S3, S4).

A two-end path-finding method like QSM finds only branch 1 if M1 and M3 are fixed as the reactant and the product structures. If one attempts to explore both branches at the same time by setting the endpoints to M3 and M4, then the QSM calculation never converges because there is no continuous path that meets those boundary conditions. The result of the (nonconverging) QSM calculation is shown in Figure 4. 4: it traces small portions of the lower branch of the NT and then jumps, as discontinuously as it can given the algorithmic constraints of the QSM method, to the upper branch of the NT. The

stationary points for the 4-well potential could be located (cf. Figure 4.2) only by choosing the second searching direction to be close to the first one, so that the two NTs branched in a similar fashion. The alternative MEP, corresponding to the reaction mechanism M1-S4-M4-S3-M3, can be located, but requires looking for NTs with very different directions, like [2.0,1.0]. Such NTs intersect the NTs in Figure 4.2 only at the reactant and product structures. If one wished to find a path from M2 to M4, then one would need to adjust the searching direction. Moreover, it is likely that only one of the two possible MEPs can be found: either M2-S1-M1-S4-M4 or M2-S2-M3-S3-M4.

A singular NT passes and branches on the valley ridge inflection point on the PES,[54] so a valley ridge inflection point might be confused with intersections of two NTs. But different branches of one NT and different NTs can be easily distinguished by examining the searching directions. Thus there is no danger of confusing valley ridge inflection points with stationary points on the PES. However, because of discontinuous trajectories, in order to find all the stationary points on a PES, one must vary not only the searching directions for the NTs, but consider several different choices for the path endpoints.

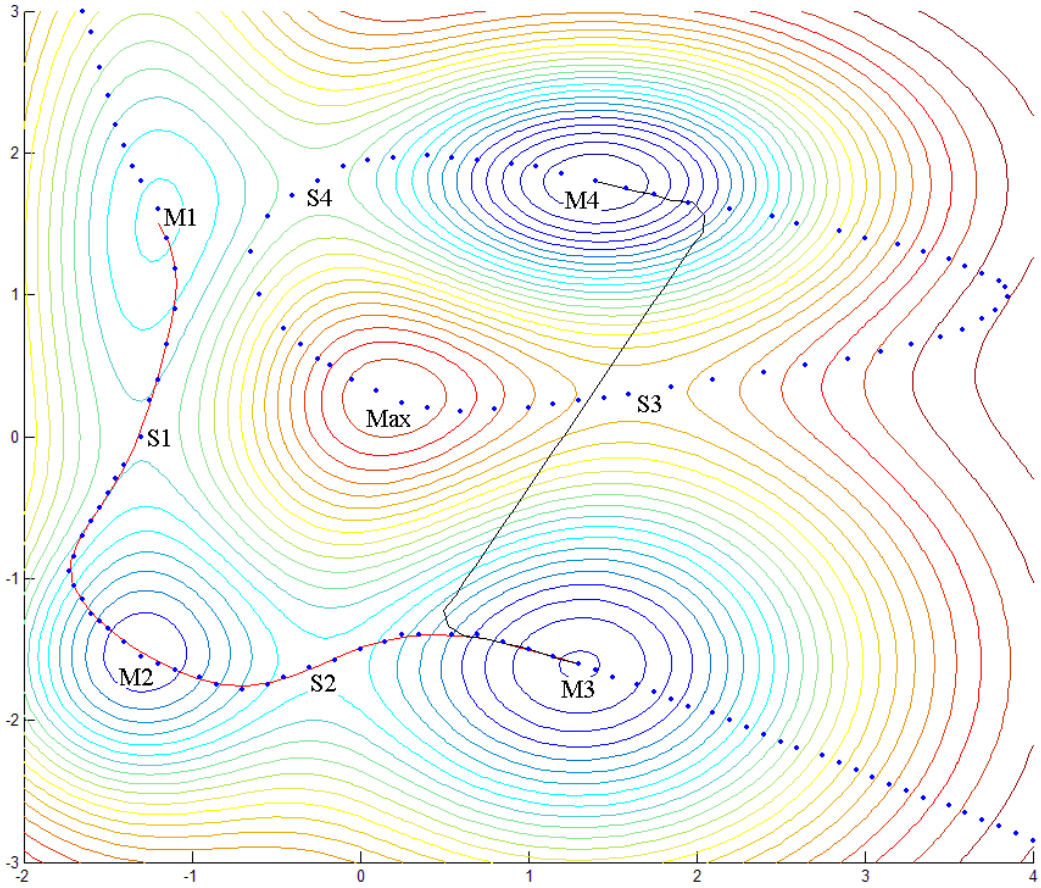


Figure 4. 4: **The dotted curves denote the discontinuous Newton trajectory on the 4-well PES found using Matlab 7.0.1; the searching direction of this NT is [-1.75, 1.0]. The solid curves correspond to the same searching direction, but the NT was found using QSM, choosing the endpoints as either M1 and M3 (the red curve on the left) or M3 and M4 (the black curve on the right). Since the NT with this searching direction is discontinuous, the QSM-NT method does not converge in the latter case.**

4.6.2 Multiple minima of $\|\mathbf{g}_\perp\|$ on the hypersurface

Recall, from section 3, that the NT is found by minimizing $\|\mathbf{g}_\perp\|$ on the hypersurface orthogonal to the searching direction. When the hypersurface is tangent to a contour line of the potential, the gradient $\mathbf{g}$ is orthogonal to the hypersurface and $\|\mathbf{g}_\perp\| = 0$.

In some cases, the hypersurface coincides with the tangent of potential contours from two potential wells. An example for the Müller-Brown potential is shown in Figure 4. 5. When the searching direction is [0.0, 1.0] (vertical), the hypersurfaces are horizontal (the blue and yellow lines in Figure 4. 5). Hypersurfaces between the two yellow lines are tangent to contour-lines of the PES in two places, so there are two places where $\|\mathbf{g}_\perp\| = 0$ on these hypersurfaces. Two-end path finding method like QSM can only find the minimum closest to the initial path, which leads to an incorrect path (black solid line in Figure 4. 5).

While the problem of discontinuous NTs is inherent in the definition of the NT, the problems we encounter when $\|\mathbf{g}_\perp\|$ has multiple minima on the hypersurface orthogonal to the searching direction is a consequence of using a two-end algorithm to find the NT. Solving the differential equation directly avoids this problem. Using a growing string algorithm, with one free end, also solves this problem.

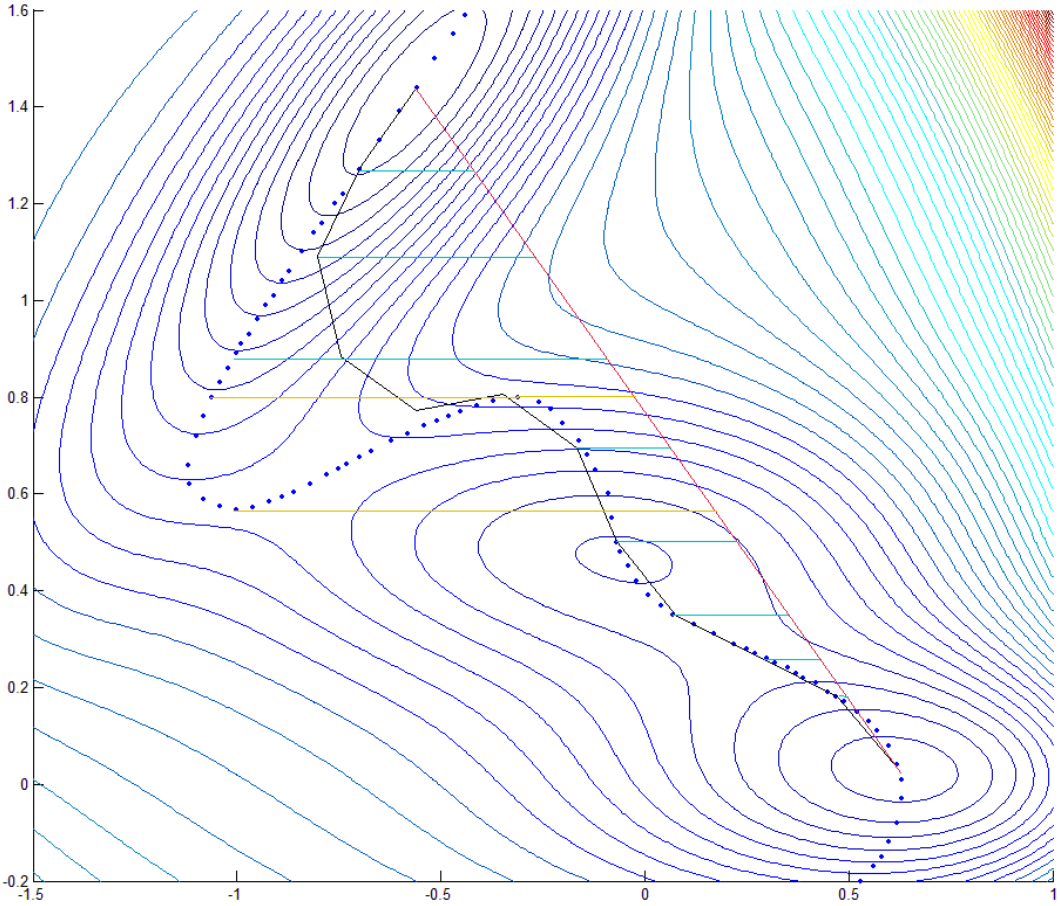


Figure 4. 5: **A case where QSM-NT method fails to find the correct Newton trajectory. The searching direction is [0.0, 1.0]. The dotted path is the complete NT found by Matlab7.0.1. The black solid line is the trajectory found using the modified QSM program. The red straight line is the initial path. The light blue lines show the hypersurfaces orthogonal to the searching direction, which is also the moving direction of the nodes during minimization.**

4.7 Conclusion

The QSM-NT method is a promising method for finding all stationary points on the PES, accordingly all alternative reaction paths linking the reactant and the product. Unlike two-end methods (which only find a single local minimum energy path) or the fast-marching method (which only finds the global minimum energy path), the basic approach pursued here can, in principle, find all possible reaction pathways. One of the most appealing features of this approach is that the density of the NTs is highest near the stationary points on the PES: the computational effort therefore concentrated in the most chemically interesting regions of the PES.

In our QSM-NT method, a modified version of the QSM program is used to find NTs for two different searching directions. The intersections between these NTs are stationary points on the PES. While the method does not always work (see section 5), when it does work, it is computationally efficient. The main problem is that of discontinuous, or incomplete, Newton trajectories. Because not every NT is complete, finding all the stationary points on the PES requires, in general, not only considering several different searching directions, but also several different choices for the path endpoints.

The computational cost of the QSM-NT method is nearly twice of QSM, or more, if more than two NTs are required to locate all the stationary points than are needed to characterize a chemical process. Like other two-end methods, however, conventional

QSM can find only the closest local MEP to the initial path, and cannot find the global MEP or other alternative paths. However, with proper setup and followed by further analysis, the QSM-NT method can find all stationary points on the PES, and therefore all reaction paths; this can reveal alternative reaction mechanisms.

Appendix 1. The Müller-Brown Potential

The Müller–Brown PES is defined by the following function,[50]

$$V(x,y)=\sum_{i=1}^{4} d(i)\cdot\exp\left[a(i)(x-x_0(i))^2+b(i)(x-x_0(i))(y-y_0(i))+c(i)(y-y_0(i))^2\right] \tag{4.7}$$

Table 4. 1: Parameters for the Müller-Brown potential.

i	1	2	3	4
$a(i)$ (Å)	-1.0	-1.0	-6.5	0.7
$b(i)$ (Å)	0.0	0.0	11.0	0.6
$c(i)$ (Å)	-10.0	-10.0	-6.5	0.7
$x_0(i)$ (Å)	1.0	0.0	-0.5	-1.0
$y_0(i)$	0.0	0.5	1.5	1.0
$d(i)$ (kcal/mol)	-200.0	-100.0	-170.0	15.0

Appendix 2. The 4-well Potential

The 4-well PES is defined by the following function,[30]

$$V(x,y) = V_0 + \sum_{i=0}^{4} d(i) e^{-a(i)(x-x_0(i))^2 - b(i)(y-y_0(i))^2} \tag{4.8}$$

Table 4. 2: Parameters for the 4-well potential.

i	0	1	2	3	4
V_0 (kcal/mol)			5.0		
$d(i)$ (kcal/mol)	0.6	3.0	1.5	3.2	2.0
$a(i)$ (Å^{-2})	1.0	0.3	1.0	0.4	1.0
$b(i)$ (Å^{-2})	1.0	0.4	1.0	1.0	0.1
$x_0(i)$ (Å)	0.1	1.3	-1.5	1.4	-1.3
$y_0(i)$ (Å)	0.1	-1.6	-1.7	1.8	1.23

Reference List

1. H. B. Schlegel, J. Comput. Chem. **24**, 1514 (2003).
2. W. Quapp, M. Hirsch, O. Imig, and D. Heidrich, J. Comput. Chem. **19**, 1087 (1998).
3. W. Quapp, J. Math. Chem. **36**, 365 (2004).
4. W. Quapp, Theor. Chem. Acc. **121**, 227 (2008).
5. S. K. Burger and W. T. Yang, J. Chem. Phys. **124**, 054109 (2006).
6. K. Fukui, Acc. Chem. Res. **14** (12), 363 (1981).
7. K. K. Irikura and R. D. Johnson, J. Phys. Chem. A **104**, 2191 (2000).
8. B. K. Dey and P. W. Ayers, Mol. Phys. **104**, 541 (2006).
9. Y. L. Liu, S. K. Burger, B. K. Dey, U. Sarkar, M. Janicki, and P. W. Ayers, in *Quantum Biochemistry*, edited by C. F. Matta (Wiley-VCH, Boston, 2010).
10. K. Ohno and S. Maeda, Chem. Phys. Lett. **384** (4-6), 277 (2004).
11. S. Maeda, Y. Watanabe, and K. Ohno, Chem. Phys. Lett. **414** (4-6), 265 (2005).
12. S. K. Burger and P. W. Ayers, Journal of Chemical Theory and Computation **6**, 1490 (2010).
13. K. Ohno and S. Maeda, Phys. Scr. **78**, 058122 (2008).
14. S. Maeda and K. Ohno, J. Phys. Chem. A **109**, 5742 (2005).
15. K. Ohno and S. Maeda, J. Phys. Chem. A **110**, 8933 (2006).
16. W. N. E, W. Q. Ren, and E. Vanden-Eijnden, J. Chem. Phys. **126**, 164103 (2007).
17. W. N. E, W. Q. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
18. L. Xie, H. Y. Liu, and W. T. Yang, J. Chem. Phys. **120**, 8039 (2004).
19. P. Maragakis, S. A. Andreev, Y. Brumer, D. R. Reichman, and E. Kaxiras, J. Chem. Phys. **117**, 4651 (2002).
20. G. Henkelman, B. P. Uberuaga, and H. Jonsson, J. Chem. Phys. **113**, 9901 (2000).
21. S. A. Trygubenko and D. J. Wales, J. Chem. Phys. **120**, 2082 (2004).
22. J. M. Carr, S. A. Trygubenko, and D. J. Wales, J. Chem. Phys. **122**, 234903 (2005).
23. E. Weinan, W. Q. Ren, and E. Vanden-Eijnden, J. Phys. Chem. B **109**, 6688 (2005).
24. B. Peters, A. Heyden, A. T. Bell, and A. Chakraborty, J. Chem. Phys. **120**, 7877 (2004).
25. G. Henkelman and H. Jonsson, J. Chem. Phys. **113**, 9978 (2000).
26. D. Sheppard, R. Terrell, and G. Henkelman, J. Chem. Phys. **128**, 134106 (2008).
27. G. Mills and H. Jonsson, Phys. Rev. Lett. **72**, 1124 (1994).

28. G. Mills, H. Jonsson, and G. K. Schenter, Surf. Sci. **324**, 305 (1995).
29. J. W. Chu, B. L. Trout, and B. R. Brooks, J. Chem. Phys. **119**, 12708 (2003).
30. B. K. Dey, M. R. Janicki, and P. W. Ayers, J. Chem. Phys. **121**, 6667 (2004).
31. S. K. Burger, Y. L. Liu, U. Sarkar, and P. W. Ayers, J. Chem. Phys. **130**, 024103 (2009).
32. S. K. Burger and P. W. Ayers, J. Chem. Phys. **132**, 234110 (2010).
33. B. K. Dey, S. Bothwell, and P. W. Ayers, J. Math. Chem. **41**, 1 (2007).
34. B. K. Dey and P. W. Ayers, Mol. Phys. **105**, 71 (2007).
35. W. Quapp, J. Chem. Phys. **122**, 174106 (2005).
36. W. Quapp, Journal of Theoretical & Computational Chemistry **8**, 101 (2009).
37. W. Quapp and D. Heidrich, Journal of Molecular Structure-Theochem **585**, 105 (2002).
38. W. Quapp, Journal of Theoretical & Computational Chemistry **2**, 385 (2003).
39. W. Quapp, J. Comput. Chem. **25**, 1277 (2004).
40. M. Hirsch and W. Quapp, Chem. Phys. Lett. **395**, 150 (2004).
41. M. Hirsch and W. Quapp, Journal of Molecular Structure-Theochem **683**, 1 (2004).
42. M. Hirsch and W. Quapp, J. Math. Chem. **36**, 307 (2004).
43. M. Hirsch and W. Quapp, Theor. Chem. Acc. **113**, 58 (2005).
44. W. Quapp, J. Comput. Chem. **28**, 1834 (2007).
45. W. Quapp, E. Kraka, and D. Cremer, J. Phys. Chem. A **111**, 11287 (2007).
46. J. M. Bofill and W. Quapp, J. Chem. Phys. **134**, 074101 (2011).
47. W. Quapp and B. Schmidt, Theor. Chem. Acc. **128**, 47 (2011).
48. J. Gonzalez, X. Gimenez, and J. M. Bofill, PCCP (13), 2921 (2002).
49. One problem is that one can have NTs that have multiple maxima (some of which are turning points, rather than transition states) or NTs that have a single maximum that is a turning point, but don't pass through the transition state. Such NTs are not candidate reaction paths. Cf. ref. [35].
50. K. Muller and L. D. Brown, Theor. Chim. Act. **53**, 75 (1979).
51. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Peersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, O. Salvetti, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S.

Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, *Gaussian03, Revision D.01*. (Gaussian Inc., Wallingford, CT, 2004).

52. A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
53. C. Lee, W. Yang, and R. G. Parr, Phys. Rev. B **37**, 785 (1988).
54. W. Quapp, M. Hirsch, and D. Heidrich, Theor. Chem. Acc. **112**, 40 (2004).

Chapter 5:

Computational study of the binding modes of caffeine to the adenosine A_{2A} receptor*

* The content of this chapter is submitted to J. Phys. Chem. B: Yuli Liu; Steven K. Burger; Esteban Vöhringer-Martinez; Paul W. Ayers; Computational study of the binding modes of caffeine to the adenosine A_{2A} receptor.

5.1 Statement of Problem

As discussed in previous chapters, chemical processes involving small molecules can be predicted and characterised from stationary points or reaction paths linking them on the potential energy surface (PES). However, for large molecules like proteins, statistical sampling techniques like molecular dynamics (MD) simulation are needed because entropic effects and dynamical conformational changes make important contributions to thermodynamic properties of the system, e.g.: the binding free energy of protein-ligand complexes.

In this chapter, using the recently solved crystal structure of the human adenosine A_{2A} receptor, we applied MM/PBSA to compare the binding mechanism of caffeine with the antagonist ZM241385. MD simulations were performed on the protein, which was embedded in a lipid membrane bilayer and then solvated with water. Four low-energy binding modes of caffeine-A_{2A} were found, all of which had similar energies. Assuming an equal contribution of each binding mode, a binding free energy of -21.2kcal/mol was calculated for caffeine. The binding free energy difference between caffeine and ZM241385 was determined to be -2.4 kcal/mol, in good agreement with the experimental value of -3.6kcal/mol. The configurational entropy contribution of -0.9 kcal/mol from multiple binding modes of caffeine helps explain how a small molecule like caffeine can compete with significantly larger molecule, ZM241385, which can form many more interactions with the receptor. We also performed residue-wise energy decomposition and

found that Phe168, Leu249 and Ile274 contribute most significantly to the binding modes of caffeine and ZM241385.

5.2 Introduction

Caffeine is the most popular psychoactive stimulant, with 1/3 of the world's population under its influence. Caffeine has a wide range of psychostimulative and neuroprotective effects on humans.[1] According to recent pharmacologic and genetic studies, caffeine's effects are primarily due to its interference with neurotransmission within the basal ganglia, and more specifically its involvement in the blockage of adenosine receptors, especially the adenosine A_{2A} receptor.[1-5]

Adenosine receptors are G protein coupled receptors (GPCR).[6] In the human central nervous system, there are four adenosine receptor subtypes: A_1, A_{2A}, A_{2B} and A_3.[7] A_1 and A_{2A} receptors play roles in the heart and the brain, regulating myocardial oxygen consumption, coronary blood flow and mediating the release of neurotransmitters such as dopamine, glutamate and acetylcholine. A_{2B} and A_3 receptors are involved in emergency processes such as inflammation, injury and immune responses.[7-13]

Caffeine is one of many xanthine derivatives that act as a non-selective antagonists on A_1 and A_{2A} receptors. Initially the stimulant effect of caffeine was credited to the blockage of A_1 rececters[14-17]. However recent pharmacological[4] and genetic studies[2]

show that the psychomotor stimulant effects from low doses of caffeine, as found in everyday beverages and food, depend on A_{2A}, not A_1, receptors.

While the normal physiological response to caffeine is interesting in its own right, there have been a number of preclinical and clinical studies that show a link between caffeine intake and a reduced risk of Parkinson's disease.[18,19] This has lead to an interest in xanthine derivates as A_{2A} antagonists for treating Parkinson's and other diseases associated with A_{2A} receptors.[20-23] Elucidating the mechanism of action of caffeine may give important insights into designing new xanthine based compounds for controlling the adenosine A_{2A} receptor.

To date, there are only two computational studies examining the binding modes of caffeine and other xanthine antagonists.[24,25] These studies used homology modeling and de novo design since no crystal structure was available for the A_{2A} receptor at the time. Recently the crystal structure of human adenosine A_{2A} receptor (PDB ID:3EML) has been solved by the Stevens group with a bound high-affinity selective antagonist, ZM241385.[6] The structure proved difficult to characterize due to the thermal instability of the A_{2A} receptor, so the receptor had to be engineered using the T4-lysozyme fusion strategy,[26-28] in which the intracellular part of the receptor (mostly the third cytoplasmic loop: Leu209 – Ala221, was replaced with lysozyme from T4 bacteriophage and the carboxy-terminal tail (Ala317 to Ser412) of the receptor was deleted. The engineered A_{2A} receptor (A_{2A}-T4L-ΔC) was shown to be a functional receptor with increased agonist binding affinity and a wild-type affinity for the antagonist.[6] The crystal structure

demonstrated that the binding pocket is perpendicular to the membrane surface. The structure also revealed the importance of several previously uncharacterized pocket residues: Phe168, Met177, and Leu249. The significance of these residues was further confirmed by site-directed mutagenesis and *in silico* mutation studies.[29]

To calculate the free energy difference between ligands, free energy perturbation (FEP)[30,31] can be used. In a FEP calculation, the free energy difference between two states is calculated by slowly perturbing one state into another via a set of mixed states. Generally the method works well if the structural changes between the initial and final states are small. In this case the structures of caffeine and ZM241385 are too different for FEP to be effective(Figure 5. 1, Figure 5. 2), so instead we used the molecular mechanics/Poisson-Boltzmann surface area (MM/PBSA) method. MM/PBSA has been shown to work well with various systems[32-37], especially for the relative binding affinities of a series of ligands bound to the same protein[36], although large standard deviations from the protein entropy term can impede discrimination of ligands with similar binding affinities.[35,38,39]

In this study we performed the following steps to determine the free energy differences between the two ligands: 1) Using A_{2A}-T4L-ΔC,[6] we docked caffeine, and, for comparison, we re-docked ZM241385. 2) After docking the ligands we ran molecular dynamics (MD) simulations in explicit solvent on selected docked poses. 3) We then determined the relative binding energy with MM/PBSA and 4) finally we performed a residue-wise decomposition of the energetic contributions to determine which residues

were conserved between selective (ZM241385) and non-selective (caffeine) antagonist binding. This analysis gives a sense of the binding poses involved with these types of antagonists. The analysis reveals how a small molecule like caffeine can compete with a significantly larger molecule, ZM241385, which is capable of forming many more interactions with the receptor.[25]

5.3 Computational Methods

5.3.1 Docking

For the protein structure we used the engineered A_{2A} receptor (A_{2A}-T4L-ΔC, PDB ID: 3EML) from which we removed the bound inhibitor ZM241385. The two ligands to be docked were caffeine (Figure 5. 1) and ZM241385 (Figure 5. 2). The molecules were first optimized with HF/6-31G* level of theory and basis set using *Gaussian03*.[40] Atomic charges were then assigned based on the molecular electrostatic potential fitting, with the aide of the *RESP* module in the *Amber10* program package[41]. The docking was done with the *AutoDock Vina* program[42], using a 32×26×28 search grid centered at (-5, 7, 60) with a spacing of 1Å. (The origin is defined in the PDB file of the crystal structure.) To search for binding modes the iterated local search global optimizer was used.[42] From the resulting binding modes we examined in detail only the top nine binding modes; the other lower scoring modes were found to be either unphysical or redundant.

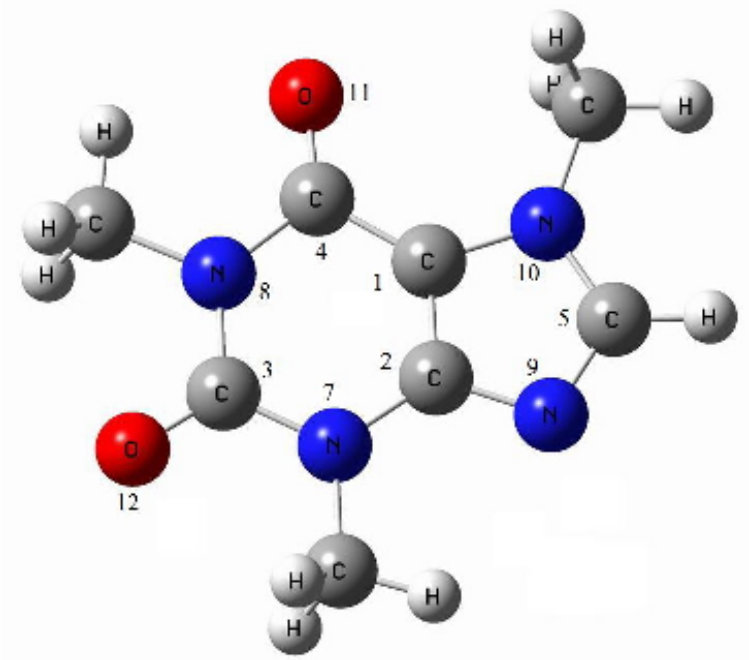


Figure 5. 1: The molecular structure and atom numbering of caffeine.

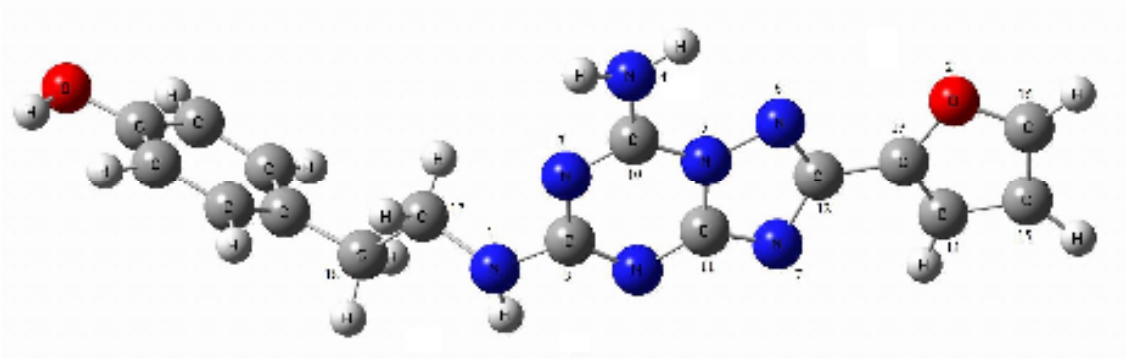


Figure 5. 2: The molecular structure and numbering of ZM241385

An analysis of the caffeine binding modes demonstrated that the most important interactions are hydrogen bonds and aromatic stacking. Specifically, the most important hydrogen-bond interactions are between hydrogen bond acceptors in caffeine (N9, O11, O12) and the hydrogen bond donors (Asn253, His278) from the receptor. There are important aromatic stacking interactions between the bicyclic core of caffeine and the benzene ring of Phe168. Five of the nine binding modes involved relatively strong hydrogen bond(s) (where the distance between the heavy atoms is less than 4.0Å) and/or aromatic stacking with Phe168 and were considered suitable for further study by MD simulations.

The observed binding modes of ZM241385 could be categorized into two types. In type I the ligand is perpendicular to the membrane surface, with its furan ring buried inside the receptor; this is the same orientation as the crystal structure, except for some minor differences due to the rotation of C17-C18 bond (Figure 5. 2). In type II the ligand is flipped 180°; the ligand is still perpendicular to the membrane surface, but its furan ring points towards the extracellular fluid. A representative example of each type of binding mode was then selected for further study by MD simulations.

5.3.2 Molecular Dynamics

MD simulations were performed on the five docked poses of caffeine, the crystal structure with ZM241385 bound (type I), the 180° flipped pose of ZM241385 (type II).

To approximate the physiological environment of the A_{2A} receptor, each selected ligand-protein complex structure was inserted into a 100×100Å POPC (palmitoyl-oleyl-phosphatidyl-choline) membrane bilayer generated using *VMD*,[43] and then solvated with two layers of TIP3P water molecules above and below the lipid bilayers in such a way that the intracellular and extracellular parts of the receptor are at least 15Å away from the solvent boundary.

The POPC membrane bilayers were parameterized to be consistent with the rest of the Amber parameters[44]. This involved optimizing POPC with HF/6-31G* using *Gaussian 03*[40], and then fitting the atomic charges using the *RESP* module of *Amber10*[41]. The rest of the parameters for POPC were generated using the general Amber force field (GAFF) using the *ANTECHAMBER* module in *Amber10*.

When building the initial structure, POPC lipid molecules within 0.8Å and water molecules within 3.8Å of the receptor were removed. The *TLEAP* module in *Amber 10* program package is used to add hydrogens and counterions. A total of 17 Cl^- counterions were added to neutralize the entire system. The final system size was 100×100×125Å. The numbers of water molecules were slightly different between the ZM241385-A_{2A} and caffeine-A_{2A} complexes. With caffeine there were 115,930 atoms, including 26,162 water molecules, 225 POPC lipid molecules and 17 Cl^- anions.

The parameters for caffeine and ZM241385 were also generated using GAFF with the *ANTECHAMBER* module, with the atomic charges determined by *RESP* fitting from a HF/6-31G* calculation. The standard Amber ff03 force field was used for the receptor.

Before running MD a series of energy minimizations were performed. First 2000 steps of minimization were done on the hydrogen atoms to remove any bad contacts, followed by 2000 steps on the water molecules. Then 22000 steps of minimization were done on protein side chains and the ligand molecule, followed by 50000 steps on the environment (POPC molecules and water molecules). To better solvate the system, short MD simulations and further minimizations were done on the environment alternatively: 160ps of MD on the water molecules using a NVT ensemble; then 20000 minimization steps on the environment (water molecules, POPC molecules and counterions); then another 100ps of MD were done on POPC molecules with a NVT ensemble; then 25000 steps of energy minimization on the environment (water molecules, POPC molecules and counterions), and finally 6000 steps of energy minimization on the entire system.

To heat up the system, a 100ps simulation was run with the NVT ensemble using the Langevin temperature regulation scheme. This was followed by 100ps with the NVT ensemble, then 100ps with the NPT ensemble. The production phase was done for 5 ns with a periodic boundary condition at 300K and 1 atm with anisotropic pressure scaling to maintain a NPT ensemble. The particle-mesh Ewald (PME) method was applied to get the correct long-range electrostatic interactions, and a nonbonded cutoff of 12Å was used. Bond stretching involving hydrogen atoms was restrained with the SHAKE algorithm, enabling the use of a 2-fs time step for MD simulation.

5.3.3 MM/PBSA Binding Energy Calculation

The relative binding energies of the binding modes of caffeine and ZM241385 bound to the A_{2A} receptor were evaluated using MM/PBSA, as implemented in *Amber10*. In MM/PBSA the binding free energy is defined as,

$$\Delta G_{bind} = \left\langle G_{complex} \right\rangle - \left(\left\langle G_{protein} \right\rangle + \left\langle G_{ligand} \right\rangle \right), \tag{5.1}$$

where the average of complex, protein, and ligand free energies are evaluated from uncorrelated snapshots in the MD trajectories. The binding free energy can be expressed as,

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{solv} - T\Delta S, \tag{5.2}$$

where ΔE_{MM} is the molecular mechanical energy difference between the bound state (complex) and unbound state (receptor and ligand), $\Delta G_{solv} = \Delta G_{PB} + \Delta G_{nonpolar}$ and $\Delta G_{nonpolar} = \gamma \text{SASA}, \gamma = 0.0072 \text{kcal}\,\text{Å}^{-2}$. T is the temperature, and ΔS is the entropy change upon binding.

The entropy term $T\Delta S$ in Eq. (5.2) is not considered in MM/PBSA calculation, due to a large standard deviation from normal mode analysis[45]. It is estimated using a simplified hindered rotor model instead.[46] We use the following equation in the MM/PBSA relative binding energy calculation,

$$\Delta E_{bind} = \Delta E_{MM} + \Delta G_{solv}. \tag{5.3}$$

To differentiate from the binding free energy ΔG_{bind}, ΔE_{bind} is used to denote the relative binding energy.

The binding free energy of caffeine is calculated using a Boltzmann weighting

$$w(i) = \frac{e^{-\Delta E_i / RT}}{\sum_{i=1}^{5} e^{-\Delta E_i / RT}} \tag{5.4}$$

for each binding mode, where R is the gas constant (8.314 $JK^{-1}mol^{-1}$) and T = 310 K (the human body temperature), and ΔE_i is the relative binding energy of binding mode i. The overall binding energy of caffeine is the sum of the weighted energy of each binding mode,

$$\Delta U_{bind} = \sum_{i=1}^{5} w(i) \cdot \Delta E_i \,. \tag{5.5}$$

The standard deviation is calculated as,

$$STD = \sqrt{\sum_{i=1}^{5} w(i) \cdot (\Delta E_i)^2} \,. \tag{5.6}$$

For each ligand, there is a configurational entropy contribution ($T\Delta S$) to the binding free energy. The configurational entropy[47] is defined as,

$$\Delta S = -R \sum_{i=1}^{N_{bind}} w(i) \cdot \ln w(i) \,, \tag{5.7}$$

where N_{bind} is the number of binding modes.

5.3.4 Residue-wise MM/GBSA energy decomposition

Residue-wise MM/GBSA energy decomposition was performed on the low-energy binding modes of caffeine (binding mode 1, 2, 4, 5) and ZM241385 (binding mode 1) to calculate the contribution of each residue to the total binding energy of the complex. In *Amber10*, only MM/GBSA is available for energy decomposition. For MM/GBSA binding energy calculations, the same free energy equation is used as the MM/PBSA method except that in equation (5.3) $\Delta G_{solv} = \Delta G_{GB} + \Delta G_{nonpolar}$, with the electrostatic part of the solvation energy calculated using the generalized Born method (GB) instead of by solving the Poisson-Boltzmann equation (PB).

5.4 Results and Discussion

5.4.1 Binding Modes

MD simulations on the five docking poses of caffeine and the two poses of ZM241385 converged to stable binding modes. The RMSD plots of the protein backbone atoms show that MD simulations on the five binding modes converge after about 1.5ns (Figure 5. 3).

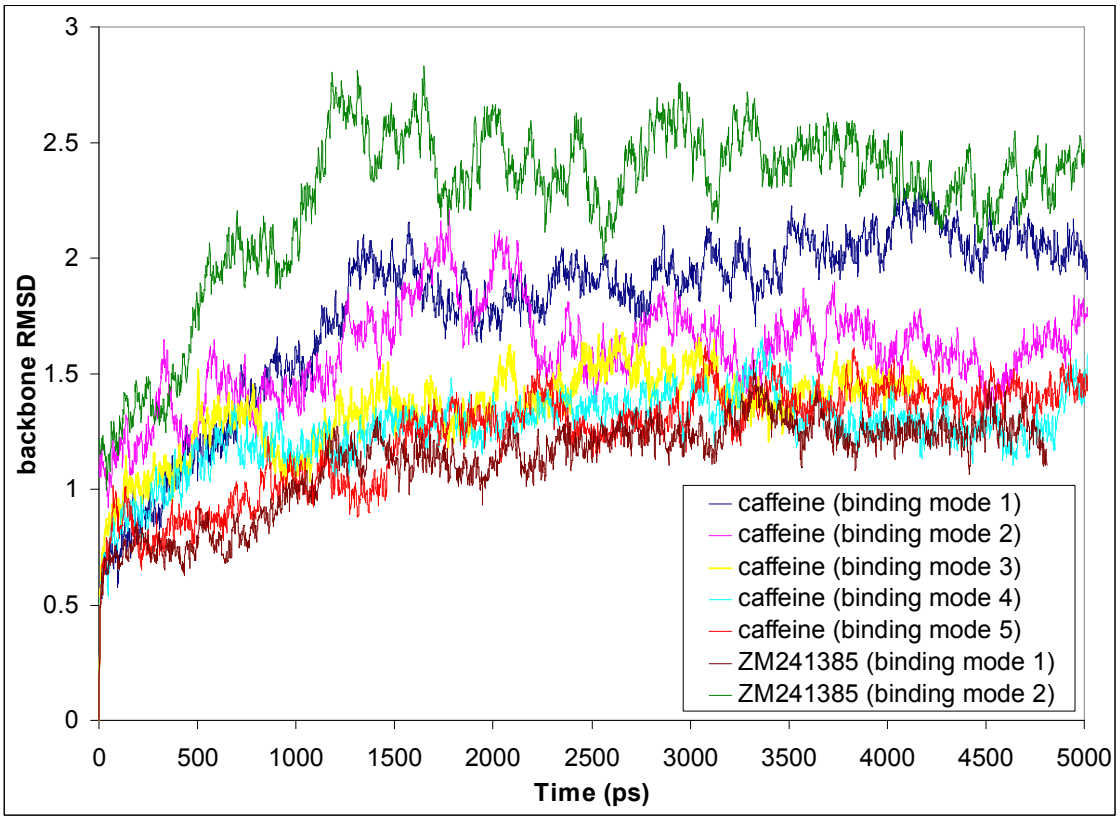


Figure 5. 3: RMSD of the backbone atoms of the receptor. Each binding mode is represented a unique color: Blue for binding mode 1 of caffeine, magenta for binding mode 2 of caffeine, yellow for binding mode 3 of caffeine, cyan for binding mode 4 of caffeine, red for binding mode 5 of caffeine, brown for binding mode 1 of ZM241385 and olive for binding mode 2 of ZM241385.

In binding mode 1 of caffeine, O12 of caffeine forms a hydrogen bond network with OG1 of Thr88 (2.54Å) and NE2 of His250 (2.92Å) (Figure 5.8). (The hydrogen bond length is denoted as the average distance between the heavy atoms over 5ns MD simulation.) Additionally, NE2 of His278 forms a hydrogen bond with O11 of caffeine (2.90Å). Other important hydrophobic interactions include residues Val84, Phe168, Trp246, Leu249 and Ile274, although the aromatic stacking between caffeine and Phe168 is broken because the aromatic rings are not parallel.

In binding mode 2 of caffeine only one hydrogen bond is formed between NE2 of His278 and O11 of caffeine (2.98Å) (Figure 5.8). During the MD simulation, water molecules get into the active site and mediate the binding of caffeine to the receptor. The hydrogen bond acceptor O12 of caffeine forms a hydrogen bond with a water molecule (2.66Å). Other important hydrophobic interactions in binding mode 2 involve residues Val84, Phe168, Leu249 and Ile274. As in binding mode 1, the aromatic stacking between caffeine and Phe168 is broken.

In binding mode 3 the most significant interaction is the aromatic stacking with Phe168 (stacking distance 3.6Å, the stacking distance is determined by the average distance of the center of mass over 5ns MD simulation). No hydrogen bond formed between caffeine and the receptor. Water molecules interfere with the binding of caffeine to the receptor for this mode, forming a hydrogen bond of 2.86Å with O11 of caffeine and interfering with the hydrogen bond between O11 of caffeine and ND2 of Asn253

(4.5Å). Other direct hydrophobic contacts include residues Val84, Glu169, Leu249, His250, and Ile274.

The most important interactions in binding mode 4 include a hydrogen bond between O11 of caffeine and NE2 of His250 (3.63Å), a hydrogen bond between O12 of caffeine and NE2 of His278 (3.06Å), and the aromatic stacking between caffeine and Phe168 (stacking distance 3.9Å) (Figure 5.8). A hydrophobic binding pocket is formed from residues Met177, Leu249, and Ile274.

In binding mode 5, N9 of caffeine forms a hydrogen bond with ND2 of Asn253 (3.7Å) and the aromatic rings of caffeine stack with Phe168 (stacking distance 3.7Å). O12 of caffeine forms a hydrogen bond with a water molecule (2.81Å) (Figure 5.8). Other important hydrophobic interactions involve Leu249 and Ile274.

Binding mode 1 of ZM241385-A_{2A} complex is the result of MD simulation on the crystal structure. In binding mode 1, Asn253 forms multiple hydrogen bonds with atoms on ZM241385 (OD1 of Asn253 to N4 of ZM241385 (2.97Å), ND2 of Asn253 to N6 of ZM241385 (3.69Å), ND2 of Asn253 to O2 of ZM241385 (3.71Å)). Phe168 stacks with the bicyclic core of ZM241385 (stacking distance 3.95Å). Other important hydrophobic interactions involve Ile66, Leu249, His250 and Ile274.

In binding mode 2, ZM241385 remained in a 180° flipped mode compared to binding mode 1. OE1 of Glu13 forms a hydrogen bond with N4 of ZM241385 (3.12Å), OH of Tyr271 forms a hydrogen bond with O2 of ZM241385 (3.72Å). Phe168 has multiple hydrophobic contacts with the ligand but its stacking with the bicyclic core of ZM241385

does not hold. Other close contacts include Val84, Trp246, Leu249, His250, Ile274, and His278.

5.4.2 Relative Binding Energy

The relative binding energies of the binding modes of caffeine and ZM241385 are shown in Table 5. 1. We can see that binding mode 1 is the most stable binding mode of caffeine. The binding energies of other 3 modes (2, 4 and 5) are within a standard deviation (~3kcal/mol) of this result, while mode 3 is 7.8kcal/mol higher. We focus our attention on the low-energy modes 1, 2, 4 and 5 and discard mode 3.

Table 5. 1: MM/PBSA calculation results for 5 binding modes of caffeine- A_{2A}, and 2 binding modes of ZM241385-A_{2A} complex.

Complex	Binding Energy[a] (kcal/mol)	Standard Deviation (kcal/mol)
Caffeine, binding Mode 1	-20.5	2.7
Caffeine, binding Mode 2[b]	-18.6	1.9
Caffeine, binding Mode 3[b]	-12.7	3.2
Caffeine, binding Mode 4	-17.8	3.1
Caffeine, binding Mode 5[b]	-17.4	2.7
ZM241385, binding mode 1	-25.4	3.5
ZM241385, binding mode 2	-14.8	3.0

[a.] The calculation of binding energy is according to Equation (5.3).
[b.] Water molecules mediate the binding through forming hydrogen bond with hydrogen bond acceptors (O11, O12 or N9) in caffeine.

Table 5. 2: Calculation of the overall binding energy of the caffeine-A_{2A} complex from the relative binding energy of each binding mode.

Binding Energy (ΔE_i, kcal/mol)	Weight $w(i)$	$w(i)\cdot\Delta E_i$	Standard Deviation (kcal/mol)	$w(i)\cdot(\Delta E_i)^2$
-20.5	0.94	-19.3	2.7	6.7
-18.6	0.04	-0.7	1.9	0.1
-12. 7	0.00	0.00	3.2	2.9E-05
-17.8	0.01	-0.2	3.1	0.1
-17.4	0.01	-0.1	2. 7	0.04
Sum	1.0000000	**-20.4**[a]	**2.7**[b]	7.0

[a.] The overall binding energy of caffeine ΔU_{bind} (equation (5.5)).

[b.] The standard deviation of the overall binding energy ΔU_{bind} (see equation (5.5) and (5.6)).

Since the energy differences between binding mode 1, 2, 4 and 5 are within one standard deviation (see Table 5. 1), we can assume that all four binding modes have the same binding energy; this assumption yields the maximum contribution from the configurational entropy. If we assume that binding mode 1, 2, 4 and 5 have same binding energy and therefore same probability ($w(i) = \frac{1}{4}, i = 1,2,4,5$) in Boltzmann distribution, then the configurational entropy contribution is,

$$T \cdot \Delta S = -RT \sum_{i=1}^{4} w(i) \cdot \ln w(i) = -RT \left(4 \times \frac{1}{4} \ln \frac{1}{4} \right) = 0.9 kcal / mol . \quad (5.8)$$

The optimum binding free energy of caffeine would then be,

$$\Delta G_{bind} = \Delta U_{bind} - T \cdot \Delta S = -21.2 kcal / mol , \quad (5.9)$$

where ΔU_{bind} is the overall binding energy of caffeine (equation (5.5) and Table 5. 2).

As for ZM241385, binding mode 1 is found to be dominant in terms of relative binding energy (Table 5. 1). So binding mode 2 can be discarded and the configurational entropy contribution to the binding free energy is zero.

The entropy upon binding due to the protein conformational change is calculated using principal component analysis (PCA).[48,49] For each binding mode, 100 uncorrelated snapshots were taken from the 5ns MD trajectory. After removing the translational and rotational motions, a mass-weighted covariance matrix of the backbone atoms was calculated from all snapshots. By diagonalizing the covariance matrix, the eigenvectors (also called the principal components) and eigenvalues can be obtained. Using the

g_anaeig tool of *Gromacs* program package,[50] the entropy was calculated from the Schlitter formula. The PCA calculation results (Supporting Information, Table 5.5) show that the entropies for four low-energy binding modes of caffeine and the dominant binding mode of ZM241385 present no significant difference during the 5ns simulation time.

The restrained internal rotation of the pocket residues and the ligand accounts for the major local conformational change of the binding pocket. A simplified hindered rotor model is applied to calculate the entropy contribution from local conformational change. Assuming 2 or 3 states of equal energy per rotatable bond in the unbound state, the entropy penalty for each restrained rotation upon binding would be $RT\ln 2$ or $RT\ln 3$ (0.4 or 0.7 kcal/mol). A multiple linear regression analysis on the experimental binding data of 45 protein-ligand complexes gave an empirical value of 0.3kcal/mol for the entropy contribution of a restrained rotor.[46] According to the average structures from 5ns MD simulation, the ZM241385-A_{2A} complex has 6 more restrained rotors than caffeine-A_{2A} upon binding, which leads to 1.8 kcal/mol entropy penalty to the relative binding free energy. Therefore the binding free energy difference between ZM241385 and caffeine is -2.4kcal/mol ($\Delta G = -25.4 - (-21.2) + 1.8 = -2.4 kcal/mol$). According to experimental measurements, ZM241385 is 397 times more potent than caffeine for binding to the adenosine A_{2A} receptor in terms of IC50.[51] Since the experimental conditions of caffeine and ZM241385 were exactly the same, the binding affinity difference is also 397 times

according to the Cheng-Prusoff equation.[52] This equals a -3.6kcal/mol difference in binding free energy ($\Delta G = 2.303RT\log\left(\frac{K_{ZM}}{K_{CAF}}\right) = -3.6kcal/mol$), which matches well with the calculated number.

ZM241385 is a long molecule and it is much larger than caffeine. An analysis of the average structure of the four low-energy binding modes of caffeine and the dominant binding mode of ZM241385 over 5ns MD has shown that the bicyclic core of ZM241385 binds at a similar position as caffeine does in multiple binding modes, and its phenol ring points to the extracellular side of the membrane surface and has closer contact with transmembrane helix II (Figure 5. 5). This explains the large energy contribution of residues from helix II (Ile66) in binding mode 1 of ZM241385 as shown in Table 5. 3.

Although caffeine is relatively small and has fewer contact centers, it can form more binding modes with the A_{2A} receptor than large molecules like ZM241385, and therefore has a favourable configurational entropy contribution to the binding free energy. We estimate that the binding affinity of caffeine is about one order of magnitude larger than it would have been were there only a single binding conformer. Larger ligands, like ZM241385, also face a larger entropic penalty upon binding because more hindered rotators are constrained; the loss of rotational freedom in ZM241385 makes it bind about one order of magnitude more weakly than it would were as a rigid ligand.

Table 5. 3: Comparison of pocket residues in 4 low-energy binding modes of caffeine-A_{2A} and the dominant binding modes of ZM241385-A_{2A}.

Complex	Binding Mode 1 of caffeine-A_{2A}	Binding Mode 2 of caffeine-A_{2A}	Binding Mode 4 of caffeine-A_{2A}	Binding Mode 5 of caffeine-A_{2A}	Binding mode 1 of ZM241385-A_{2A}
Residues with energy contribution > 0.8 kcal/mol	Val84	Val84	**Phe168**	**Phe168**	Ile66
	Phe168	**Phe168**	Met177	**Leu249**	**Phe168**
	Leu249	**Leu249**	**Leu249**	Asn253	**Leu249**
	His250	**Ile274**	His250	**Ile274**	His250
	Ile274	His278	**Ile274**		Asn253
	His278		His278		**Ile274**
Other residues within 4.5Å	Ala63	Val55	Ala63	Ala63	Ala63
	Thr88	Ala59	Ile66	Ile66	Ser67
	Met177	Ala63	Val84	Ser67	Val84
	Asn181	Ile66	Leu85	Val84	Leu85
	Trp246	Leu85	Thr88	Leu85	Thr88
		Leu87	Glu169	Met174	Glu169
		Thr88	Trp246	Met177	Met177
		Trp246	Asn253	Trp246	Trp246
		Ser277	Met270	His250	Leu267
				Met270	Tyr271
				Tyr271	

Residues on different transmembrane (TM) helices are highlighted in different colours: TM II - green, TM III - yellow, TM V - cyan, TM VI - magenta, TM VII – grey.

5.4.3 MM/GBSA energy decomposition results

Residue-wise MM/GBSA energy decomposition calculations are performed on the low-energy binding modes of caffeine (mode 1, 2, 4, 5) and ZM241385 (mode 1). (The high-energy binding modes (binding mode 3 of caffeine and binding mode 2 of ZM241385) are discarded and are not discussed further.) For the binding modes of caffeine, Val84, Phe168, Leu249, His250, Ile274 and His278 make the most important contributions to the binding energy, followed by Ala63, Leu85, Thr88, Met177, Trp246, and His250. Trp246 is considered the key to a "toggle-switch" activation mechanism of the human adenosine A_{2A} receptor.[6,53] The rotameric position of Trp246 is thought to control the equilibrium between the active and inactive states of the receptor. According to the MM/GBSA energy decomposition results as shown in Figure 5. 4, Trp246 has important interactions with the ligand in all binding modes of caffeine-A_{2A} and ZM241385-A_{2A}.

Phe168, Leu249 and Ile274 have an energy contribution greater than 1.2kcal/mol in all four binding modes of caffeine-A_{2A} and the binding mode of ZM241385-A_{2A} (Figure 5. 4). Thus, Phe168, Leu249 and Ile274 can be considered as conserved residues for selective and non-selective antagonist binding. The binding pockets are formed by residues on the transmembrane helix II, III, V, VI and VII, which is clearly shown by colour representations of the transmembrane helices in Table 5. 3, Figure 5. 6, Figure 5. 7 and Figure 5. 8

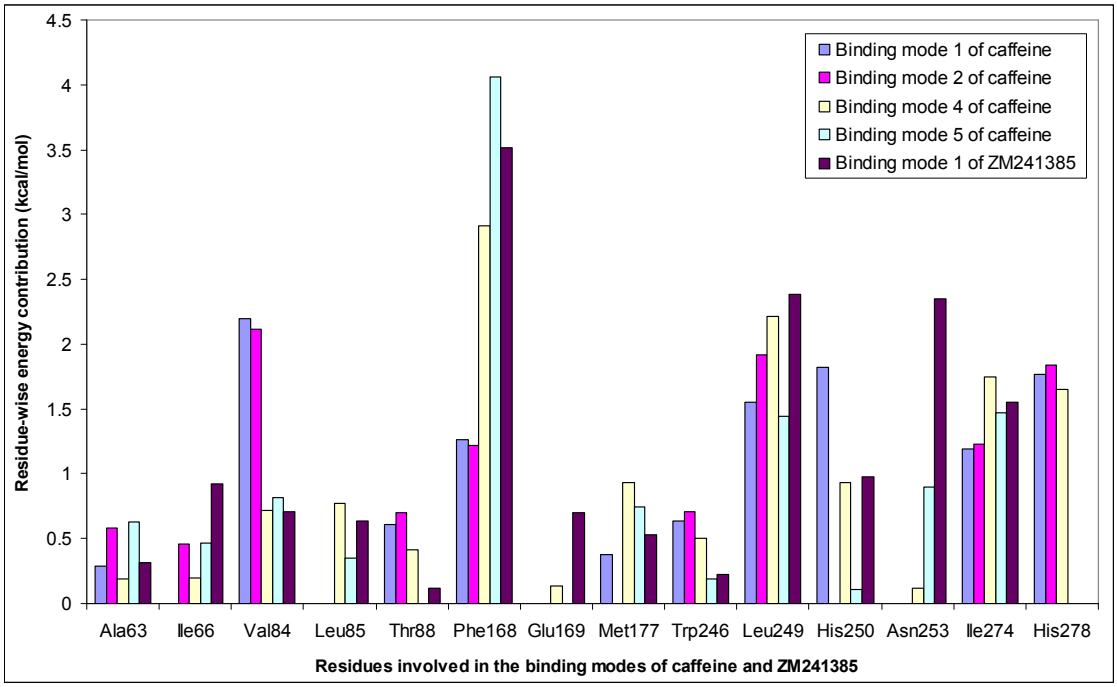


Figure 5. 4: Residue-wise energy contribution of pocket residues in the binding modes of caffeine and ZM241385.

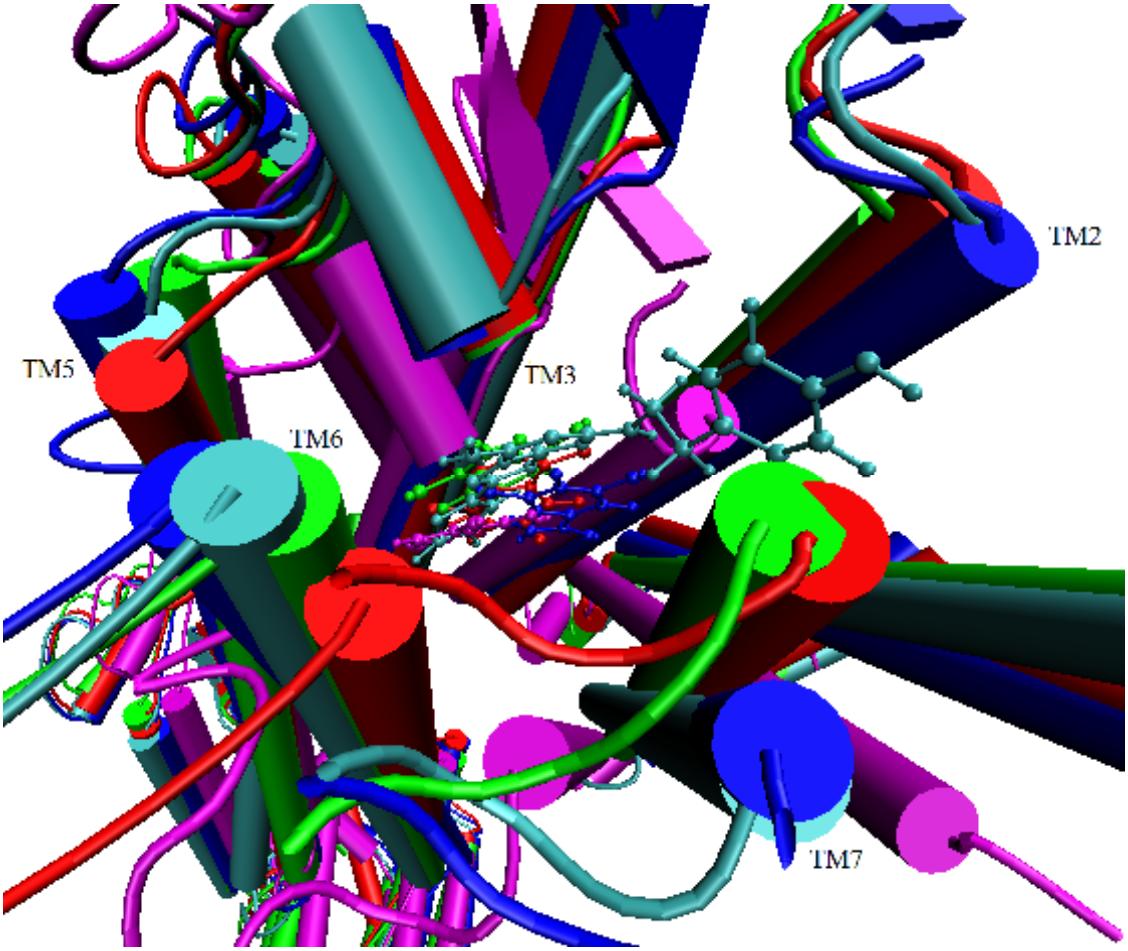


Figure 5. 5: A superimposition of the 4 low-energy binding modes of caffeine (mode 1, 2, 4, 5) and the dominant binding mode of ZM241385 (mode 1). Binding mode 1 of caffeine is shown in red, binding mode 2 of caffeine is shown in blue, binding mode 4 of caffeine is shown in green, binding mode 5 of caffeine is shown in purple. Binding mode 1 of ZM241385 is shown in cyan.

5.4.4 Comparison with site-directed mutagenesis studies

Previous site-directed mutagenesis studies indicate the significance of residues Val84, Phe168, Glu169, Met177, Leu249, His250, Asn253, Ile274 and His278 in agonist and antagonist binding with the adenosine A_{2A} receptor (Table 5. 4).[29,55-60] (Details about the numbering scheme in the parenthesis can be found in ref. 54.) Mutating Phe168, Glu169, Met177, Leu249 and Asn253 to Ala (A) impedes antagonist (ZM241385) binding.[29,56] Mutation of Val84 to Leu (L) impedes the binding of xanthine-type ligands but other antagonists are not affected.[58] Mutation of His278 to Tyr (Y) decreases the binding affinity of theophylline.[57] Mutation of His250, Ile274 and His278 to Ala (A) abolishes antagonist binding.[55,56]

According to our MM/GBSA energy decomposition results, the pocket residues are categorized to the following three significance levels based on their contributions to the binding energy: >1.2kcal/mol, >0.8kcal/mol. and >0.5kcal/mol. (Table 5. 4). Phe168, Leu249, and Ile274 (Figure 5. 4) make contributions over 1.2kcal/mol to the binding energies in four low-energy binding modes (mode 1, 2, 4, 5) of caffeine-A_{2A} and the dominant binding mode (mode 1) of ZM241385-A_{2A}. As shown in Table 5. 4, previous site-directed mutagenesis studies show that the mutation of these residues to Ala (A) completely abolished agonist and antagonist binding.[29,56] Residues with >0.8kcal/mol energy contribution in the binding mode(s) of caffeine-A_{2A} (e.g., Val84, Met177, His250, Asn253, and His278) have been shown to abolish antagonist or xanthine-type antagonist

binding in site-directed mutagenesis studies.[29,56-58] Other important residues (>0.5kcal/mol energy contribution) include Ala63, Leu85, Thr88, Glu169 and the "toggle-switch" residue Trp246. Mutation of Thr88 to Ala (A), Ser (S), Arg (R), Asp (D) and Glu (E) decreased agonist binding substantially, but not antagonist binding.[60] Mutation of Glu169 to Ala (A) impeded agonist and antagonist binding.[59] No mutagenesis data is available for Ala63, Leu85 and Trp246.

A comparison between our MM/GBSA energy decomposition has shown that pocket residues with >1.2kcal/mol energy contribution (Phe168, Leu249, Ile274) are crucial in agonist and antagonist binding according previous site-directed mutagenesis studies; residues with >0.8kcal/mol energy contribution are important in xanthine-type antagonist binding; residues with >0.5kcal/mol energy contribution either have some impact on agonist (Thr88) or antagonist (Glu169) binding, or no site-directed mutagenesis data is available (Ala63, Leu85, Trp246) for comparison. Although residues with >0.5kcal/mol energy contribution are part of the binding pocket, we postulate that they are not crucial to antagonist binding, which may be the reason that no site-directed mutagenesis study has been done on Ala63, Leu85 and Trp246.

Table 5. 4 Important residues involved in the binding of caffeine and ZM241385 to the adenosine A_{2A} receptor, comparison of MD simulation results and site-directed mutagenesis results.

Residues	MD simulation results	Site-directed mutagenesis results
Ala63	>0.5kcal/mol energy contribution to binding mode 2 and 5 of caffeine-A_{2A}.	N/A
Val84	>0.5kcal/mol energy contribution to binding mode 1, 2, 4, 5 of caffeine-A_{2A} and binding mode 1 of ZM241385-A_{2A}.	L: impede xanthine type ligand binding.[58]
Leu85	>0.5kcal/mol energy contribution to binding mode 4 of caffeine-A_{2A}.	N/A
Thr88	H-bond with caffeine in binding mode 1, >0.5kcal/mol energy contribution in binding mode 1 and 2 of caffeine-A_{2A}.	A/S/R/D/E: decrease agonist but not antagonist activity.[60]
Phe168	>1.2kcal/mol energy contribution to binding mode 1, 2, 4, 5 of caffeine-A_{2A} and binding mode 1 of ZM241385-A_{2A}.	A: abolish both agonist and antagonist binding and receptor activity.[29]
Glu169	>0.5kcal/mol energy contribution in binding mode 1 of ZM241385-A_{2A}.	A: loss of agonist and antagonist binding.[59]
Met177	>0.5kcal/mol energy contribution in binding mode 4, 5 of caffeine-A_{2A}, and binding mode 1 of ZM241385-A_{2A}.	A: impede antagonist but not agonist binding.[29]
Trp246	>0.5kcal/mol energy contribution in binding mode 1, 2, 4 of caffeine-A_{2A}.	N/A
Leu249	>1.2kcal/mol energy contribution in binding mode 1, 2, 4, 5 of caffeine-A_{2A} and binding mode 1 of ZM241385-A_{2A}.	A: abolish both agonist and antagonist binding.[29]
His250	H-bond with caffeine in binding mode 1 and 4, >0.8kcal/mol energy contribution in binding mode 1, 4 of caffeine-A_{2A} and binding mode 1 of ZM241385-A_{2A}.	A: loss of agonist and antagonist binding.[56,58]
Asn253	H-bond with caffeine and >0.8kcal/mol energy contribution in binding mode 5 of caffeine-A_{2A}, H-bonds with ZM241385 and 2.4kcal/mol energy contribution in binding mode 1 of ZM241385-A_{2A}.	A: loss of agonist and antagonist radioligand binding.[56]
Ile274	>1.2kcal/mol energy contribution in binding mode 1, 2, 4, 5 of caffeine-A_{2A} and binding	A: loss of agonist and antagonist binding.[56]

	mode 1 of ZM241385-A_{2A}.	
His278	H-bond with caffeine in binding mode 1 and, >1.2kcal/mol energy contribution in binding mode 1, 2, 4 of caffeine-A_{2A}.	A: loss of agonist and antagonist binding.[56,57] Y: decrease the binding affinity of theophiline[57]

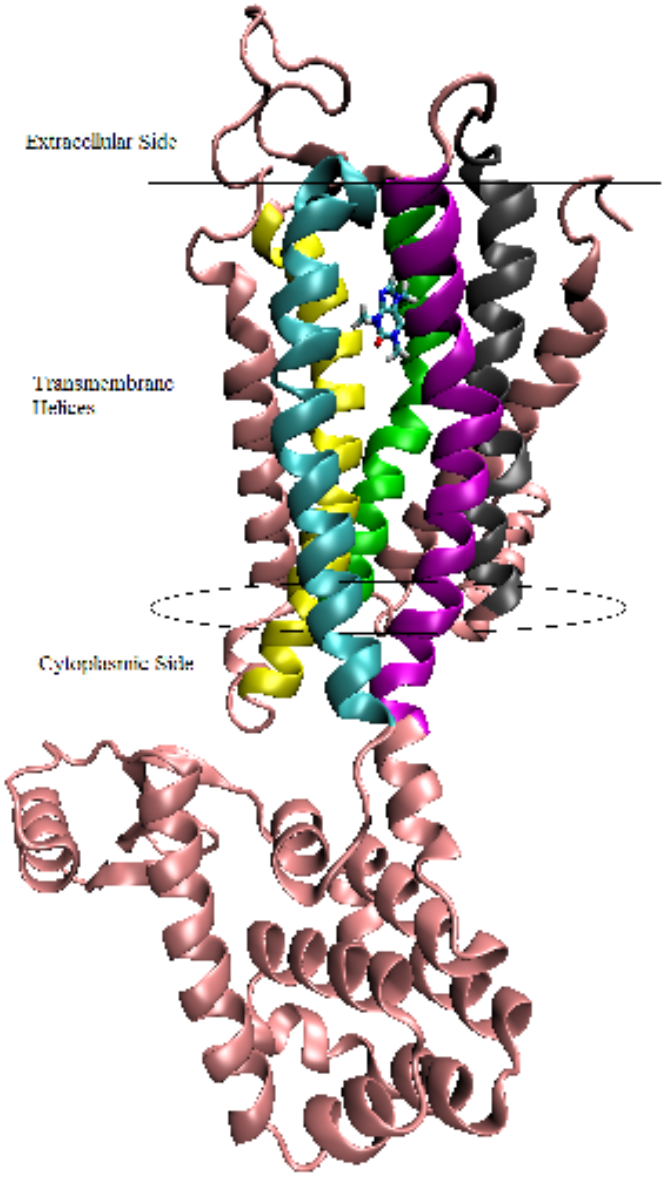


Figure 5. 6: Caffeine binding cavity. The five transmembrane (TM) helices that define the caffeine binding cavity are illustrated with different colors: TM II – green, TM III – yellow, TM V – cyan, TM VI – magenta, TM VII – grey. Everything else is in pink.

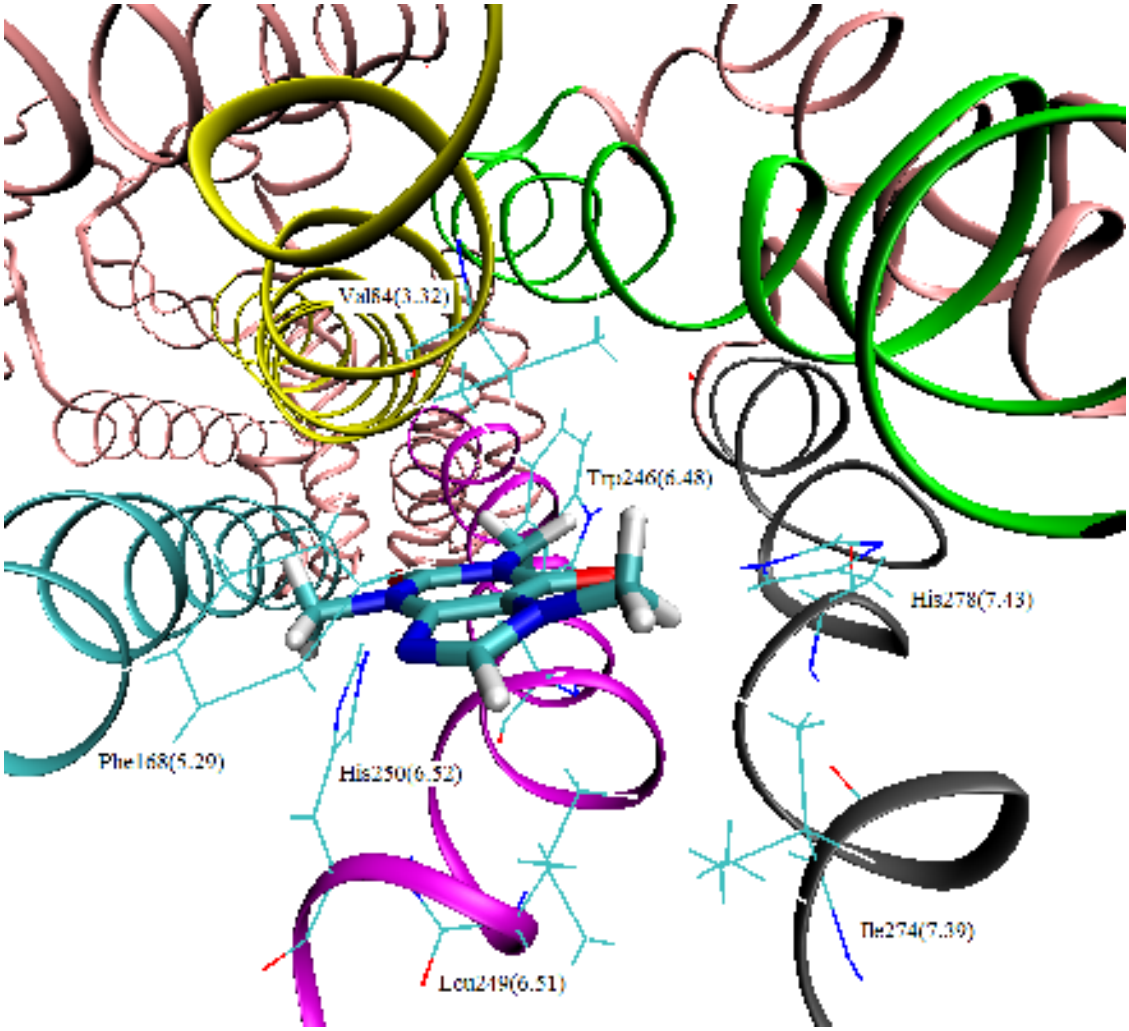


Figure 5. 7: Caffeine binding cavity, extracellular view. The five transmembrane (TM) helices that define the caffeine binding cavity are illustrated with different colors: TM II – green, TM III – yellow, TM V – cyan, TM VI – magenta, TM VII – grey. Some important pocket residues are labelled (residue numbering see Ref. 54) and shown in Lines. Everything else is in pink.

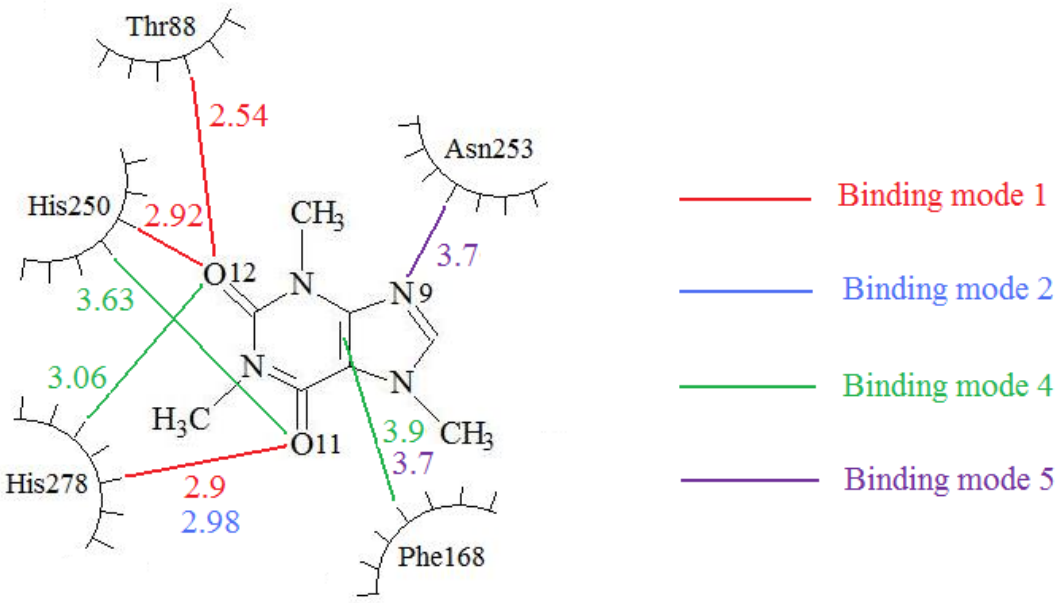


Figure 5. 8: Interactions between caffeine and the pocket residues. Only hydrogen bonds and aromatic stacking distances are shown. The hydrogen bond length is denoted as the average distance between the heavy atoms over 5ns MD simulation. The stacking distance is determined by the average distance of center of mass of Phe168 and caffeine over 5ns MD simulation.

5.5 Conclusion

Based on the recently solved crystal structure of the engineered human adenosine A_{2A} receptor, we docked caffeine to the receptor and performed MD simulations on selected docking poses. Five stable binding modes were found and the relative binding energies were calculated using MM/PBSA method. Except for binding mode 3, the binding energies of all other binding modes are within the range of standard deviation, which indicates that none of these binding modes are dominant, and all contribute to the binding free energy. Although caffeine is relatively small, with fewer contact centers to form strong interaction with the adenosine A_{2A} receptor, it is capable of forming more binding modes, which leads to a favourable configurational entropy contribution to the binding free energy of the complex.

Residue-wise MM/GBSA energy decomposition showed Phe168, Leu249 and Ile274 make significant contribution (>1.2kcal/mol) to the binding modes of both ligands, and were clearly conserved through selective and non-selective antagonist binding. These three residues and others, which made major contributions (>0.8kcal/mol), were in agreement with the results of site-directed mutagenesis studies.

The recognition of binding modes and associated pocket residues from our computational work provides more details and better understanding of the binding mechanism of caffeine to the adenosine A_{2A} receptor, which present important insight for

the development of treatments for Parkinson's disease and other diseases associated with the adenosine A_{2A} receptor.

Supporting Information

The Schlitter Formula for entropy is,

$$S = 0.5R\sum_{i} \ln\left[1 + (kTe^2 / \hbar^2)E(i)\right], \tag{5.10}$$

where R is the gas constant, k is the Boltzmann constant, T is the temperature, $\hbar$ is the reduced Planck's constant, and $E(i)$ is the i^{th} eigenvalue of the mass-weighted covariance matrix.

Table 5. 5: PCA calculation results for 4 low-energy binding modes of caffeine and the dominant binding mode of ZM241385.

Complex	Caffeine, binding mode 1	Caffeine, binding mode 2	Caffeine, binding mode 4	Caffeine, binding mode 5	ZM241385, binding mode 2
Entropy (kcal/mol.K)	0.687	0.695	0.662	0.685	0.678

Reference List

1. Fisone, G.; Borgkvist, A.; Usiello, A. Caffeine as a psychomotor stimulant: mechanism of action. *Cellular and Molecular Life Sciences* **2004,** *61* (7-8), 857-872.

2. Huang, Z. L.; Qu, W. M.; Eguchi, N.; Chen, J. F.; Schwarzschild, M. A.; Fredholm, B. B.; Urade, Y.; Hayaishi, O. Adenosine A2(A), but not A(1), receptors mediate the arousal effect of caffeine *Nature Neuroscience* **2005,** *8* (7), 858-859.

3. Ledent, C.; Vaugeois, J. M.; Schiffmann, S. N.; Pedrazzini, T.; ElYacoubi, M.; Vanderhaeghen, J. J.; Costentin, J.; Heath, J. K.; Vassart, G.; Parmentier, M. Aggressiveness, hypoalgesia and high blood pressure in mice lacking the adenosine A(2a) receptor. *Nature* **1997,** *388* (6643), 674-678.

4. Svenningsson, P.; Nomikos, G. G.; Ongini, E.; Fredholm, B. B. Antagonism of adenosine A(2A) receptors underlies the behavioural activating effect of caffeine and is associated with reduced expression of messenger RNA for NGFI-A and NGFI-B in caudate-putamen and nucleus accumbens. *Neuroscience* **1997,** *79* (3), 753-764.

5. Vaugeois, J. M.; El Yacoubi, M.; Costentin, J.; Ledent, C.; Parmentier, M. Mice not aroused by caffeine. *M S-Medecine Sciences* **1997,** *13* (12), 1496-1498.

6. Jaakola, V. P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y. T.; Lane, J. R.; IJzerman, A. P.; Stevens, R. C. The 2.6 Angstrom Crystal Structure of a Human A(2A) Adenosine Receptor Bound to an Antagonist. *Science* **2008,** *322* (5905), 1211-1217.

7. Dunwiddie, T. V.; Masino, S. A. The role and regulation of adenosine in the central nervous system. *Annual Review of Neuroscience* **2001,** *24*, 31-55.

8. Feoktistov, I.; Wells, J. N.; Biaggioni, I. Adenosine A(2B) receptors as therapeutic targets. *Drug Development Research* **1998,** *45* (3-4), 198-206.

9. Hasko, G.; Csoka, B.; Nemeth, Z. H.; Vizi, E. S.; Pacher, P. A(2B) adenosine receptors in immunity and inflammation. *Trends in Immunology* **2009,** *30* (6), 263-270.

10. Jacobson, K. A.; Gao, Z. G. Adenosine receptors as therapeutic targets. *Nature Reviews Drug Discovery* **2006,** *5* (3), 247-264.

11. Kaiser, S. M.; Quinn, R. J. Adenosine receptors as potential therapeutic targets. *Drug Discovery Today* **1999,** *4* (12), 542-551.

12. Manjunath, S.; Sakhare, P. M. Adenosine and adenosine receptors: Newer therapeutic perspective. *Indian Journal of Pharmacology* **2009,** *41* (3), 97-105.

13. Varani, K.; Vincenzi, F.; Tosi, A.; Targa, M.; Masieri, F. F.; Ongaro, A.; De Mattei, M.; Massari, L.; Borea, P. A. Expression and functional role of adenosine receptors in regulating inflammatory responses in human synoviocytes. *British Journal of Pharmacology* **2010,** *160* (1), 101-115.

14. Nikodijevic, O.; Sarges, R.; Daly, J. W.; Jacobson, K. A. Behavioral-Effects of A1-Selective and A2-Selective Adenosine Agonists and Antagonists - Evidence for Synergism and Antagonism. *Journal of Pharmacology and Experimental Therapeutics* **1991,** *259* (1), 286-294.

15. Basheer, R.; Strecker, R. E.; Thakkar, M. M.; McCarley, R. W. Adenosine and sleep-wake regulation. *Progress in Neurobiology* **2004,** *73* (6), 379-396.

16. Porkka-Heiskanen, T.; Alanko, L.; Kalinchuk, A.; Stenberg, D. Adenosine and sleep. *Sleep Medicine Reviews* **2002,** *6* (4), 321-332.

17. Adrien, J. Adenosine and regulation of sleep. *Revue Neurologique* **2001,** *157* (11), S7-S11.

18. Cognato, G. P.; Agostinho, P. M.; Hockemeyer, J.; Muller, C. E.; Souza, D. O.; Cunha, R. A. Caffeine and an adenosine A(2A) receptor antagonist prevent memory impairment and synaptotoxicity in adult rats triggered by a convulsive episode in early life. *Journal of Neurochemistry* **2010,** *112* (2), 453-462.

19. Gongora-Alfaro, J. L. Caffeine as a preventive drug for Parkinson's disease: epidemiologic evidence and experimental support. *Revista de Neurologia* **2010,** *50* (4), 221-229.

20. Varani, K.; Gessi, S.; Dalpiaz, A.; Ongini, E.; Borea, P. A. Characterization of A(2A) adenosine receptors in human lymphocyte membranes by [H-3]-SCH 58261 binding. *British Journal of Pharmacology* **1997,** *122* (2), 386-392.

21. Baraldi, P. G.; Tabrizi, M. A.; Bovero, A.; Avitabile, B.; Preti, D.; Fruttarolo, F.; Romagnoli, R.; Varani, K.; Borea, P. A. Recent developments in the field of A(2A) and A(3) adenosine receptor antagonists. *European Journal of Medicinal Chemistry* **2003,** *38* (4), 367-382.

22. Baraldi, P. G.; Fruttarolo, F.; Tabrizi, M. A.; Preti, D.; Romagnoli, R.; El-Kashef, H.; Moorman, A.; Varani, K.; Gessi, S.; Merighi, S.; Borea, P. A. Design, synthesis, and biological evaluation of C-9- and C-2-substituted pyrazolo[4,3-e]-1,2,4-triazolo[1,5-c] pyrimidines as new A(2A) and A(3) adenosine receptors antagonists. *Journal of Medicinal Chemistry* **2003,** *46* (7), 1229-1241.

23. Varani, K.; Abbracchio, M. P.; Cannella, M.; Cislaghi, G.; Giallonardo, P.; Mariotti, C.; Cattabriga, E.; Cattabeni, F.; Borea, P. A.; Squitieri, F.; Cattaneo, E. Aberrant A(2A) receptor function in peripheral blood cells in Huntington's disease. *FASEB Journal* **2003,** *17* (12), 2148-+.

24. Ye, Y.; Wei, J.; Dai, X.; Gao, Q. Computational studies of the binding modes of A(2A) adenosine receptor antagonists. *Amino Acids* **2008,** *35* (2), 389-396.

25. Poltev, V. I.; Rodriguez, E.; Grokhlina, T. I.; Deriabina, A.; Gonzalez, E. Computational Study of the Molecular Mechanisms of Caffeine Action: Caffeine Complexes With Adenosine Receptors. *International Journal of Quantum Chemistry* **2010,** *110* (3), 681-688.

26. Engel, C. K.; Chen, L.; Prive, G. G. Insertion of carrier proteins into hydrophilic loops of the Escherichia coli lactose permease. *Biochimica et Biophysica Acta-Biomembranes* **2002,** *1564* (1), 38-46.

27. Rosenbaum, D. M.; Cherezov, V.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Yao, X. J.; Weis, W. I.; Stevens, R. C.; Kobilka, B. K. GPCR engineering yields high-resolution structural insights into beta(2)-adrenergic receptor function. *Science* **2007,** *318* (5854), 1266-1273.

28. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G. F.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C. High-resolution crystal structure of an engineered human beta(2)-adrenergic G protein-coupled receptor. *Science* **2007,** *318* (5854), 1258-1265.

29. Jaakola, V. P.; Lane, J. R.; Lin, J. Y.; Katritch, V.; IJzerman, A. P.; Stevens, R. C. Ligand Binding and Subtype Selectivity of the Human A(2A) Adenosine

Receptor Identification and Characterisation of Essential Amino Acid Residues. *Journal of Biological Chemistry* **2010,** *285* (17), 13032-13044.

30. Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *The Journal of Chemical Physics* **1954,** *22* (8), 1420-1426.

31. Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *Journal of Chemical Physics* **2003,** *119* (11), 5740-5761.

32. Brown, S. P.; Muchmore, S. W. High-throughput calculation of protein-ligand binding affinities: Modification and adaptation of the MM-PBSA protocol to enterprise grid computing. *Journal of Chemical Information and Modeling* **2006,** *46* (3), 999-1005.

33. Huo, S. H.; Wang, J. M.; Cieplak, P.; Kollman, P. A.; Kuntz, I. D. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: Insight into structure-based ligand design. *Journal of Medicinal Chemistry* **2002,** *45* (7), 1412-1419.

34. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S. H.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E. Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research* **2000,** *33* (12), 889-897.

35. Kongsted, J.; Ryde, U. An improved method to predict the entropy term with the MM/PBSA approach. *Journal of Computer-Aided Molecular Design* **2009,** *23* (2), 63-71.

36. Kuhn, B.; Kollman, P. A. Binding of a diverse set of ligands to avidin and streptavidin: An accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *Journal of Medicinal Chemistry* **2000,** *43* (20), 3786-3791.

37. Weis, A.; Katebzadeh, K.; Soderhjelm, P.; Nilsson, I.; Ryde, U. Ligand affinities predicted with the MM/PBSA method: Dependence on the simulation method and the force field. *Journal of Medicinal Chemistry* **2006,** *49* (22), 6596-6606.

38. Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. Validation and use of the MM-PBSA approach for drug discovery. *Journal of Medicinal Chemistry* **2005,** *48* (12), 4040-4048.

39. Pearlman, D. A. Evaluating the molecular mechanics Poisson-Boltzmann surface area free energy method using a congeneric series of ligands to p38 MAP kinase. *Journal of Medicinal Chemistry* **2005,** *48* (24), 7796-7807.

40. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. 2004, Gaussian 03, Revision C.02, Gaussian Inc., Wallingford CT.

41. Case D.A.; Darden T.A.; Cheatham T.E.; Simmerling C.L.; Wang J.; Duke R.E.; Luo R.; Crowley M.; Walker R.C.; Zhang W.; Merz K.M.; Wang B.; Hayik S.Roitberg A.; Seabra G.; Kolossvary I.; Wong K.F.; Paesani F.; Vanicek J.; Wu X.; Brozell S.R.; Steinbrecher T.; Gohlke H.; Yang L.; Tan C.; Mongan J.; Hornak V.; Cui G.; Mathews D.H.; Seetin M.G.; Sagui C.; Babin V.; Kollman P.A. 2008, AMBER 10, University of California, San Francisco.

42. Trott, O.; Olson, A. J. Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *Journal of Computational Chemistry* **2010,** *31* (2), 455-461.

43. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics* **1996,** *14* (1), 33-38.

44. Jojart, B.; Martinek, T. A. Performance of the general amber force field in modeling aqueous POPC membrane bilayers. *Journal of Computational Chemistry* **2007,** *28* (12), 2051-2058.

45. Kongsted, J.; Ryde, U. An improved method to predict the entropy term with the MM/PBSA approach. *Journal of Computer-Aided Molecular Design* **2009,** *23* (2), 63-71.

46. Andrew R.Leach The Use of Molecular Modelling and Chemoinformatics to Discover and Design New Molecules. In *Molecular Modelling: Principles and Applications*, Pearson: 2010; pp 640-726.

47. David Chandler Statistical Mechanics. In *Introduction to Modern Statistical Mechanics*, Oxford University Press: 1987; pp 54-85.

48. Haas, J.; Vohringer-Martinez, E.; Bogehold, A.; Matthes, D.; Hensen, U.; Pelah, A.; Abel, B.; Grubmuller, H. Primary Steps of pH-Dependent Insulin Aggregation Kinetics are Governed by Conformational Flexibility. *Chembiochem* **2009,** *10* (11), 1816-1822.

49. Schlitter, J. Estimation of Absolute and Relative Entropies of Macromolecules Using the Covariance-Matrix. *Chemical Physics Letters* **1993,** *215* (6), 617-621.

50. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation* **2008,** *4* (3), 435-447.

51. Yan, X.; Koos, B. J.; Kruger, L.; Linden, J.; Murray, T. F. Characterization of [I-125]ZM 241385 binding to adenosine A(2A) receptors in the pineal of sheep brain. *Brain Research* **2006,** *1096*, 30-39.

52. Yung-Chi, C.; Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochemical Pharmacology* **1973,** *22* (23), 3099-3108.

53. Schwartz, T. W.; Frimurer, T. M.; Holst, B.; Rosenkilde, M. M.; Elling, C. E. Molecular mechanism of 7TM receptor activation - A global toggle switch model. *Annual Review of Pharmacology and Toxicology* **2006,** *46*, 481-519.

54. Kristiansen, K. Molecular mechanisms of ligand binding, signaling, and regulation within the superfamily of G-protein-coupled receptors: molecular

modeling and mutagenesis approaches to receptor structure and function. *Pharmacology & Therapeutics* **2004,** *103* (1), 21-80.

55. Kim, S. K.; Gao, Z. G.; Van Rompaey, P.; Gross, A. S.; Chen, A.; Van Calenbergh, S.; Jacobson, K. A. Modeling the adenosine receptors: Comparison of the binding domains of A(2A) agonists and antagonists. *Journal of Medicinal Chemistry* **2003,** *46* (23), 4847-4859.

56. Kim, J. H.; Wess, J.; Vanrhee, A. M.; Schoneberg, T.; Jacobson, K. A. Site-Directed Mutagenesis Identifies Residues Involved in Ligand Recognition in the Human A(2A) Adenosine Receptor. *Journal of Biological Chemistry* **1995,** *270* (23), 13987-13997.

57. Gao, Z. G.; Jiang, Q. L.; Jacobson, K. A.; IJzerman, A. P. Site-directed mutagenesis studies of human A(2A) adenosine receptors - Involvement of Glu(13) and His(278) in ligand binding and sodium modulation. *Biochemical Pharmacology* **2000,** *60* (5), 661-668.

58. Jiang, Q. L.; Lee, B. X.; Glashofer, M.; Vanrhee, A. M.; Jacobson, K. A. Mutagenesis reveals structure-activity parallels between human A(2A) adenosine receptors and biogenic amine G protein-coupled receptors. *Journal of Medicinal Chemistry* **1997,** *40* (16), 2588-2595.

59. Kim, J. H.; Jiang, Q. L.; Glashofer, M.; Yehle, S.; Wess, J.; Jacobson, K. A. Glutamate residues in the second extracellular loop of the human A(2a) adenosine receptor are required for ligand recognition. *Molecular Pharmacology* **1996,** *49* (4), 683-691.

60. Jiang, Q. L.; Vanrhee, A. M.; Kim, J. H.; Yehle, S.; Wess, J.; Jacobson, K. A. Hydrophilic side chains in the third and seventh transmembrane helical domains of human A(2A) adenosine receptors are required for ligand recognition. *Molecular Pharmacology* **1996,** *50* (3), 512-521.

Chapter 6:

pK_a calculation of Lys115 in Acetoacetate Decarboxylase*

* The content of this chapter is in preparation for publication: Yuli Liu; Steven K. Burger; Paul W. Ayers; pKa calculation of Lys115 in Acetoacetate Decarboxylase.

6.1 Statement of Problem

Ionisable residues play important roles in proteins' structure and function. Protein protonation patterns affect fundamental processes such as protein folding, substrate binding, and enzyme catalysis. The pH-dependent properties of proteins are of great importance in most enzyme-catalyzed reactions.[1] The knowledge of the protonation state, more specifically the pK_a values of key ionisable residues in a protein, is a prerequisite to study enzymatic reaction mechanisms.

Proton binding is strongly influenced by the protein environment, so the pK_a of ionisable residues in the protein can be very different from the pK_a of the amino acid molecules in solution. To rigorously determine the pK_a of an ionisable residue, one needs to compute the free energy difference between the protonated and deprotonated state of the residue in the protein environment. Accurately performing this calculation often requires computing a fictious chemical reaction path whereby the protein is gradually deprotonated. Such calculations are very demanding and can be error-prone, so there are also a variety of models, of varying degrees of empiricism, for computing the pK_a's of the residues in a protein. In this chapter, we use acetoacetate decarboxylase (AADase) as a test case and study many different methods for calculating protein pK_a's.

AADase is a prototypal enzyme for pK_a shifts of active site residues.[11] Lys115 in the active site of AADase is the key residue that catalyzes the decarboxylation reaction for acetone production, and the protonation state of Lys115 is crucial for the activity of the

enzyme. Experimental studies reveal that the pK_a value of Lys115 in AADase has shifted from 10.5 to 5.9.[12,13,14] The recent crystal structure of AADase exhibits a novel protein folding that shows that Lys115 is located in the bottom of a hydrophobic cone, where it is almost entirely solvent inaccessible.[15] Moreover, the crystal structure shows no hydrogen bonds or close-range charge-charge interaction involving Lys115. Most computational methods can provide accurate pK_a prediction for surface residues and residues with small pK_a shifts, but not for a residue in a novel environment like Lys115. Lys115 in AADase provides the computational community a challenge in pK_a prediction methods.

In this chapter we use three different types of pK_a prediction methods to calculate the pK_a of Lys115 in AADase: molecular dynamics/thermodynamic integration (MD/TI) method with implicit solvent, the multiconformation continuum electrostatics (MCCE) method, and the empirical method PROPKA. As we expected, the pK_a prediction of Lys115 depends on the right protonation patterns of other ionisable groups, especially the close-by Glu76. Unfortunately, the above-mentioned pK_a prediction methods do not explicitly sample the protonation patterns of other ionisable residues. When Glu76 is deprotonated, all three methods give the wrong pK_a value for Lys115.

According to previous site-directed mutagenesis studies, the mutation of Glu76 (negatively charged if unprotonated) to Gln (neutral) causes no change in K_m (the Michaelis constant), which suggests that Glu76 has no effect on the pK_a shift of Lys115. We postulate that the pK_a of Glu76 is also shifted so that Glu76 is protonated (neutral) in

AADase. If protonated Glu76 is used in MD/TI calculation, the pK_a of Lys115 is predicted to be 5.3, which agrees well with the experimental value of 5.9.

6.2 Introduction

AADase catalyzed acetone and butanol production (AB fermentation) was one of the first large-scale industrial fermentation processes.[1] The production of butanol in a microbial fermentation was first reported by Pasteur in 1861, and the production of acetone by fermentation was reported by Schardinger in 1905. In 1914, Weizmann developed a culture later named Clostridium acetobutylicum that could produce acetone and butanol at high yields from a variety of starchy substances. This process was widely applied during World War I and World War II due to the large demand of acetone as solvent for the production of nitrocellulose. Production of solvents by the AB fermentation almost completely ceased during the early 1960s. However, in the last decade there has been renewed interest in AADase-catalyzed production of solvents because the biological fermentation process is environmentally friendly and uses renewable source materials..

The mechanism of the enzyme-catalyzed decarboxylation was studied by Westheimer and coworkers starting in the 1940s.[2-5] Inspired by the mechanism of the decarboxylation of dimethylacetoacetic acid catalyzed by aniline[7,8], they proposed that the enzymatic decarboxylation took place through the formation of a Schiff base intermediate between

the enzyme and the substrate[2,6]. Using radioactive acetoacetate and sodium borohydride to label and trap the intermediate, followed by hydrolysis of the resulting protein, a single radioactive peptide was isolated, and therefore the sequence of amino acids in the active site of the AADase was solved.[9,10] Lys115 was recognized as the key catalytic residue involved in the Schiff base formation.

The optimum pH value for AADase catalyzed decarboxylation is around 6. Since the pK_a of the protonated ε-amino group of free lysine is 10.5, Lys115 should be completely protonated at pH 6. However, the proposed catalytic mechanism of AADase requires a free amino group, not an ammonium salt group, as a nucleophile for the formation of Schiff base.[11] Therefore, the pK_a of Lys115 must have shifted in the protein environment. Two independent experimental methods had been performed to measure the pK_a of Lys115 in AADase: the reporter group method[12,13] and a kinetic method[14]. The experimental pK_a value of Lys115 in AADase is reported to be around 5.9, which indicates a large pK_a shift of –4.6.

Many researchers have tried to explain this large pK_a shift. Westheimer proposed that the large pK_a shift of Lys115 resulted from its proximity to the positively charged ε-ammonium group of Lys116, which destabilised the protonated state of Lys115. (This hypothesis is historically significant; it was the first time microenvironment effects were invoked in enzymology.[15]) However, the recent crystal structure of AADase (PDB ID:3BH2, AADase refers to CaAADase in this manuscript) exhibits a previously unknown fold, where Lys115 is located in the bottom of a hydrophobic cone, and Lys116

is facing away from Lys115, and with the two lysine ζ-N separated by 14.8Å.[15] The large distance between the ζ-N atoms implies that the large pK_a shift of Lys115 is not from Coulombic destabilisation, but from the properties of the hydrophobic cone environment.

For many decades AADase had been cited as the prototypical example of pK_a shifts of active site residues.[16] Westheimer's electrostatic microenvironment proposal is widely cited and taught in textbooks. With the unveiling of the AADase crystal structure, Westheimer's proposal about the pK_a shift of Lys115 has been replaced by new hypothesis, which indicates that the pK_a shift is mainly due to the desolvation effects induced by the hydrophobic environment. Confirmation of this hypothesis using computational methods is interesting and challenging.

In the last two decades many methods have been developed to predict the pK_a shifts of the ionisable amino acids. These methods can be categorized into four types: Poisson-Boltzmann equation based methods[17-19], such as MCCE (multiconformation continuum electrostatics)[20-22] and MEAD (macroscopic electrostatics with atomic details)[23,24]; empirical methods[25-27] such as PROPKA; quantum mechanics method on selected cluster models[28]; and the molecular dynamics (MD) or quantum mechanics/molecular mechanics(QM/MM) free energy simulation methods[27,29], such as thermodynamic integration (TI)[30] and free energy perturbation (FEP)[31,32].

In PB equation methods, the protein and the solvent are given separate dielectric constants, and the charge distribution of the protein is usually represented by the partial charges from a molecular mechanics force field. The electrostatic potential is determined

by solving the PB equation and the work of mutating the ionisable groups from the solvent to the protein environment (also called solvation energy) can be solved for calculating the pK_a shift from solvent to protein environment**.** PB equation-based methods can yield accurate pK_a values in some cases, but they are dependent on the charge distribution and protein conformation and they are sensitive to the choice of protein dielectric constant.

Empirical methods start with the factors that are believed to influence protein pK_a values (e.g., in PROPKA, the key terms are hydrogen bonds, desolvation effects, and charge-charge interactions), then parameterize these factors based on experimentally available pK_a values. The accuracy of empirical methods depends on fitting procedure and the choice of reference protein pK_a values to which the parameters are fit. Such methods can be unreliable for residues in environments very different from any residue in the training set.

The MD and QM/MM free energy simulation methods calculate the deprotonation free energy difference of the ionisable groups in solution and in the protein environment, and accordingly give the pK_a shifts from aqueous solution to the protein environment. The free energy simulation methods are the most fundamental ways to calculate the pK_a values. The MD and QM/MM free energy simulation methods provide a fully atomistic simulation of the entire system. The protein electrostatics and other nonbonded interactions are modeled explicitly. These are also the most expensive methods.

In this manuscript we report our attempts to use molecular dynamics/thermodynamic integration (MD/TI), MCCE and PROPKA to calculate the pK_a shift of Lys115 in AADase. In the following Section 2, we give a detailed description of each method. Then we present our calculation results in Section 3, followed by comparison and discussion of the results in Section 4. In section 5 we conclude that the pK_a shift of Lys115 is most likely due to the fact that Glu76 is protonated in AADase.

6.3 Methods

6.3.1 Molecular dynamics/thermodynamic integration (MD/TI) for calculating pK_a shifts

6.3.1.1 The calculation of pK_a shifts and free energy differences

The pK_a value is proportional to the deprotonation free energy of an ionisable group,

$$pK_a = \frac{\Delta G}{2.3026RT}. \tag{6.1}$$

The deprotonation free energy is the free energy difference between the unprotonated and protonated states of the system.

According to the thermodynamic cycle (Figure 6.1), the pK_a shift (ΔpK_a) of an ionisable group from the aqueous solution to the protein environment can be denoted as the difference in the deprotonation free energies ($\Delta\Delta G$),

$$\Delta pK_a = \frac{\Delta\Delta G}{2.3026RT} = \frac{\Delta G_p - \Delta G_s}{2.3026RT}, \quad (6.2)$$

where ΔG_p is the deprotonation free energy in the protein environment, that is, the free energy difference between the unprotonated and protonated state of the protein. The protein is solvated in implicit or explicit water. ΔG_s is the deprotonation free energy of the model compound in solution, where again solvation effects can be modelled with either explicit (molecular mechanics-based) or implicit (continuum solvation) water models. The model compound is defined as the amino acid in a dipeptide chain, and both sides of the backbone of the dipeptide chain are capped with methyl groups (Figure 6.2).

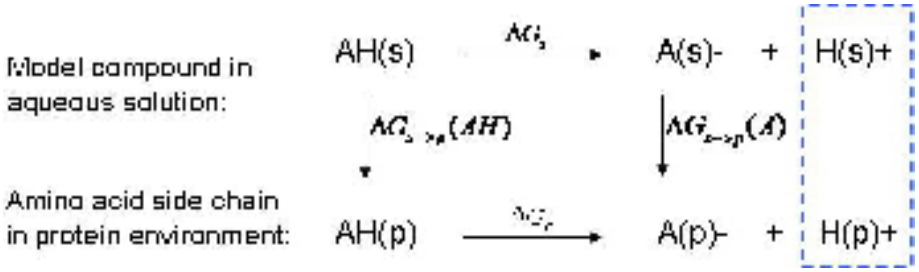


Figure 6. 1 The thermodynamic cycle

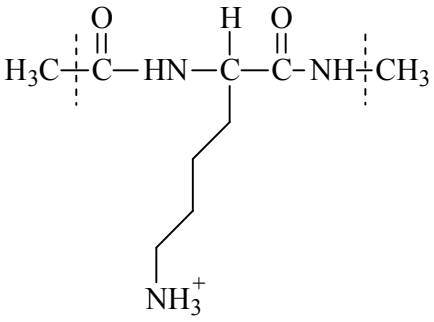


Figure 6. 2 The model compound (ACE-LYS-NME)

6.3.1.2 Thermodynamic integration (TI) for calculating the deprotonation free energies

The free energy difference between the protonated (initial) state and deprotonated (perturbed) state of the ionisable group can be calculated using thermodynamic integration. Denote the energy function of the initial state as V_i and the energy function of the final perturbed state as V_f. As we gradually mutate the system from the initial state to the final state, the energy function of the system can be represented as a linear combination of V_i and V_f through a coupling parameter λ,

$$V(\lambda) = (1-\lambda)V_i + \lambda V_f . \tag{6.3}$$

Changing λ from 0 to 1 represents mutation from state i to f. The values of λ correspond to a hybrid system that consists of a mixture of the initial and final energy functions. The free energy difference (Gibbs free energy for NPT ensemble or Helmholtz free energy for NVT ensemble) between state i and f is given by,

$$\Delta G = \int_{\lambda=0}^{\lambda=1} \left(\frac{\partial H}{\partial \lambda} \right)_{\lambda} d\lambda = \int_{\lambda=0}^{\lambda=1} \left\langle \frac{dV}{d\lambda} \right\rangle_{\lambda} d\lambda , \tag{6.4}$$

where the brackets indicate ensemble average over the MD simulations for a given λ value. In practice a numerical integration method (typically Gaussian quadrature) is used to calculate the integral in Eq. (6.4):

$$\Delta G = \sum_i \omega_i \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_i , \tag{6.5}$$

In this paper the energy function is modelled using a molecular mechanics potential-energy function. The potential function involves the bonded terms, electrostatic terms, and van der Waals terms. For the deprotonation process, the bonded terms associated with the disappearing of the proton are internal to the amino acid, so they are the same for the model compound and the protein, and consequently the bonded terms are cancelled out in calculating $\Delta\Delta G$ and ΔpK_a. The van der Waals terms are negligible because the molecular mechanics force field assigns small Lennard-Jones parameters to hydrogen atoms. Therefore, most of the deprotonation free energy arises from the electrostatic changes: the loss of a positive charge from deprotonation and the rearrangement of charges on the remaining atoms in the amino acid upon protonation.

The deprotonation process can be simply modelled as the loss of one positive charge and the rearrangement of partial charges on the atoms of the amino acids as shown in Figure 6.3. The numbers outside of the parentheses are the partial charges on the atoms of the amino acid in the protonated form (denoted LYS in Amber force field), and the numbers inside the parentheses are the partial charges on the atoms of the deprotonated amino acid (denoted LYN in Amber force field).

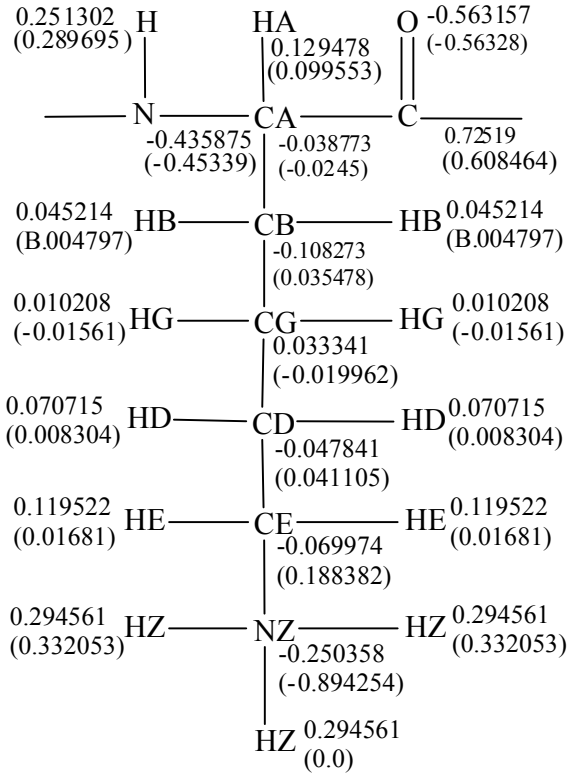


Figure 6. 3 Mutation of charges on the lysine group during thermodynamic integration. Charges outside of the parenthesis are the charges on protonated lysine group, charges inside the parenthesis are the charges on deprotonated lysine group.

6.3.1.3 Thermodynamic integration with Glu76 unprotonated

The thermodynamic integration simulations are based on 5 λ values ($\lambda = 0, 0.1127, 0.5, 0.8873, 1$). According to the weights of Gaussian quadrature ($\omega = 0, 0.27777, 0.44444, 0.27777, 0$, corresponding to the 5 λ values), the weights for the two end points ($\lambda = 0$ and $\lambda = 1$) are 0, so actually we only do simulations on three windows $\lambda = 0.1127, 0.5, 0.8873$. The protein deprotonation free energy is calculated based on MD simulations on chain A of the AADase crystal structure (PDB ID: 3BH2). All other ionisable groups in the protein are assigned to their natural protonation states at pH=6.0 (the optimum pH value for AADase catalysis). For example, all glutamic residues are unprotonated, including Glu76 (pK_a of free glutamic acid is 4.07); all arginines are protonated (pK_a of free arginine is 12.48), all histidines are doubly protonated (pK_a of histidine is 6.04), etc. The model compound deprotonation free energy is calculated based on MD simulations on the model compound as shown in Figure 6.2. The simulations were performed in generalized Born (GB) implicit solvent model[33] using the *Amber 10* program package[34] and the standard FF03 Amber force field[33,35]. For each λ value the system was equilibrated for 150ps with 10kcal/mol harmonic restraint on the protein backbone and the restraint was reduced to 1kcal/mol for the 3ns production phase. The *TLEAP* module in *Amber 10* program package is used to add hydrogens, counterions and prepare for parameter files and input geometry files. SHAKE algorithm is used to

freeze bond length involving hydrogen atoms so that a time step of 2fs can be used for the MD simulation.

For each λ value, the corresponding $\left\langle \frac{dV}{d\lambda} \right\rangle_{\lambda}$ is calculated based on the MD simulation over 3ns. The deprotonation free energies of the model compound and the protein can then be calculated using the Gaussian quadrature formula as shown in Eq. (6.5).

6.3.1.4 Thermodynamic integration with Glu76 protonated

Previous site-directed mutagenesis study has shown that Glu76Gln mutant has the same K_m as the wild type AADase, indicating that replacing Glu76 with an neutral residue gives no pK_a shift for Lys115.[15] Glu76 is 4.3Å away from the ζ-N atoms of Lys115 in chain A. If it is unprotonated as other pK_a prediction methods indicate,[36] the negative charge on Glu76 will favour the protonated state of lysine. Therefore we postulate that the pK_a of Glu76 is also shifted so that Glu76 is protonated in AADase. In this MD/TI calculation, the protonation patterns of other ionisable groups are the same as the MD/TI calculation in Section 6.3.1.3, except for Glu76. The MD simulations were performed using the same protocol for thermodynamic integration.

6.3.2 pK_a calculation using MCCE

Poisson-Boltzmann (PB) equation based methods calculate the pK_a shift of an ionisable group from aqueous solution to the protein environment by estimating the free energy difference of moving the unprotonated and protonated states of the ionisable group from aqueous solution to the protein environment. According to the thermodynamic cycle (Figure 6.1),

$$\Delta pK_a = \frac{\Delta\Delta G}{2.3026RT} = \frac{\Delta G_p - \Delta G_s}{2.3026RT} = \frac{\Delta G_{s\to p}(A) - \Delta G_{s\to p}(AH)}{2.3026RT}, \quad (6.6)$$

where $\Delta G_{s\to p}$ is the free energy change from moving the ionisable group from the aqueous solution to the protein environment. (The unprotonated state is denoted A and the protonated state is denoted AH.) $\Delta G_{s\to p}$ is often called the reaction field energy or solvation energy. $\Delta G_{s\to p}$ is estimated using the electrostatic potential obtained by solving PB equation,

$$\Delta G_{s\to p} = \frac{1}{2}\sum_i q_i(\phi_i^{\varepsilon} - \phi_i^{80}). \quad (6.7)$$

In PB equation based methods the protein is usually defined as a region of a low dielectric constant ($\varepsilon = 4 \sim 20$) and surrounded by solvent water with a high dielectric constant (80). The accuracy of PB equation based methods requires that one choose an appropriate protein dielectric constant and a representative conformation of the protein. In MCCE method, the side chain flexibility from extra hydrogen bond orientations (e.g., due to the ambiguousness of O and N atom positions in the amide group of Asn and Gln,

or the C and N atom positions in the imidazol ring of His from the crystal structure) and hydroxyl rotamers are considered when generating conformers of the protein. The preselected conformers are subjected to Monte Carlo sampling to generate a Boltzmann distribution of conformers. One conformer of each residue constitutes a microstate. Instead of solving the PB equation on one single conformation (e.g., the crystal structure), the PB equations are solved for M conformers using the standard PB solver Delphi[37,38]. And an $M \times M$ conformer-conformer pairwise electrostatic interaction matrix is obtained; this allows us to determine the dielectric boundary for each microstate. More details about the MCCE method can be found in Ref 21,22.

QUICK MCCE calculation only makes isosteric rotamers by swapping O with N atoms in the amide group of Asn and Gln, or CD2 with ND1 and CE2 with NE1 atoms in His, so there are about 2.5 conformers per residue. FULL MCCE calculation considers all possible side chain flexibilities and makes about 20 conformers per residue (about 50 conformers per ionisable residue, 15 conformers per polar residue and 5 conformers per non-polar residue). Both QUICK and FULL conformer searches are performed on chain A of the crystal structure of AADase using MCCE2.4 program package[22]. For both QUICK and FULL MCCE calculations the pK_a values are calculated using two different dielectric constants of the protein: 4 and 8.

6.3.3 pK_a calculation using PROPKA 2.0

PROPKA determines the pK_a of ionisable group empirically by parameterizing the following determinants: global (GlobalDes) and local (LocalDes) desolvation effects, side-chain (SC-HB) and backbone (BB-HB) hydrogen bonds, and Coulomb interactions between charged groups (chg-chg), as shown in Eq. (6.8),

$$\Delta pK_a = \Delta pK_{GlobalDes} + \Delta pK_{LocalDes} + \Delta pK_{SC-HB} + \Delta pK_{BB-HB} + \Delta pK_{chg-chg} . \quad (6.8)$$

The formulae for the pK_a determinants and the associated parameters can be found in Ref. 25,26.

The pK_a of Lys115 was first calculated based on chain A of the crystal structure of AADase alone. A comparison of all 4 monomers in the crystal structure (PDB ID: 3BH2) has shown that there are some differences in the active site, probably due to the resolution of X-ray crystallography. We also performed PROPKA calculation on the other 3 monomers (chains B, C and D) in the crystal structure. Since AADase is believed to be a dodecamer in solution, we built the dodecamer of AADase from the tetramer structure provided in the crystal structure, based on the symmetry information provided in the PDB file. PROPKA calculations were performed on the dodecamer, and then the average pK_a value was obtained from the twelve monomers.

6.4 Results and Discussion

6.4.1 Results from MD/TI pK_a calculations

6.4.1.1 With Glu76 unprotonated

The $\langle dU/d\lambda \rangle_\lambda$ value for each λ is shown in Table 6.1 for the model compound and AADase. The difference between the deprotonation free energies of the model compound and AADase is -1.14kcal/mol, therefore the pK_a shift of Lys115 from the aqueous solution to the protein environment is predicted to be -0.83. Even though Glu76 is unprotonated and carries a negative charge, the pK_a shift of Lys115 is in the right direction, but with a large error compared to the experimental pK_a shift of -4.6.

Table 6. 1: MD/TI results with Glu76 unprotonated

λ	ω	$\langle dU/d\lambda \rangle_\lambda$ over 3ns (kcal/mol)	
		Model compound	AADase
0.1127	0.27777	14.95	17.61
0.5	0.44444	14.59	13.44
0.8873	0.27777	14.22	9.30
ΔG		14.59	13.45
$\Delta\Delta G$		-1.14	
ΔpK_a		-0.8	

6.4.1.2 With Glu76 protonated

Results from MD/TI with Glu76 protonated are shown in Table 6.2. The difference between the deprotonation free energies of the model compound and the AADase is

calculated as -7.22kcal/mol using Gaussian quadrature formula in Eq. (6.5), and the pK_a shift of Lys115 from the aqueous solution to the protein environment is calculated as -5.3 according to Eq.(6.2). We can see that, with Glu76 protonated (neutral), the pK_a shift of Lys115 matches well with the experimental value of -4.6. This result is supported by the site-directed mutagenesis study that showed that replacing Glu76 by a neutral residue didn't change the pK_a.

Table 6. 2: MD/TI results with Glu76 protonated

λ	ω	$\langle dU/d\lambda \rangle_\lambda$ over 6ns (kcal/mol)	
		Model compound	AADase
0.1127	0.27777	14.95	10.63
0.5	0.44444	14.59	7.38
0.8873	0.27777	14.22	4.17
ΔG		14.59	7.39
$\Delta\Delta G$		-7.2	
ΔpK_a		-5.2	

6.4.2 Results from MCCE calculations

The results of QUICK and FULL MCCE calculations with dielectric constants 4 and 8 are shown in Table 6.3. With MCCE method a dielectric constant of 4 is suggested for large proteins (>200 residues), and a dielectric constant of 8 is suggested for small soluble proteins. AADase is a large protein with 244 residues in one chain. As we can see in Table 6.3, FULL MCCE calculation with a dielectric constant 4 gives the best pK_a

value of 8.90, with a pK_a shift of -1.6. The extreme sensitivity of MCCE to dielectric constant is disconcerting, and was also noted by Ishikita.[36]

Table 6. 3: pK_a values of Lys115 from MCCE calculations.

Protein Dielectric Constants	QUICK	FULL
$\varepsilon = 4$	9.4	8.9
$\varepsilon = 8$	12.0	12.5

6.4.3 Results from PROPKA calculations

AADase is a homododecameric enzyme. The crystal structure (PDB ID: 3BH2) provides a tetramer structure, with its monomer labelled as chain A, B, C and D, respectively. Using the symmetry information and the tetramer structure provided in the PDB file, we built the dodecamer using PyMol. PROPKA calculations were performed on the 4 single chains in separate forms and on the dodecamer (see Table 6.4). The average pK_a value of Lys115 from the single chain is 9.3, with a pKa shift of -0.8; and the average pK_a value of Lys115 in the dodecamer is 8.1, with a pKa shift of -2.4 (Table 6.4).

Table 6. 4: pK_a values of Lys115 from PROPKA calculations.

Protein Environment	Chain	pK_a value	pK_a shift	Desolvation		Coulomb (Glu76)	Average pK_a value	Average pK_a shift
				Global	Local			
Single chain	A	9.8	-0.7	-2.22	-0.63	2.13	9.3	-0.8
	B	9.8	-0.7	-2.58	-0.49	2.40		
	C	8.6	-1.9	-2.26	-0.42	0.78		
	D	9.1	-1.4	-2.23	-0.42	1.20		
Dodecamer	A	8.5	-2.0	-3.46	-0.63	2.13	8.1	-2.4
	B	8.9	-1.6	-3.52	-0.49	2.40		
	C	7.4	-3.1	-3.44	-0.42	0.78		
	D	7.8	-2.7	-3.45	-0.42	1.20		

The breakdown of the contributions to the PROPKA results is shown in Table 6.4. This reveals that the pK_a shift of Lys115 is due to the Coulomb interaction with the negative charged residue Glu76 in the close proximity and due to global and local desolvation effects. There is no backbone or side-chain hydrogen bond associated with the amino group of Lys115, and there are only two charged groups within 10Å of NZ of Lys115: the negative Glu76 (if unprotonated) and the positive Arg29 (if protonated). The positive NH1 and NH2 atoms of Arg29 are over 7Å away, so according to the empirical PROPKA formula for charge-charge interaction, the unfavourable charge-charge interaction with the protonated form of Lys115 is negligible. The distance of Glu76 to Lys115 in the 4 monomers varies from 2.7Å (chain B) to 5.0Å (chain C), which explains the pK_a difference between the monomers. The major contribution to Lys115 pK_a shift is the desolvation effects. The global and local desolvation effects favour the pKa shift (−4.0 contribution on average), but the Coulomb interaction from Glu76 is against the

pKa shift (1.6 contribution by average). Because PROPKA uses the natural protonation state of other ionisable residues, the negative charge on the supposedly unprotonated Glu76 stablizes the protonated form of Lys115. Previous site-directed mutagenesis studies on charged residues (Arg29/Glu76 to Gln) indicate that these residues do not have significant impact on the pK_a shift of Lys115. The contribution of 1.6 pK_a unit of from Glu76 in the PROPKA results does not agree with the site-directed mutagenesis results, which casts doubt on the protonation state of Glu76. If we leave out the Coulomb interaction with Glu76, as suggested by Ho based on the site-directed mutagenesis studies,[15] then the pK_a shift from desolvation effects only (−4.0) is very close to the experimental pK_a shift of −4.6.

The difference in the pK_a shift of Lys115 in a single AADase chain and in the dodecamer is from the global desolvation effect. As shown in the PROPKA results, Lys115 is obviously buried deeper in the dodecamer than in the monomer.

6.5 Comparison and discussion of results

In this study, we calculated the pK_a shift of Lys115 in AADase, a buried residue in a hydrophobic environment using three different types of methods: MD/TI with implicit and explicit solvent, MCCE and PROPKA. In previous benchmarking pK_a prediction study[39], the MD/TI method with implicit solvent model, MCCE and PROPKA all achieved an overall RMSD of 1.4.[39] Another benchmarking study compared PROPKA to

the results from the PB equation based methods MCCE, MEAD, and UHBD.[40] They claim that among these methods, PROPKA is more accurate for Asp, Glu, Lys and Tyr with RMSD values of 0.934, 0.849, 0.260 and 1.001, while MCCE is more accurate for His with an RMSD of 1.522.

When using the natural protonation pattern of other ionisable residues, MD/TI calculation with implicit solvent model, FULL MCCE calculation with dielectric constant 4, and PROPKA calculations on the biologic unit (the dodecamer) all predicted the correct direction of the pK_a shift direction. However, none of these methods predicted the right protonation state of Lys115 in AADase (Table 5). If we assume that Glu76 is protonated (neutral), then in MD/TI calculation, the pK_a of Lys115 is calculated as 5.3 (highlighted in Table 6.5), which gives the right protonation state for Lys115 at the optimum catalysis pH value of 6.0 of AADase. Without the contribution of the negative charge from Glu76, PROPKA predicts the pK_a of Lys115 to be 6.5, which is also close to the experimental pK_a value of 5.9. We can postulate that the pK_a of Glu76 is shifted so that Glu76 is protonated in AADase. We are running calculations to test this hypothesis.

Table 6. 5: Comparison of results from different methods

Methods	pK_a of Lys115	ΔpK_a	Experimental pK_a Value	Experimental pK_a shift
MD/TI with Glu76 unprotonated	9.7	-0.8	5.9	-4.6
MD/TI with Glu76 protonated	**5.3**	**-5.2**		
FULL MCCE with $\varepsilon = 4$	8.9	-1.6		
PROPKA average on the dodecamer	8.1	-2.4		

In our pK_a calculations the protonation patterns cannot be explicitly sampled according to site-site interactions of all ionisable residues. Instead the protonation states of all other ionisable residues are pre-assigned according to the pK_a values of the corresponding free amino acids (referred as the natural protonation pattern)**.** Since the pK_a values of other ionisable residues might shift due to the protein environment, the assignment of protonation states could be wrong, which could lead to wrong pK_a prediction for the target residue. For example, PROPKA results show that Glu76 contributes 1.6 pK_a units to the pK_a shift of Lys115, provided that Glu76 is unprotonated and carries a negative charge. But according to site-directed mutagenesis studies, Glu76 does not show significant effect on the pK_a shift of Lys115, which indicate that Glu76 is protonated. In this case, we have to adjust the protonation state of Glu76 in our MD/TI calculation to achieve reliable results.

A recent computational work on calculating the pK_a value of Lys115 in AADase claimed that the pK_a of Lys115 was calculated as 5.73 using MEAD with protein dielectric constant of 4.[36] In MEAD calculation, the pK_a of the target residue is broken into two parts: the first part is pH-independent, including the Born solvation energy and the interaction of the charge on the target residue and the background charge from non-ionisable residues, with all other ionisable residues neutralized; the second part includes the interaction between the target residue and other ionisable residues. In the second part, the ensemble of the protonation patterns of all other ionisable residues is sampled using MD or MC simulation. The sampling of protonation patterns is referred as titration. MEAD calculation circumvents the problem of pre-assigned protonation patterns of ionisable residues in the protein. But it still has other problems. For example, due to the large computational cost of sampling the protonation patterns of all ionisable residues, usually only the protonation patterns of some close residues are sampled. The calculation results are very sensitive to the choice of ionisable residues. According to our experience, excluding one residue from the titration process, the pK_a value of the target residue could vary from 0.1 to 7.5. The MEAD calculation result is very sensitive to the choice of protein dielectric constant as well. According to Ishikita,[36] with protein dielectric constant of 4, the pK_a of Lys115 in AADase was calculated as 5.73, which agrees very well with the experimental value of 5.9. But with protein dielectric constant of 6, the pK_a of Lys115 was calculated as 8.04, which does not predict the right protonation state of Lys115. According to Ishikita's MEAD calculation results,[36] Glu76 is unprotonated and

has 6.5pK_a unit contribution to the pK_a shift of Lys115, but this contribution is neutralized by the -4.4pKa unit contribution of Arg29. These results do not seem consistent with the site-directed mutagenesis study.

The PROPKA calculations reveal that it is important to include the full biological unit, and not just the protein monomer when using empirical protein pK_a models. This is true even for residues like Lys115, which is not especially close to the surface of the monomer. This is in contrast to standard practice, where only the monomer is used. While the predicted pK_a shift from PROPKA was too small, it was in the right direction, and results were closer when the full dodecamer was used. PROPKA results are in good agreement with experiment when the electrostatic contribution from the negative charge on Glu76 is omitted from the calculation. This supports our hypothesis that the large pK_a shift in Lys115 is partly due to desolvation and partly due to the fact Glu76 is protonated in AADase.

6.6 Conclusion

Using the recently solved crystal structure, the pK_a value of Lys115 in AADase has been calculated using three different kinds of pK_a prediction methods: MD/TI, MCCE and PROPKA. Among those the MD/TI calculation with protonated Glu76 gave the best result compared to the experimental value.

We postulate that the large errors are due to the site-site interactions from other ionisable residues whose protonation patterns are not sampled in our calculations. This can be overcome with MD/TI by running TI with different protonation state of nearby ionisable residues. Motivated by the site-directed mutagenesis study, we postulate that Glu76 is protonated, and used protonated Glu76 for the MD/TI calculation. The resulting pK_a value of 5.3 for Lys115 agrees well with the experimental value of 5.9.

Reference List

1. Jones, D. T.; Woods, D. R. Acetone-Butanol Fermentation Revisited. *Microbiological Reviews* **1986,** *50* (4), 484-524.

2. Hamilton, G. A.; Westheimer, F. H. On the mechanism of the enzymatic decarboxylation of acetoacetate. *Journal of the American Chemical Society* **1959,** *81* (23), 6332-6333.

3. Seltzer, S.; Hamilton, G. A.; Westheimer, F. H. Isotope Effects in the Enzymatic Decarboxylation of Oxalacetic Acid. *Journal of the American Chemical Society* **1959,** *81* (15), 4018-4024.

4. Steinberger, R.; Westheimer, F. H. THE METAL ION CATALYZED DECARBOXYLATION OF DIMETHYLOXALOACETIC ACID. *Journal of the American Chemical Society* **1949,** *71* (12), 4158-4159.

5. Westheimer, F. H.; Jones, W. A. The Effect of Solvent on Some Reaction Rates. *Journal of the American Chemical Society* **1941,** *63* (12), 3283-3286.

6. Fridovich, I.; Westheimer, F. H. On the Mechanism of the Enzymatic Decarboxylation of Acetoacetate. II. *Journal of the American Chemical Society* **1962,** *84* (16), 3208-3209.

7. Pedersen, K. J. Studies of Complex Formation between Aniline and Picrate Ion by Solubility Measurements. *Journal of the American Chemical Society* **1934,** *56* (12), 2615-2619.

8. Pedersen, K. J. Amine Catalysis of the Ketonic Decomposition of Dimethylacetoacetic Acid. *Journal of the American Chemical Society* **1938,** 60 *(*3), 595-601.
9. Lederer, F.; Coutts, S. M.; Laursen, R. A.; Westheimer, F. H. Acetoacetate Decarboxylase. Subunits and Properties*. Bi*ochemistry* **1966, 5** *(*3), 823-833.

10. Laursen, R. A.; Westheimer, F. H. The Active Site of Acetoacetate Decarboxylase. Jo*urnal of the American Chemical Society* **1966, 8**8 *(*14), 3426-3430.

11. Westheimer, F. H. Coincidences, Decarboxylation, and Electrostatic Effects. Te*trahedron* **1995, 5**1 *(*1), 3-20.

12. Kokesh, F. C.; Westheimer, F. H. Reporter group at the active site of acetoacetate decarboxylase. II. Ionization constant of the amino group. Jo*urnal of the American Chemical Society* **1971,** 93 *(*26), 7270-7274.

13. Frey, P. A.; Kokesh, F. C.; Westheimer, F. H. Reporter group at the active site of acetoacetate decarboxylase. I. Ionization constant of the nitrophenol. Jo*urnal of the American Chemical Society* **1971,** 93 *(*26), 7266-7269.

14. Westheimer, F. H.; Schmidt, D. E. pK of the lysine amino group at the active site of acetoacetate decarboxylase. *Biochemistry* **1971,** 10 *(*7), 1249-1253.

15. Ho, M. C.; Menetret, J. F.; Tsuruta, H.; Allen, K. N. The origin of the electrostatic perturbation in acetoacetate decarboxylase. *Nature* **2009,** 45*9 (*7245), 393-U107.

16. Highbarger, L. A.; Gerlt, J. A.; Kenyon, G. L. Mechanism of the reaction catalyzed by acetoacetate decarboxylase. Importance of lysine 116 in determining the pK(a) of active-site lysine 115. *Biochemistry* **1996,** 35 *(*1), 41-46.

17. Yang, A. S.; Honig, B. On the Ph-Dependence of Protein Stability. *Journal of Molecular Biology* **1993,** 23*1 (*2), 459-474.

18. Bashford, D.; Karplus, M. Pkas of Ionizable Groups in Proteins - Atomic Detail from A Continuum Electrostatic Model. *Biochemistry* **1990,** 29 *(*44), 10219-10225.

19. Yang, A. S.; Gunner, M. R.; Sampogna, R.; Sharp, K.; Honig, B. On the Calculation of Pk(A)S in Proteins. *Proteins-Structure Function and Genetics* **1993,** 15 *(*3), 252-265.

20. Alexov, E. G.; Gunner, M. R. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophysical Journal* **1997,** 72 *(*5), 2075-2093.

21. Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. *Biophysical Journal* **2002, 8**3 *(*4), 1731-1748.

22. Song, Y. F.; Mao, J. J.; Gunner, M. R. MCCE2: Improving Protein pK(a) Calculations with Extensive Side Chain Rotamer Sampling. *Journal of Computational Chemistry* **2009, 3**0 *(*14), 2231-2247.

23. Bashford, D.; Gerwert, K. Electrostatic Calculations of the Pka Values of Ionizable Groups in Bacteriorhodopsin. *Journal of Molecular Biology* 19**92, 2**24 *(*2), 473-486.

24. Bashford, D. Macroscopic electrostatic models for protonation states in proteins. *Frontiers in Bioscience* **2004, 9**, 1082-1099.

25. Bas, D. C.; Rogers, D. M.; Jensen, J. H. Very fast prediction and rationalization of pK(a) values for protein-ligand complexes. *Proteins-Structure Function and Bioinformatics* **2008, 7**3 *(*3), 765-783.

26. Li, H.; Robertson, A. D.; Jensen, J. H. Very fast empirical prediction and rationalization of protein pK(a) values. *Proteins-Structure Function and Bioinformatics* **2005, 6**1 *(*4), 704-721.

27. Li, H.; Robertson, A. D.; Jensen, J. H. Protein PKA predictions: From QM/MM/LPBE and QM/LPBE methods to an empirical method. *Abstracts of Papers of the American Chemical Society* **2004, 2**2*8,* U230.

28. Li, H.; Robertson, A. D.; Jensen, J. H. The determinants of carboxyl pK(a) values in Turkey ovomucoid third domain. *Proteins-Structure Function and Bioinformatics* **2004, 5**5 *(*3), 689-704.

29. Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. Prediction and rationalization of protein pK(a) values using QM and QM/MM methods. *Journal of Physical Chemistry A* **2005, 1**0*9 (*30), 6634-6643.

30. Simonson, T.; Carlsson, J.; Case, D. A. Proton binding to proteins: pK(a) calculations with explicit and implicit solvent models. *Journal of the American Chemical Society* **2004, 1**2*6 (*13), 4167-4180.

31. Kaukonen, M.; Soderhjelm, P.; Heimdal, J.; Ryde, U. Proton transfer at metal sites in proteins studied by quantum mechanical free-energy perturbations. *Journal of Chemical Theory and Computation* **2008, 4** *(*6), 985-1001.

32. Li, G. H.; Cui, Q. pK(a) calculations with QM/MM free energy perturbations. *Journal of Physical Chemistry B* **2003, 1**0*7 (*51), 14521-14528.

33. Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins-Structure Function and Bioinformatics* **2004, 5**5 *(*2), 383-394.

34. Case D.A.; Darden T.A.; Cheatham T.E.; Simmerling C.L.; Wang J.; Duke R.E.; Luo R.; Crowley M.; Walker R.C.; Zhang W.; Merz K.M.; Wang B.; Hayik S.Roitberg A.; Seabra G.; Kolossvary I.; Wong K.F.; Paesani F.; Vanicek J.; Wu X.; Brozell S.R.; Steinbrecher T.; Gohlke H.; Yang L.; Tan C.; Mongan J.; Hornak V.; Cui G.; Mathews D.H.; Seetin M.G.; Sagui C.; Babin V.; Kollman P.A. 2008, AMBER 10, University of California, San Francisco.

35. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G. M.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J. M.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry* **2003, 2**4 *(*16), 1999-2012.

36. Hiroshi Ishikita . Origin of the pKa $_{sh}$ift of the catalytic lysine in acetoacetate decarboxylase. *FEBS Letter* 584, 3464-3468. **2010**.

37. Nicholls, A.; Honig, B. A Rapid Finite-Difference Algorithm, Utilizing Successive Over-Relaxation to Solve the Poisson-Boltzmann Equation. *Journal of Computational Chemistry* **1991, 1**2 *(*4), 435-445.

38. Rocchia, W.; Alexov, E.; Honig, B. Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *Journal of Physical Chemistry B* **2001, 1**0*5 (*28), 6507-6514.

39. Stanton, C. L.; Houk, K. N. Benchmarking pK(a) prediction methods for residues in proteins. *Journal of Chemical Theory and Computation* **2008, 4** *(*6), 951-966.

40. Davies, M.; Toseland, C.; Moss, D.; Flower, D. Benchmarking pKa prediction. *BMC Biochemistry* **2006, 7** *(*1), 18.

41. Burger, S. K.; Ayers, P. W. A parameterized, continuum electrostatic model for predicting protein pKa values. *Proteins* **2011, 7**9 *(*7), 2044-2052.

Chapter 7:

Summary and Future Work

7.1 Summary

This thesis represents my Ph.D. work on developing new methods to elucidate chemical reactions. Three different methods were discussed. For small molecules, it is computationally feasible to find minimum energy reaction path on the potential energy surface (PES), with the PES computed from quantum mechanical models. This gives very detailed information about the mechanism of the chemical reaction. Biological systems are too large, and entropic effects are too important, for the full reaction path to be determined. Instead, free-energy differences between key structures are computed by sampling the PES with molecular dynamics (MD); in these applications, the PES is modelled using the ball-and-spring-type models known as molecular mechanics.

Chapter 2 and chapter 3 introduce the fast marching method (FMM) for finding the minimum energy path (MEP). FMM is shown to be one of the most general and reliable surface-walking algorithms for finding MEP. Without any prior knowledge about the PES, it can always find the global MEP. Unfortunately, FMM is an expensive method. Therefore, in chapters 2 and 3, some improvements to the original FMM method were made, increasing its accuracy and efficiency.

Chapter 4 presents the QSM-NT method for finding all stationary points on the PES. Usually the path-finding methods can only find one reaction path, and the reliability of the results depends on the initial guess. QSM-NT can find all stationary points, accordingly all alternative reaction paths, which could be a great advantage while

studying reactions with several alternative reaction mechanisms or when trying to analyze and compare different postulated mechanisms. QSM-NT was proven to be efficient and reliable through successful applications to analytical potentials and chemical reactions, however it is not a "black box" method and, unlike FMM, sometimes fails.

For complex biological systems, the properties of the system can no longer be represented by a single state, but by averaging over all possible microstates consistent with given restraints instead. Statistical sampling methods, such as MD simulation, are used to calculate the ensemble average of the system. Chapter 5 and chapter 6 present our work on studying biological processes using MD simulation. In these cases, the reaction path is not found; instead only the free-energy difference between key stationary points on the free-energy surface is determined. Chapter 5 computes a simple free-energy of binding where no covalent bonds are created or formed. Chapter 6 looks at a simple chemical reaction, the deprotonation of an amino acid in an enzyme.

Chapter 5 presents a comprehensive computational study on the binding modes of caffeine bound to the adenosine A_{2A} receptor. Molecular docking was used to build candidate structures of the caffeine-A_{2A} complex. Then 5ns MD simulations were performed on the selected docking poses in an approximated physiological environment and 5 stable binding modes were found. The relative binding free energy of each binding mode was calculated using MM/PBSA method and compared to the binding free energy of the ZM241385-A_{2A} complex. Significant pocket residues were identified using MM/GBSA energy decomposition and compared to the results of previous site-directed

mutagenesis studies. This computational study brings important insight for the targeted drug design of the adenosine A_{2A} receptor.

Chapter 6 presents the pK_a calculation of Lys115 in Acetoacetate decarboxylase (AADase) with three different types of pK_a calculation methods: the molecular dynamics/thermodynamic integration (MD/TI) method, a Poisson-Boltzmann equation based method (MCCE), and an empirical method (PROPKA). Using the natural protonation states of other ionisable residues, none of the three methods could predict the correct protonation state of Lys115. But if Glu76 is protonated as indicated by previous site-directed mutagenesis studies, MD/TI predicts the pK_a of Lys115 to be 5.3, which predicts the right protonation state of Lys115 and agrees well with the experimental value of 5.9. In PROPKA, we modelled protonated Glu76 (neutral) by zeroing the charge-charge interaction between Lys115 and Glu76. The predictedpK_a of Lys115 is then 6.5, which is also close to the experimental value. The case study on Lys115 in AADase has shown that the right protonation pattern of other ionisable residues, especially the nearby ones, is crucial to the pK_a prediction of the target residue, which brings important insight to future pK_a predictions and development of new pK_a calculation methods.

7.2 Future Work

Finding the reaction path is very important for studying the mechanisms of gas phase reactions. The existing path-finding methods require a good initial guess to locate the

desired path, otherwise they need to explore a significant portion of the PES, which is very expensive computationally. Our work on developing new path-finding methods is only the start of the long journey. There are still lots of work to do on the two methods proposed in this thesis.

FMM is a very general and reliable method. Without prior knowledge of the PES, it can always find the global MEP. But it is an expensive method. Moving least square enhanced Shepard interpolation has been applied to reduce the computational cost. FMM has been successfully applied to analytical PES and small gas-phase chemical reactions. To apply FMM to larger systems, we can use the parallel FMM and compute many points on the surface at once. The current FMM method still has an exponential dependence on the dimensionality of the PES, however, so parallelizing the program will not make it possible to look at systems with very many reactive degrees of freedom like proteins with ten or more ionisable residues. At this point, FMM is restricted to small systems. Medium-sized systems could be accessed if a better interpolation method could be designed, so that fewer *ab initio* calculations were required. For complex biological systems with only a few reactive bonds (e.g., enzyme reactions), FMM can be interfaced with QM/MM program packages such as Sigma to explore the reaction path.

The QSM-NT method can find all stationary points on the PES, accordingly all alternative reaction paths. The pitfalls of this method include: 1) discontinuous Newton trajectories might impede locating all stationary points, and 2) multiple minima on the hyperplane might lead to the wrong path. The first problem is inherent to the character of

Newton trajectory. The workaround is to try more searching directions and to locate more Newton trajectories and their intersections. The second problem is associated with the path-finding algorithm (QSM). Using a growing string algorithm (GSM) can solve this problem. GSM is more expensive than QSM, so the next step of this program would be to design a method that automatically switches from QSM to GSM when the QSM calculations is failing to converge to the NT.

The computational study on the binding modes of caffeine bound to the adenosine A_{2A} receptor is a self-contained project. It reveals, however, the difficulty of computing the entropic contribution to the binding free energy. New computational methods for computing the entropic contribution to protein-ligand binding should be developed.

The pK_a calculations on Lys115 in AADase reveal that the site-site interactions from other ionisable residues play important roles on the pK_a shift of Lys115. Although our calculation results agree with the experimental measurements and results from previous site-directed mutagenesis studies, additional calculations on the pK_a values of close ionisable residues are required to confirm our hypothesis. In particular, it is important to calculate the pK_a of Glu76, Arg29, Glu61, Arg59 and Lys116 using MD/TI using an implicit solvent model. Understanding this complicated network of protonations can guide the design of new empirical models for pK_a prediction. Such models are particularly interesting because MD/TI is slow and tedious, and it is not a reasonable method for computing all the pK_a's in a protein.

Appendix:

List of Abbreviations

(1) Potential Energy Surface (PES)

(2) Quantum Mechanics (QM)

(3) Molecular Mechanics (MM)

(4) Quantum Mechanics/Molecular Mechanics (QM/MM)

(5) Intrinsic Reaction Coordinate (IRC)

(6) Steepest Descent Path (SDP)

(7) Minimum Energy Path (MEP)

(8) Fast Marching Method (FMM)

(9) Newton Trajectory (NT)

(10) String Method (SM)

(11) Quadratic String Method (QSM)

(12) Molecular Dynamics (MD)

(13) Molecular Mechanics/Molecular Dynamics (MM/MD)

(14) Molecular Mechanics/Poisson Boltzmann Surface Aread (MM/PBSA)

(15) Molecular Mechanics/General Born Surface Aread (MM/GBSA)

(16) AcetoAcetate Decarboxylase (AADase)

(17) Molecular Dynamics/Thermodynamic Integration (MD/TI)

(18) Multi-Conformation Continuum Electrostatic (MCCE)

(19) Nudged Elastic Band (NEB)

(20) Growing String Method (GSM)

(21) Transition State (TS)

(22) Eigenvector Following (EF)

(23) Gradient Extremal Following (GEF)

(24) Reduced Gradient Following (RGF)

(25) Scaled Hypersphere Search (SHS)

(26) G protein coupled receptor (GPCR)

(27) Free Energy Perturbation (FEP)

(28) Particle-Mesh Ewald (PME)

(29) Macroscopic Electrostatics with Atomic Details (MEAD)

(30) Thermodynamic Integration (TI)