PARAMETERIZATION OF DFTB3 AND APPLICATIONS IN STUDYING ENZYMATIC MECHANISM WITH QM/MM FREE ENERGY SIMULATIONS

by

Xiya Lu

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Chemistry)

at the

University of Wisconsin–Madison

2017

Date of final oral examination: 05/26/17

The dissertation is approved by the following members of the Final Oral Committee:

Qiang Cui, Professor, Chemistry
Jordan Ryan Schmidt, Associate Professor, Chemistry
Ned Sibert, Professor, Chemistry
Silvia Cavagnero, Professor, Chemistry
M. Thomas Record, Jr., Professor, Biochemistry

Parameterization of DFTB3 and Applications in Studying Enzymatic Mechanism with QM/MM Free Energy Simulations

Xiya Lu

Under the supervision of Professor Qiang Cui

At the University of Wisconsin–Madison

Abstract

Due to the higher computational cost relative to pure molecular mechanical (MM) simulations, hybrid quantum mechanical/molecular mechanical (QM/MM) free energy simulations particularly require a careful consideration of balancing computational cost and accuracy. For metalloenzymes that feature a dynamical active site, a direct sampling of the QM/MM potential function is important. Therefore, it is crucial to develop a low-level QM/MM potential that provides a semi-quantitative description of the underlying potential energy surface. In particular, we have been pursing an approximate density functional theory, density functional tight binding (DFTB), for biologically relevant systems. The latest version of the DFTB methodology, referred to as DFTB3, provides encouraging results for a fairly broad class of systems of biological interest. We report the parameterization of DFTB3, for sulfur, phosphorus, magnesium and zinc. The parameters are benchmarked against DFT, ab initio, and available experimental data. These parameters laid the ground work for using DFTB3/MM simulations with enhanced sampling techniques to study the chemical processes in enzyme, such as ATP hydrolysis in myosin. In normal myosin function, ATP hydrolysis is tightly coupled to the conformational state of the motor domain. The studies provide new details of the catalytic mechanism of ATP hydrolysis in the pre-powerstroke state and identify factors that regulate the hydrolysis activity in different conformational states.

To my parents and my husband for their unbounded love and support

Acknowledgments

I express my deepest gratitude to my Ph.D. advisor, Prof. Qiang Cui for the continuous support of my Ph.D study, for his motivation and immense knowledge. His curiosity and passion for science have constantly inspired me on my learning path. His advice on both research as well as on my career have been priceless. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. JR Schmidt, Prof. Ned Sibert, Prof. Silvia Cavagnero, and Prof. Tom Record for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

I thank my fellow labmates and friends for the stimulating discussions and all the fun we have had in the past years.

A special thanks to my family. Words cannot express how grateful I am to my mother-in law, father-in-law, my mother, and father for all of the support. At the end I would like express appreciation to my beloved husband for all of the sacrifices that youve made on my behalf.

Contents

Abstract i

1 Introduction **1**

1.1 Overview 4

2 Parameterization of DFTB3/3OB for Sulfur and Phosphorus for biological application **11**

2.1 Introduction 11

2.2 Theory 14

2.3 Parametrization of Sulfur and Phosphorus 16

2.3.1 Parameters for General Application 16

2.3.2 Special Parametrization for Phosphate Hydrolysis Reactions 19

2.4 Benchmarks and Discussion 20

2.4.1 Sulfur 21

2.4.1.1 Atomization Energies and Geometries 21

2.4.1.2 Vibrational Frequencies 22

2.4.1.3 Reaction Energies 23

2.4.1.4 Proton Affinities 25

2.4.1.5 Hydrogen Bonding Energies 26

2.4.1.6 Proton Transfer Barriers 28

2.4.1.7 Non-covalent interactions 31

2.4.2 Phosphorus 32

2.4.2.1 Atomization Energies and Geometries 32

2.4.2.2 Vibrational Frequencies 35

2.4.2.3 Reaction Energies . . . 37
2.4.2.4 Proton Affinities and Hydrogen Bonding Energies . . . 38
2.4.2.5 Hydrolysis reactions . . . 40
2.5 Conclusion . . . 43

3 Parameterization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications 55

3.1 Introduction . . . 55
3.2 Computational Methods . . . 57
3.2.1 Theory . . . 57
3.2.2 Reference Systems and Parameter Fitting . . . 58
3.2.3 Condensed Phase Benchmark Calculations . . . 60
3.2.3.1 Mg^{2+} and Zn^{2+} Ion Solvation and Relative pK_a Calculations . . . 62
3.2.3.2 Myosin-II Enzyme Setup: test for a Mg site . . . 63
3.2.3.3 Alkaline Phosphatase Enzyme setup: test for a bimetallic zinc site . . . 65
3.3 Results and Discussions . . . 66
3.3.1 Magnesium . . . 66
3.3.1.1 Geometries . . . 66
3.3.1.2 Sequential Bond Dissociation Energies . . . 67
3.3.1.3 Proton Affinities . . . 69
3.3.1.4 Large Test Set . . . 70
3.3.2 Zinc . . . 72
3.3.2.1 Geometries . . . 72
3.3.2.2 Sequential Bond Dissociation Energies . . . 74
3.3.2.3 Proton affinities . . . 74
3.3.2.4 Mixed Ligands . . . 77
3.3.2.5 Large Test Set . . . 79
3.3.3 Di-Nuclear Mg/Zn Active Sites . . . 80
3.3.4 Mg^{2+} and Zn^{2+} Ions Solvation Structure and Relative pK_a . . . 84

3.3.5 Structural Flexibility of the Active Site in Myosin . . . 89
3.3.6 Force Fields Comparison in AP Active Site . . . 89
3.4 Concluding Remarks . . . 91

4 Charging Free Energy Calculations Using the Generalized Solvent Boundary Potential (GSBP) and Periodic Boundary Condition 105

4.1 Introduction . . . 105
4.2 Computational Methods . . . 109
4.2.1 Charging free energy of ions in solution . . . 109
4.2.2 Oxidation free energy for a metal site in a protein . . . 112
4.2.3 Analysis of differences between PBC and GSBP simulations . . . 114
4.2.3.1 Analysis of electrostatic potential in a spherical water droplet . . . 115
4.2.3.2 Poisson-Boltzmann analysis for the smooth dielectric boundary in GSBP . . . 115
4.3 Results and Discussions . . . 116
4.3.1 Charging free energy of ions in a water droplet: correction terms for PME and GSBP simulations . . . 116
4.3.2 Charging free energy of Na^+ in a water droplet with additional charge: smooth boundary in GSBP simulations . . . 121
4.3.3 Oxidation free energy for the metal site in a fixed CueR protein . . . 123
4.3.4 Oxidation free energy for the metal site in a flexible protein . . . 127
4.4 Concluding Remarks . . . 131
4.5 Supporting Information . . . 133
4.5.1 Energy gap distribution . . . 133
4.5.2 Statistical for the Poisson-Boltzmann (PB) correction for the smooth dielectric boundary approximation in GSBP . . . 135

5 Recent Progress and Challenges in QM/MM Free Energy Simulations 142

5.1 Introduction . . . 142
5.2 Theory and Methods . . . 144
5.2.1 Reaction free energy profile and pathways . . . 144

5.2.1.1 Sampling along pre-determined order parameters . . . 144
5.2.1.2 Free energy path sampling . . . 146
5.2.2 Alchemical and multi-level free energy computations . . . 149
5.2.2.1 Alchemical free energy simulations . . . 149
5.2.2.2 Multi-level QM/MM free energy methods . . . 152
5.3 Model and realistic examples . . . 156
5.3.1 Reaction free energy profile and pathways . . . 156
5.3.1.1 Umbrella sampling and metadynamics . . . 156
5.3.1.2 String method in collective variables . . . 161
5.3.2 Alchemical and multi-level free energy computations . . . 164
5.3.2.1 Solvation free energy . . . 164
5.3.2.2 Dual-level QM/MM PMF . . . 166
5.4 Concluding Remarks . . . 172

6 Regulation and Plasticity of Catalysis in Enzymes: Insights from Analysis of Mechanochemical Coupling in Myosin 184

6.1 Introduction . . . 184
6.2 Computational Methods . . . 189
6.2.1 Basic QM/MM setup . . . 189
6.2.2 A Closed PR Structural Model . . . 190
6.2.3 Metadynamics . . . 191
6.2.4 String Method . . . 192
6.2.5 Average Electrostatic Potential Analysis . . . 193
6.3 Results . . . 193
6.3.1 The Reactant and Product Structures in the Pre-powerstroke State . . . 193
6.3.2 Reaction Pathways in the Pre-powerstroke State . . . 194
6.3.3 Higher free energy pathways in the pre-powerstroke state . . . 199
6.3.4 ATP Hydrolysis in the Post-rigor State . . . 200
6.3.5 ATP hydrolysis in the closed post-rigor model . . . 202

6.4 Discussion . . . 206
6.4.1 Mechanism of ATP hydrolysis in the pre-powerstroke state of myosin . . . 206
6.4.2 Factors that regulate ATPase activity in myosin . . . 208
6.4.3 Broader Implications: importance of positioning, electrostatics and plasticity in catalysis 210
6.5 Concluding Remarks . . . 212
6.6 Supporting Information . . . 213
6.6.1 Collective Variables in the Metadynamics Simulations . . . 213
6.6.2 The Projection Problem in Metadynamics Simulations for Myosin . . . 213
6.6.3 Collective Variables in the String Calculations . . . 215
6.6.4 Comparison between Metadynamics and String Calculations for the W2-E459 Pathway in the Pre-powerstroke State . . . 217
6.7 Electrostatic Polarization due to Glu459 is Captured by a Simple Model . . . 218
6.7.1 Calibration of DFTB3 with B3LYP calculations . . . 220
6.7.2 Exothermicity of Hydrolysis . . . 221
6.7.3 Transition State Energetics in the Enzyme . . . 222
6.7.4 Comparison of active site structure and fluctuations between the pre-powerstroke state and the closed post-rigor model . . . 224
6.7.5 RMSD between reaction pathways in the pre-powerstroke state and the closed post-rigor model . . . 225
6.7.6 W2 Pathway in the Pre-powerstroke State . . . 227
6.7.7 W2-S181 Pathway in the Pre-powerstroke State . . . 228
6.7.8 S236 pathway in the Pre-powerstroke State . . . 229
6.7.9 S236-W3-S181 pathway in the Pre-powerstroke State . . . 230
6.7.10 S236 pathway in the Post-rigor State . . . 231
6.7.11 W2-E459 pathway in the ”closed” Post-rigor State . . . 232

7 Concluding Remarks **241**

APPENDICES **245**

A List of Publications **246**

There is no number in this footnote

List of Tables

2.1 Overview of Electronic Parameters (in Atomic Units if not Unitless)[a] 18
2.2 Parameters Defining the Repulsive Potentials [a] 19
2.3 Mean and Maximum Absolute Deviations from G3B3 Atomization Energies and Heats of Formations and B3LYP/cc-pVTZ geometric properties for Our Sulfur Test Set 22
2.4 Selected Vibrational Frequencies in cm^{-1} 24
2.5 Deviation for Reaction Energies of Neutral Closed Shell Sulfur Containing Molecules Compared to G3B3[a] 25
2.6 Proton Affinities for Sulfur Containing Species in kcal/mol: Deviations in Comparison to G3B3[a] 27
2.7 Hydrogen Bonding Energies in kcal/mol: Deviations in Comparison to G3B3 28
2.8 Heavy Atom Distances in Å for Small Sulfur Containing Dimers: Deviations in Comparison to MP2-CP/G3large 29
2.9 Mean and Maximum Absolute Deviations for Small Neutral Closed-Shell Phosphorus Containing Molecules[a] 36
2.10 Mean and Maximum Absolute Deviations for 9 Closed-Shell Phosphate Anions in Comparison to B3LYP/aug-cc-pVTZ Geometries 36
2.11 Selected Vibrational Frequencies in cm^{-1} 37
2.12 Deviations for 16 Reaction Energies of Neutral Closed Shell Phosphorus Containing Molecules Compared to G3B3[a] 38
2.13 18 Proton Affinities for Phosphorus Containing Molecules in kcal/mol: Deviation of DFTB in Comparison to G3B3[a] 39
2.14 Hydrogen Bonding Energies for Phosphates in kcal/mol: Deviations in Comparison to G3B3[a] 39

2.15 Hydrogen Bonding Distances in Å for Small Phosphorus Containing Systems: Deviations in Comparison to B3LYP/aug-cc-pVTZ 40

2.16 Deviations of Exothermicities and Barrier Heights in Comparison to MP2/G3Large Single Point Calculations at B3LYP/6-31+G(d,p) relaxed geometries for 37 Elementary Steps in the Hydrolysis of MMP and DMP[a] 42

3.1 Overview of Electronic Parameters (in Atomic Units if not Unitless) [a] 59

3.2 Parameters Defining the Repulsive Potentials[a] 61

3.3 Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Magnesium Test Set 67

3.4 Sequential Bond Dissociation Energies of Magnesium Containing Molecules Compared to G3B3[a] 68

3.5 Gas Phase Proton Affinities of Magnesium Containing Molecules Compared to G3B3[a] 69

3.6 Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Magnesium Large Test Set 71

3.7 Bond Dissociation Energies of Magnesium Containing Large Molecules Compared to G3B3[a] . 72

3.8 Gas Phase Proton Affinities of Magnesium Containing Large Molecules Compared to G3B3[a] . 73

3.9 Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Zinc Test Set 74

3.10 Sequential Bond Dissociation Energies of Zinc Containing Molecules Compared to B3LYP/aug-cc-pVTZ[a] 75

3.11 Gas Phase Proton Affinities of Zinc Containing Molecules Compared to B3LYP/aug-cc-pVTZ[a] 76

3.12 Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Zinc Mixed Ligand Test Set 77

3.13 Bond Dissociation Energies of Zinc Containing Mixed Ligand Molecules Compared to B3LYP/aug-cc-pVTZ[a] 78

3.14 Gas Phase Proton Affinities of Zinc Containing Mixed Ligand Molecules Compared to B3LYP/aug-cc-pVTZ[a] 78

3.15 Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Zinc Large Test Set 79
3.16 Bond Dissociation Energies of Zinc Containing Large Molecules Compared to B3LYP/aug-cc-pVTZ[a] 80
3.17 Gas Phase Proton Affinities of Zinc Containing Large Molecules Compared to B3LYP/aug-cc-pVTZ[a] 81
3.18 Calculated activation barrier (in kcal/mol) for diester hydrolysis in the active site model of AP in the gas phase[a] 86

4.1 Uncorrected charging free energy (in kcal·mol^{-1}) calculated using PBC and GSBP boundary conditions[a] 117
4.2 The average potential (in kcal·mol^{-1}·e^{-1}) in spherical water droplets and the derived correction terms for PME/GSBP charging free energy simulations 118
4.3 Relative charging free energy for cations (in kcal·mol^{-1}) calculated using PBC and GSBP boundary conditions 121
4.4 Calculated charging free energy (in kcal·mol^{-1}) of Na^+ with different net charges in the systems using PBC and GSBP boundary conditions, and Poisson-Boltzmann correction for the GSBP results 122
4.5 Oxidation free energy (in kcal·mol^{-1}) for a metal site M in a rigid CueR protein calculated using PBC and GSBP boundary conditions 125
4.6 Oxidation free energy (in kcal·mol^{-1}) of a metal site M in a rigid CueR protein calculated with different net charges using PBC and GSBP boundary conditions, and Poisson-Boltzmann correction for the GSBP results 126
4.7 Oxidation free energy (in kcal·mol^{-1}) of a metal site M in CueR using PBC and GSBP with different parts of protein fixed 129
4.8 Standard deviation of Poisson-Boltzmann (PB) correction for the energy gaps from different λ windows of GSBP (20 Å) simulations and for the charging/oxidation free energy 136

5.1 Relative Solvation Free Energy $\Delta\Delta G_{gas\rightarrow aq}$(MM→QM) (in kcal/mol) with Different Size of QM Region[a] 165

5.2 Free energy correction and variance (in parentheses) of multi-level free energy estimators[a] . . 169

6.1 The average electrostatic potential difference ($\Delta\phi$ in kcal·mol^{-1}·e^{-1}) between sites of interest due to the protein environment (excluding QM atoms and water molecules) in the reactant and transition state configurations for the pre-powerstroke and closed post-rigor structures 205

6.2 Identity of proton donor, acceptor and potential proton shuttling group used in the definition of ζ in the metadynamics simulations of different reaction pathways. 213

6.3 Definition of the collective variables in the string method 216

6.4 Energy difference between the product and the reactant in the model-ADP system obtained at DFTB3 and B3LYP/aug-cc-pvtz level in the unit of kcal/mol 221

6.5 Potential Energy Barriers from DFTB3/MM and B3LYP/6-31+G(d,p)/MM calculations for different reaction mechanisms in the pre-powerstroke state 223

List of Figures

2.1 Problematic geometries for DFTB3/3OB with significant difference in angles or connectivity (i.e., the position of the hydrogen involved in hydrogen bonding) in comparison to counterpoise corrected MP2/G3large. 28

2.2 Proton transfer barriers. The energy is given relative to the energy of (SH^- + XH) for the upper row, and relative to (SH_2 + XH_2^+) for the lower row, where X=SH,OH,NH_2. The vertical axes shows the distance between sulfur and the shared hydrogen atom (rSH). The color code is: black=MP2/G3large, green=B3LYP/6-31+G(d,p), blue=PBE/6-31+G(d,p), red=DFTB3/3OB, light red=DFTB3/3OB/H–N-mod, the heavy atom distance is 3.7, 3.6, 3.8 Å for the upper row and 3.6, 3.2, 3.6 Å for the lower row. 30

2.3 Potential energy curves for the proton transfer between a hydronium ion and a sulfonic acid calculated at DFTB3/3OB (red) and B3LYP/aug-cc-pVTZ (green) levels, respectively. The energies are relative to infinitely separated reactants. 31

2.4 Potential energy curves for selected non-covalent interactions involving sulfur atoms studied in Ref.72; the format is similar to Fig. 2 of Ref.72. Although the artificially attractive interaction remains with DFTB3/3OB, the magnitude of interaction is significantly reduced compared to DFTB3/MIO. 33

2.5 Optimized structures for anti-parallel (AP) and parallel (P) thiophene dimers at different levels of theory. The S-S distance is given in Å. 34

3.1 Optimized structures of Mg·AT(D)P·water model molecules at different QM levels. Distance are given in angstroms. The numbers without parentheses were obtained at the B3LYP aug-cc-pVTZ level; those in the parentheses are DFTB3 values. The B3LYP optimized structures are depicted in colors whereas DFTB3 optimized structures are colored in blue. 71

3.2 Optimized structures of di-nuclear Mg^{2+}/Zn^{2+} complexes of simplified protein active site (a) PDB ID 1A49; (b) PDB ID 4ECQ; (c) PDB ID 3L4K; (d) PDB ID 1AMP. Distance are given in angstroms. The numbers without parentheses are obtained at the B3LYP 6-31+G(d,p) level; those with parentheses are DFTB3 values. 83

3.3 AP active site gas phase model with $MpNPP^{-}$, $MmNPP^{-}$ and MPP^{-}. Distance are given in angstroms. The numbers without parentheses are obtained at the B3LYP 6-31+G(d,p) level; those in the parentheses are DFTB3 values. (a) $MpNPP^{-}$ reactant state; (b) $MpNPP^{-}$ transition state; (c) $MmNPP^{-}$ reactant state; (d) $MmNPP^{-}$ transition state; (e) MPP^{-} reactant state; (f) MPP^{-} transition state; 85

3.4 Radial distribution function of water oxygen around Mg^{2+}/Zn^{2+} in 20 Å water droplet. 87

3.5 The free energy derivatives as the function of λ for (a) Mg^{2+}-water droplet simulations with the small QM region; (b) Zn^{2+}-water droplet simulations with the small QM region; (c) Mg^{2+}-water droplet simulations with the large QM region; (d) Zn^{2+}-water droplet simulations with the large QM region. 88

3.6 Structural properties of the Myosin active site during equilibrium QM/MM MD Simulations with different nucleotide chemical states (ATP or ADP·Pi). (a) Overlay of average DFTB3/MM structure and crystal structure (gray) of Myosin-II motor domain active site in the ATP state. The VO_4 moiety in the crystal structure (PDB code 1VOM) is replaced by a PO_3 phosphate group and a presumptive lytic water molecule. Instantaneous distances between Mg^{2+} and oxygen atoms in its four nonwater ligands at (b) ATP state; (c) ADP·Pi state. Ser237 refers to O^{Ser237} and Thr186 refers to O^{Thr186}. (d) PMF calculations for the ATP/ADP·Pi states. The reaction coordinate is the distance between Mg^{2+} and O^{Ser237}. 90

3.7 Structural properties of the AP active site during equilibrium QM/MM MD Simulations. A snapshot for the reactant state with (a) Coulomb MM force field (b) SLEF1 zinc MM force field (c) DFTB3/3OB (d) overlay of crystal and DFTB3/MM structure (blue), with average key distances in angstroms labeled. Asp369, His370 and His412 are omitted for clarity. Substrate inorganic phosphate in the crystal structure (3CMR) is not shown. 92

4.1 A scheme that illustrates the corrections necessary for GSBP and PBC (PME) simulations for electrostatic free energy simulations (modified based on Fig. 1 of Ref.(8)). The filled dot indicates the reference charge and unfilled dots for other charges; the single-headed arrows indicate polarization of the nearby solvent/protein dipoles. For PBC (PME) simulations, the corrections need to be considered include(36, 49) type B, which is due to the self energy and periodicity induced solvation terms, and type C, which is due to the use of an inappropriate particle (P) based summation scheme for Coulomb interactions; for systems studied here, only the type C correction (ΔG_{P-M} in Table 4.2) is found to be large and considered. For GSBP simulations, the corrections include those due to preferentially oriented water molecules (indicated by red arrows) at the inner/outer boundary (i.e., over-polarization, ΔG_{s-pol} in Table 4.2) and the smooth dielectric boundary approximation (corrected with Poisson-Boltzmann calculations, see Tables 4.4 and 4.6). 108

4.2 The CueR protein used as the model system for oxidation free energy calculations with different boundary conditions. (a) The protein is a homodimer and shown in the ribbon form. One of the metal sites is highlighted by a ellipse and the metal ion M along with its Cys ligands are shown in the ball-and-stick form. (b) A closer view of the metal site (color coding for the ribbon: red, acidic; blue, basic; green, polar; white, nonpolar). A bonded model is parameterized for the metal and its two Cys ligands. Three oxidation states of the metal site are studied, the reduced state ($\lambda = 0$, partial charge of M is +0.1158 e), the oxidized state ($\lambda = 1$, partial charge of M is +1.1158 e) and the doubly-oxidized state ($\lambda = 2$, partial charge of M is +2.1158 e). 113

4.3 The orientation distribution of water molecules in bulk (based on PBC calculations), an ODL ensemble and two other water droplet simulations; for the last three cases, orientations for water molecules within 1.5 Å from the droplet surface are shown. The orientation is defined as the angle between the water dipole and the normal vector of the local surface. 119

4.4 Radial distribution function of water oxygen around M in PBC and GSBP (20 Å) simulations of a rigid CueR protein. Different boundary conditions give reasonably similar water distributions. (a) M in the reduced state ($\lambda = 0$) with a van der Waals radius of 1.68 Å; (b) M in the oxidized state ($\lambda = 1$) with a van der Waals radius of 1.68 Å; (c) M in the reduced state ($\lambda = 0$) with a van der Waals radius of 0.68 Å; (d) M in the oxidized state ($\lambda = 1$) with a van der Waals radius of 0.68 Å. 124

4.5 Properties of PBC and GSBP simulations with a partially or fully flexible CueR. (a) Radial distribution function of water oxygen around M (oxidized state, $\lambda = 1$) in PBC simulations with different parts of the CueR protein held fixed. (b) A snapshot that illustrates the GSBP (20 Å) setup; protein atoms beyond 20 Å of M are fixed (cyan), those within 20 Å of M are flexible (yellow) and the bulk continuum is indicated with a gray background. (c) Overlay of +8Å-fixed and +8Å-relaxed structures for the oxidize state after 1 ns production run. The flexible protein region from the +8Å-fixed structure is shown in yellow, and that in the +8Å-relaxed structure is shown in purple; the fixed protein region is colored in blue. Clearly, the metal binding loop responds to the metal oxidation and therefore should be allowed to relax for reliable oxidation free energy calculations (see Table 4.7 and text for discussion). 128

4.6 Radial distribution function of water oxygen around M in PBC (flexible or 20Å-fixed) and GSBP (20 Å) simulations. (a) M in the oxidized state ($\lambda = 1$). (b) M in the doubly-oxidized state ($\lambda = 2$). Different boundary conditions lead to reasonably similar water distributions. . . 130

4.7 Energy gap distributions for $\lambda = 0$, $\lambda = 1$ and $\lambda = 2$ windows in (a) fully flexible PBC, (b) GSBP (20 Å), and (c) rigid protein PBC simulations. For the $\lambda = 1$ windows, the distributions shown correspond to the energy gap between oxidized and reduced states; the distributions for the energy gap between doubly-oxidized and oxidized states also closely follow Gaussian statistics and are not shown here. 134

4.8 (a) Linear fits of the free energy derivatives as functions of λ for the rigid protein and fully flexible protein PBC simulations. (b) Time-dependence of free energy derivatives for different λ windows for the fully flexible protein PBC simulation. 135

4.9 The distribution of PB corrections for the energy gaps from two representative windows of GSBP(20 Å) simulations. (a) charging free energy of Na^+ in the presence of five Cl^-, (b) rigid protein simulations. For the partially flexible protein simulations, the distributions for the PB corrections are similar and therefore not shown. 136

5.1 Thermodynamic cycles used to compute the solvation free energy of a QM solute in (a) a pure MM environment and (b) when nearby solvent molecules are also described at the QM level. In both cases, a pure MM solvation free energy calculation (bottom horizontal process) is used as a reference; to obtain the desired QM solvation free energies, various vertical processes need to be studied in which the solute (colored red) or a micro droplet (colored yellow) is converted between a MM and a QM model either in the gas phase or in an MM solution environment (in blue). 150

5.2 Two different routes to compute the free energy change as the potential function is switched between a low level and a high level. (a-b) illustrate the routes for alchemical free energy calaculations and (c-d) illustrate the corresponding calculations along a path (e.g., an optimized string). In the NBB route, a low-level window (e.g., λ_1^L) is connected with a neighboring high-level (e.g., λ_0^H) by re-weighting the λ_0^L data. In the FEP(+BAR) route, free energy change due to a switch in the potential function is calculated only for the same (e.g., λ_0) window; different λ/s windows (e.g., λ_0/λ_1 or s_0/s_1) are connected with the low-level potential function using BAR. See text for the corresponding equations for the free energy changes (compare Eqs. 5.16 and 5.18). 153

5.3 Results for the proton transfer in a model channel (see Fig. 5.4) from 1D REUS simulation along the ζ coordinate. (a) Convergence of the PMF as a function of simulation time per replica; (b) Illustration of the exchange behavior of replicas during the REUS simulation, using replica 2 as an example. 158

5.4 2D metadynamics simulations for proton transfer in a model channel to illustrate the coupling between hydration level change and proton transfer. (a)-(b) Snapshots for the reactant and transition state regions, respectively. The "bulk" MM water molecules are shown in blue, and the QM water molecules are colored by atom type. The dipoles are shown in green. (c)-(d) Two dimensional PMF from well-tempered metadyanmics simulations using DFTB3/3OB(133) (barrier 9.8 kcal/mol) and DFTB3/3OBw(54) (barrier 11.5 kcal/mol), respectively. ζ describes the proton transfer and s describes the level of channel hydration (see text). 159

5.5 Study of a proton transfer process inspired by the D-channel in cytochrome c oxidase. (a) Representative structures along the optimized string with key distances (in Å) labeled (water molecules are excluded for clarity); (b) PMF for the first step of proton transfer (approximately corresponding to 1$\rightarrow$3 in panel a) studied by umbrella sampling; (c) PMFs for the first step of proton transfer studied by the string method with different collective variables and restraining force constants (indicated in the legend in kcal/mol/Å^2). 162

5.6 Potential of mean force (PMF) computed via standard umbrella sampling (US), and perturbation of the PMF following the NBB and FEP routes (see Fig. 5.2) for (a) intra-molecular proton transfer in malondiadehyde in solution (b) p-nitrophenyl phosphate (pNPP^{2-}) dissociation in solution. For the NBB and FEP calculations, the "low-level" method is taken to be DFTB3/MM for both reactions, while the "high-level" method is AM1/MM for (a) and B3LYP/MM for (b). For the NBB/FEP perturbations at the end-points, only two reactant windows and two transition state windows are used, as shown in Fig. 5.2, leading to "corrected" barrier heights indicated by dashed lines (also see text). For (a), $s_0 = -1.0$ Å, $s_1 = -0.9$ Å, $s_{n-1} = -0.1$ Å, s_n=0.0 Å; for (b), $s_0 = 1.6$ Å, $s_1 = 1.7$ Å, $s_{n-1} = 3.1$ Å, s_n=3.2 Å. 167

5.7 Corrected free energy barrier height via NBB/FEP using different numbers of data points per window. Convergence plots are shown for (a) intra-molecular proton transfer in malondiadehyde in solution (b) p-nitrophenyl phosphate (pNPP^{2-}) dissociation in solution. NBB/FEP at end points (i.e., reactant and transition state windows) only requires adequate overlaps between s_0-s_1 and s_{n-1}-s_n windows, thus the results are less dependent on the number of data points. 168

5.8 Distribution of energy difference between the low- and high-levels of theory, ΔU^{LH}, using trajectories from the two levels of theory; also plotted is the integrand of the FEP expression using the low-level trajectory (i.e., $exp(-\beta\Delta U^{LH})P_L(\Delta U^{LH})$). The ΔU^{LH} data are shifted with a different amount in (a-d) for plotting purposes; see Table 5.2 for absolute free energy perturbation results. The plots are for: (a) reactant in malondiadehyde proton transfer (b) transition state in malondiadehyde proton transfer (c) reactant in $pNPP^{2-}$ dissociation (d) transition state in $pNPP^{2-}$ dissociation. In (a)-(b), $P_{AM1}(\Delta U)$ has a similar shape from $P_{DFTB3}(\Delta U)\exp(-\beta\Delta U)$), thus FEP and NBB calculations yield reliable barrier corrections. In (c)-(d), the ΔU^{LH} distributions are remarkably broader and data at different levels have almost vanishing overlap, thus it seems challenging for FEP and NBB to obtain reliable barrier corrections, although the Gaussian nature of the ΔU^{LH} distributions (fit shown explicitly in green) suggests that a linear response model works well (see Table 5.2). 171

6.1 (a) Overview of the myosin motor. The main structural elements involved in the mechanochemical coupling between ATP hydrolysis and mechanical motion are shown for the PPS conformation in color ribbon; these are Switch I, Switch II and P-loop in the nucleotide binding site, and the Relay helix connected to the Converter; the rest of myosin is drawn in light gray; ATP is drawn as color spheres. The displacements of the converter and the Relay in the transition to the PR structure are indicated in dark gray. (b) Expanded view of the nucleotide pocket centered on the QM region. Local transition in the active site that largely involves the displacement of the SwII motif, leading to the break of the salt-bridge between Glu459 and Arg238 as well as displacement of Gly457 away from the γ-phosphate of ATP. Switch I and P-loop remain largely intact between the PR and PPS states. 186

6.2 The reactant and product states of ATP hydrolysis in myosin simulated by DFTB3/MM (a) Reactant ATP. (b) Product ADP·P_i. Ligands of Mg^{2+} are omitted for clarity. Key hydrogen bonds are highlighted with dotted lines. Note that these structures do not represent energy-optimized stationary points but rather selected structures in the trajectory. 194

6.3 Representative structures along the MFEP obtained from the W2-E459 mechanism. The structures are numbered according to the locations identified in Fig. 6.5. 195

6.4 Rearrangement of hydrogen-bond associated with W3 in the W2-E459 mechanism. Key hydrogen bonds are highlighted with dotted lines. 196

6.5 (a) The 1D free energy profile along the MFEP obtained from W2-E459 mechanism. Each dot corresponds to an image along the string. (b) Values of the important distances along the MFEP. 197

6.6 (a) Free energy profile along the MFEP obtained from W2 mechanism. Each dot corresponds to an image along the string. (b) Proton transfer from W1 to W2. (c) Proton transfer from H_3O^+ to $HP_\gamma O_4^{2-}$. Note that these structures do not represent energy-optimized stationary points but selected structures in the specific string image. 198

6.7 (a) Free energy profile along the MFEP obtained from the W2-S181 mechanism. Each dot corresponds to an image along the string. (b) Proton transfer from W1 to W2. (c) Concerted proton transfer from H_3O^+ to S181 and from S181 to $HP_\gamma O_4^{2-}$. Note that these structures do not represent energy-optimized stationary points but selected structures in the MD trajectories. 199

6.8 (a) The 1D free energy profile along the MFEP obtained from the S236 mechanism. Each dot corresponds to an image along the string. (b) Transition structure in which W1 transfers one of its protons to Ser236, which in turn transfers its hydroxyl proton to the γ-phosphate 200

6.9 (a) Free energy profile along the MFEP obtained from the S236-W3-S181 mechanism. (b) Transition structure in which proton transfers along the proton wire W1→Ser236→W3→S181→$P_\gamma O_3^-$ .201

6.10 (a) The 1D free energy profile along the MFEP obtained from the S236 mechanism in the post-rigor state. Each dot corresponds to an image along the string. (b) Transition structure in which proton transfers between W1 and S236, and S236 and γ-phosphate. 201

6.11 (a) Free energy profile for the W2-E459 mechanism in the closed post-rigor structural model. (b) Proton transfer from W1 to the carboxylate group in E459 via W2. (c) Proton transfer from -COOH in E459 to $H_\gamma PO_4^{2-}$ through W2. The structures are taken from MD simulation trajectories. 202

6.12 (a) Comparison of water distribution around carboxylate oxygen of Glu459 (the one closer to water W1) in the PPS and closed PR structures. (b) Free energy profile for water reorganization in the closed PR state. (c) Initial state of water orientation. (d) W1 and W2 reorient their dipoles before hydrolysis. 203

6.13 Superposition of average positions of several active site residues in the (a) reactant and (b) TS windows in the PPS and closed PR (yellow) simulations. The most visible difference is that Ser456 adopts different orientations in the reactant state window due to differences in the solvation environment. 207

6.14 The 2D free energy surface shown in (a)(b) is obtained from metadynamics for S236 pathway in the pre-powerstroke state. (a) The reactive trajectory is projected on the surface as the solid black line. The structures in the reactive trajectory are used as the initial guess in the following string calculation. (b) The converged string is illustrated on the surface as the solid black line. Each circle corresponds to an image along the string. The color scale denotes free energy in the unit of kcal/mol. (c) The representative pre-product structure in which hydroxyl group in the inorganic phosphate is about to rotate towards to ADP. (d) The desired transition state discovered in the MFEP in the string method. 215

6.15 The 2D free energy surface obtained from metadynamics for W2-E459 pathway in the pre-powerstroke state. The converged string is illustrated on the surface as the solid black line. Results obtained from metadynamics and string match each other well regarding to the mechanism and free energy barrier. Each circle corresponds to an image along the string. The color scale denotes free energy in the unit of kcal/mol. 218

6.16 Proton transfer in a model system consisting of acetate and water molecules to illustrate the impact of polarization effects. The acetate and water molecules are colored by atom type, and the MM point charge is colored in yellow. Atoms fixed are denoted by asterisk. (a) - (c) Snapshots for the reactant, transition state and product in the two-water model, respectively. (d) - (f) Snapshots for the reactant, transition state and product in the three-water model, respectively. 220

6.17 Model ADP hydrolysis reaction in the gas phase 221

6.18 Distances between key residues in the reactant state 224
6.19 Distances between key residues in the transition state 225
6.20 2D RMSD between the pre-powerstroke state images and the ”closed” post-rigor state images 226
6.21 Representative structures along the MFEP obtained from the W2 mechanism in the pre-powerstroke state. 227
6.22 Representative structures along the MFEP obtained from the W2-S181 mechanism in the pre-powerstroke state. 228
6.23 Representative structures along the MFEP obtained from the S236 mechanism in the pre-powerstroke state. 229
6.24 Representative structures along the MFEP obtained from the S236-W3-S181 mechanism in the pre-powerstroke state. 230
6.25 Representative structures along the MFEP obtained from the S236 mechanism in the post-rigor state. 231
6.26 Representative structures along the MFEP obtained from the W2-E459 mechanism in the “closed” post-rigor state. 232

Chapter 1

Introduction

Using theoretical and computational approaches to understand catalysis in solution and biomolecules has a long history that dates back to the beginning of the development of quantum chemistry(1). More systematic efforts began with the pioneering developments of molecular dynamics simulations of biomolecules(2) and hybrid quantum mechanical/molecular mechanical (QM/MM) methodologies(3), as recognized by the Nobel prize of Chemistry in 2013. After decades of methodology developments(4–14), QM/MM simulations have become a powerful tool for the analysis of enzyme catalysis(15–20) and can aide effectively in exciting applications such as drug design(21), enzyme/protein engineering(22–24) and systems biology(14). Despite much success, one may argue that the field of computational enzymology is still at its infancy(25, 26), in the sense that it remains challenging to make quantitative predictions on the catalytic proficiency of an arbitrary enzyme with an entirely generic computational approach. Therefore, much work is needed to push forward QM/MM based techniques toward the next level of maturity and robustness.

It is well appreciated that the key challenge is to strike the proper balance between potential energy accuracy and the degree of conformational sampling. This is particularly the case for QM/MM based free energy simulations, since the cost of QM calculations is substantially higher than that of classical force field based calculations. For example, metalloenzymes whose active site is particularly dynamical in nature, allosterically regulated, and/or features a significant degree of solvent accessibility; good examples include biomolecular motors(27, 28), DNA repair enzymes(29–31), ion transporters(32–34) and enzymes that feature a high degree of catalytic promiscuity(35–40). On one hand, the involvement of metal ions calls for the use of a reliable QM approach; on the other hand, the particular dynamical nature of these systems requires adequate sampling.

One possible approach is to adopt the QM/MM free energy path approach developed by Yang and co-workers(11), in which the fluctuations of the QM region and MM region are decoupled, and extensive sampling is only done for the MM degrees of freedom in the presence of frozen QM atoms (also see Ref. (41, 42)) This approach makes it possible to use ab initio QM/MM potentials, which are likely required for quantitative accuracy, especially for metalloenzymes. However, for systems that feature significant anharmonic motions for the reactive degrees of freedom, such as extensive rearrangement of hydrogen bonding networks during the reaction, decoupling the QM and MM fluctuations and treating the QM fluctuations with a harmonic model is unlikely an adequate approximation. For similar reason, full QM cluster models of metalloenzymes, although tremendously successful for cases that feature a relatively rigid active site(43, 44), are unlikely predictive for the metalloenzyme systems of interest to us.

We focus on problems in which a direct sampling of the QM/MM potential function is important. Therefore, it is crucial to develop a low-level QM/MM potential that provides a semiquantitative description of the underlying potential energy surface. One approach is to develop system-dependent semi-empirical QM/MM potentials, based on either minimal energy path(MEP)(45–47), force-matching(48) or para-dynamics(49). The alternative approach, which is what we follow, is to develop a general-purpose approximate QM method so that refitting is (ideally) not required when studying a new system. In particular, we have been pushing forward an approximate density functional theory, density functional tight binding (DFTB(50, 51)), for biologically relevant systems. The latest version of the DFTB methodology, referred to as DFTB3(52),provides encouraging results for a fairly broad class of systems of biological interest; the accuracy is often comparable to density functional theory with the generalized gradient approximation (DFT/GGA) and a double-zetaplus-polarization quality basis set, while the computational cost is similar to conventional semi-empirical methods(53) such as AM1 and PM3, making it routine to conduct nano-second simulations based on DFTB3/MM potentials. The DFTB3 method in the current form is most reliable for structural properties. For energetic properties, the results are less robust as compared to DFT/GGA(54, 55) but can often be improved to satisfying accuracy with single point energy calculations at high QM(/MM) level or multi-level QM/MM free energy methods, making DFTB3 a promising low-level approach in dual-level QM/MM free energy simulations. This is particularly notable for metalloenzymes(38, 39, 56) because other semi-empirical methods are often problematic for metal ions, even for structural properties.

One approach to explore reaction mechanism(s) is to search MEP(s), also known as the intrinsic reaction coordinate(57). For reactions in the liquid phase or biomolecules, which feature very rugged potential energy surfaces, a small number of MEPs are often insufficient for a quantitative description of the reaction(9, 45, 58), since the potential energy profiles can vary greatly among them and it's not trivial to evaluate the weights of these MEPs in terms of contribution to the reaction flux at a finite temperature(59). The alternative is to determine the path(s) on a free energy surface. Since QM/MM calculations are generally more expensive compared to pure MM calculations, it is particularly worthwhile to carefully consider the best free energy practices for QM/MM simulations. When a low-level of QM theory, such as a semi-empirical method, is adequate, both potential of mean force and alchemical free energy simulations can be readily carried out with several established techniques, such as umbrella sampling, metadynamics and free energy perturbation. For complex reactions, minimum free energy path (MFEP) methods have become more prevalent. Finally, it is productive to integrate metadynamics with finite temperature string type of calculations to explore the mechanism of complex reactions; metadynamics using well-chosen CVs can be very effective at identifying new reaction channels and therefore provide good initial guess for string calculations, which are local in nature and therefore rely on well-chosen initial pathways.

Besides, since semi-empirical QM methods have generally lower transferability compared to *ab initio* methods, it is important to carefully benchmark the method used for the kind of reaction(s) under investigation. Apart from the reliability of the QM method, factors like the size of the QM region, QM/MM partitioning, the scheme for computing QM/MM interactions, the amount of configurational sampling and the accuracy of the MM force field, all affect the quantitative accuracy of a QM/MM simulation. For processes like long-range PT in proteins or involving net charge change in the total charge, the issues reliable treatment of electrostatics stand out. While the effect of long-range electrostatics can be included, with a large degree of accuracy, by using an approach like the Generalized Solvent Boundary Potential (GSBP)(60, 61), which helps avoid explicit sampling of the parts of the protein away from the region of interest while taking into account the electrostatic effect of these atoms and of the bulk solvent using a Poisson-Boltzmann (PB) approach, it is important to carry out careful validations of approximate enzyme models.

1.1 Overview

In Chapter 2 and 3, we report the parameterization of the approximate density functional tight binding method, DFTB3, for sulfur, phosphorus, magnesium and zinc. The parameterization is done in a framework consistent with our previous 3OB set established for O, N, C and H, thus the resulting parameters can be used to describe a broad set of organic and biologically relevant molecules. The 3*d* orbitals are included in the parameterization and the electronic parameters are chosen to minimize errors in the atomization energies/reaction energies. The parameters are tested using a fairly diverse set of molecules of biological relevance, focusing on the geometries, reaction energies, proton affinities and hydrogen bonding interactions of these molecules. DFTB3/MM simulations for solution and two enzyme systems are carried out and compare to available experimental Xray data. The results of DFTB3/3OB are compared to those from DFT (B3LYP and PBE), *ab initio* (MP2, G3B3) and several popular semi-empirical methods (PM6 and PDDG).

Due to the higher computational cost relative to pure MM simulations, QM/MM free energy simulations particularly require a careful consideration of balancing computational cost and accuracy. Chapter 4 presents several recent developments in free energy methods most relevant to QM/MM simulations and discuss several topics motivated by these developments using simple but informative examples that involve processes in water. For chemical reactions, we highlight the value of invoking enhanced sampling technique (e.g., replica-exchange) in umbrella sampling calculations and the value of including collective environmental variables (e.g., hydration level) in metadynamics simulations; we also illustrate the sensitivity of string calculations, especially free energy along the path, to various parameters in the computation. Alchemical free energy simulations with a specific thermodynamic cycle are used to probe the effect of including the first solvation shell into the QM region when computing solvation free energies. For cases where high-level QM/MM potential functions are needed, we analyze the perturbative correction to low-level QM/MM free energy results.

To better understand the impact of various approximations in the Generalized Solvent Boundary Potential (GSBP), in Chapter 5, we compare charging free energies in both solution and protein systems calculated with GSBP and periodic boundary condition (PBC) simulations. For simple ions in solution, we find good agreements between GSBP and PBC charging free energies, once the relevant correction terms are taken into consideration. For GSBP, the smooth dielectric boundary approximation inherent to GSBP tends to make

significant errors when the inner region is featured with a high net charge. The error can be corrected with Poisson-Boltzmann calculations using snapshots from GSBP simulations in a straightforward and robust manner. Due to the more complex charge and solvent distributions, the magnitudes of corrections in protein simulations appear to be different from those derived for solution simulations, leading to uncertainty in directly comparing absolute charging free energies from PBC and GSBP simulations for protein systems. The relative charging/oxidation free energies, however, are robust.

Myosins are remarkable biomolecular machines that utilize the chemical energy of ATP hydrolysis to perform diverse biological functions, such as cellular transport, signaling and muscle contraction(62). In normal myosin function, ATP hydrolysis is tightly coupled to the conformational state of the motor domain. The postrigor state binds ATP tightly but is incapable of ATP hydrolysis. On the other hand, ATP is hydrolyzed efficiently in the pre-powerstroke state.

In Chapter 6, we focus on identifying the main features of the pre-powerstroke conformation that facilitate ATP hydrolysis. Several factors have been discussed in previous studies, which include: (1) orientation of the lytic water along the "in-line" attack configuration(63); (2) polarization of the γ phosphate via nearby polar/charged interactions(64–66) or affecting β-γ phosphate orientation(67); (3) length of water wire that bridges the general base and the γ phosphate(68); (4) stability of hydrogen bonding network in the active site, which may rely on the interaction network surrounding the active site and therefore structural state of the distal region(28).

The mechanism of ATP hydrolysis in the myosin motor domain is analyzed using a combination of QM/MM simulations and enhanced sampling techniques. The ATP hydrolysis activity is found to depend on the positioning of nearby water molecules to conduct an efficient in-line attack. A network of polar residues facilitates proton transfer and charge redistribution during hydrolysis. Residues beyond the immediate neighborhood of the active site contribute the structural support of the active site motifs. Comparison of the observed hydrolysis pathways and the corresponding free energy profiles provide new details of the catalytic mechanism of ATP hydrolysis in the pre-powerstroke state and identify factors that regulate the hydrolysis activity in different conformational states.

Chapter 7 provides the major conclusions of the work reported in this thesis. Each chapter is accompanied by its respective bibliography. A list of publications from the research work in this thesis is provided

in Appendix A.

References

[1] L. PAULING, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry, 3rd Ed.*, Cornell University Press, Ithaca, 1960.

[2] J. A. MCCAMMON, B. R. GELIN, and M. KARPLUS, *Nature* **267**, 585 (1977).

[3] A. WARSHEL and M. LEVITT, *J. Mol. Biol.* **103**, 227 (1976).

[4] M. J. FIELD, P. A. BASH, and M. KARPLUS, *J. Comput. Chem.* **11**, 700 (1990).

[5] K. B. LIPKOWITZ and D. B. BOYD, editors, *J. Gao, In Reviews in Computational Chemistry VII*, VCH: New York, 1995.

[6] T. VREVEN and K. MOROKUMA, *J. Comput. Chem.* **21**, 1419 (2000).

[7] R. A. FRIESNER and V. GUALLAR, *Annu. Rev. Phys. Chem.* **56**, 389 (2005).

[8] J. L. GAO, S. H. MA, D. T. MAJOR, K. NAM, J. Z. PU, and D. G. TRUHLAR, *Chem. Rev.* **106**, 3188 (2006).

[9] D. RICCARDI, P. SCHAEFER, Y. YANG, H. YU, N. GHOSH, X. PRAT-RESINA, P. KONIG, G. LI, D. XU, H. GUO, M. ELSTNER, and Q. CUI, *J. Phys. Chem. B* **110**, 6458 (2006).

[10] Y. ZHANG, *Theor. Chem. Acc.* **116**, 43 (2006).

[11] H. HU and W. T. YANG, *Annu. Rev. Phys. Chem.* **59**, 573 (2008).

[12] H. M. SENN and W. THIEL, *Angew. Chem. Int. Ed.* **48**, 1198 (2009).

[13] S. C. L. KAMERLIN, M. HARANCZYK, and A. WARSHEL, *J. Phys. Chem. B* **113**, 1253 (2009).

[14] E. BRUNK and U. ROTHLISBERGER, *Chem. Rev.* **115**, 6217 (2015).

[15] J. L. GAO and D. G. TRUHLAR, *Annu. Rev. Phys. Chem.* **53**, 467 (2002).

[16] G. MONARD and J. K. M. MERZ, *Acc. Chem. Res.* **32**, 904 (1999).

[17] S. SHAIK, S. COHEN, Y. WANG, H. CHEN, D. KUMAR, and W. THIEL, *Chem. Rev.* **110**, 949 (2010).

[18] M. GARCIA-VILOCA, J. GAO, M. KARPLUS, and D. G. TRUHLAR, *Science* **303**, 186 (2004).

[19] V. C. NASHINE, S. HAMMES-SCHIFFER, and S. J. BENKOVIC, *Curr. Opin. Chem. Biol.* **14**, 644 (2010).

[20] M. W. VAN DER KAMP and A. J. MULHOLLAND, *Biochem.* **52**, 2708 (2013).

[21] R. LONSDALE and A. J. MULHOLLAND, *Curr. Top. Med. Chem.* **14**, 1339 (2014).

[22] M. T. REETZ, M. PULS, J. D. CARBALLEIRA, A. VOGEL, K. E. JAEGER, T. EGGERT, W. THIEL, M. BOCOLA, and N. OTTE, *ChemBioChem* **8**, 106 (2007).

[23] M. P. FRUSCHICHEVA, M. J. L. MILLS, P. SCHOPF, M. K. SINGH, R. B. PRASAD, and A. WARSHEL, *Curr. Opin. Chem. Biol.* **21**, 56 (2014).

[24] J. WIETEK, J. S. WIEGERT, N. ADEISHVILI, F. SCHNEIDER, H. WATANABE, S. P. TSUNODA, A. VOGT, M. ELSTNER, T. G. OERTNER, and P. HEGEMANN, *Science* **344**, 409 (2014).

[25] D. A. KRAUT, K. S. CARROLL, and D. HERSCHLAG, *Annu. Rev. Biochem.* **72**, 517 (2003).

[26] J. G. ZALATAN and D. HERSCHLAG, *Nat. Chem. Biol.* **5**, 516 (2009).

[27] H. L. SWEENEY and A. HOUDUSSE, *Annu. Rev. Biophys.* **39**, 539 (2010).

[28] Y. YANG, H. YU, and Q. CUI, *J. Mol. Biol.* **381**, 1407 (2008).

[29] C. X. SONG and C. HE, *Trends in Biochem. Sci.* **38**, 480 (2013).

[30] C. YI, G. JIA, G. HOU, Q. DAI, G. ZHENG, X. JIAN, C. G. YANG, Q. CUI, and C. HE, *Nature* **468**, 330 (2010).

[31] C. S. NABEL and R. M. KOHLI, *Science* **333**, 1229 (2011).

[32] V. R. KAILA, M. V. VERKHOVSKY, and M. WIKSTRÖM, *Chem. Rev.* **110**, 7062 (2010).

[33] P. GOYAL, J. LU, S. YANG, M. R. GUNNER, and Q. CUI, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18886 (2013).

[34] P. GOYAL, S. YANG, and Q. CUI, *Chem. Sci.* **6**, 826 (2015).

[35] O. KHERSONSKY, C. ROODVELDT, and D. S. TAWFIK, *Curr. Opin. Chem. Biol.* **10**, 498 (2006).

[36] J. K. LASSILA, J. G. ZALATAN, and D. HERSCHLAG, *Annu. Rev. Biochem.* **80**, 669 (2011).

[37] A. C. BABTIE, S. BANDYOPADHYAY, L. F. OLGUIN, and F. HOLLFELDER, *Angew. Chem. Int. Ed.* **48**, 3692 (2009).

[38] G. H. HOU and Q. CUI, *J. Am. Chem. Soc.* **134**, 229 (2012).

[39] G. H. HOU and Q. CUI, *J. Am. Chem. Soc.* **135**, 10457 (2013).

[40] F. DUARTE, B. A. AMREIN, and S. C. L. KAMERLIN, *Phys. Chem. Chem. Phys.* **15**, 11160 (2013).

[41] T. KOSUGI and S. HAYASHI, *J. Chem. Theory Comput.* **8**, 322 (2012).

[42] T. H. ROD and U. RYDE, *J. Chem. Theory Comput.* **1**, 1240 (2005).

[43] P. E. M. SIEGBAHN and F. HIMO, *J. Biol. Inorg. Chem.* **14**, 643 (2009).

[44] M. R. A. BLOMBERG, T. BOROWSKI, F. HIMO, R. Z. LIAO, and P. E. M. SIEGBAHN, *Chem. Rev.* **114**, 3601 (2014).

[45] Q. CUI and M. KARPLUS, *J. Am. Chem. Soc.* **124**, 3093 (2002).

[46] K. NAM, Q. CUI, J. GAO, and D. M. YORK, *J. Chem. Theory Comput.* **3**, 486 (2007).

[47] Y. YANG, H. YU, D. YORK, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **4**, 2067 (2008).

[48] Y. ZHOU and J. PU, *J. Chem. Theory Comput.* **10**, 3038 (2014).

[49] N. V. PLOTNIKOV, S. C. L. KAMERLIN, and A. WARSHEL, *J. Phys. Chem. B* **115**, 7950 (2011).

[50] M. ELSTNER, D. POREZAG, G. JUNGNICKEL, J. ELSNER, M. HAUGK, T. FRAUENHEIM, S. SUHAI, and G. SEIFERT, *Phys. Rev. B* **58**, 7260 (1998).

[51] M. ELSTNER, T. FRAUENHEIM, and S. SUHAI, *THEOCHEM* **632**, 29 (2003).

[52] M. GAUS, Q. CUI, and M. ELSTNER, *J. Chem. Theory Comput.* **7**, 931 (2011).

[53] W. THIEL, *Adv. Chem. Phys.* **93**, 703 (1996).

[54] V. MLYNSKY, P. BANAS, J. SPONER, M. W. VAN DER KAMP, A. J. MULHOLLAND, and M. OTYEPKA, *J. Chem. Theory Comput.* **10**, 1608 (2014).

[55] M. GAUS, X. LU, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **10**, 1518 (2014).

[56] D. K. CHAKRAVORTY, B. WANG, C. W. LEE, D. P. GIEDROC, and K. M. MERZ, JR., *J. Am. Chem. Soc.* **134**, 3367 (2012).

[57] K. FUKUI, *Acc. Chem. Res.* **14**, 363 (1981).

[58] Y. K. ZHANG, J. KUA, and J. A. MCCAMMON, *J. Phys. Chem. B* **107**, 4459 (2003).

[59] D. J. WALES, *Energy Landscapes*, Cambridge University Press, 2003.

[60] W. IM, S. BERNÈCHE, and B. ROUX, *J. Chem. Phys.* **114**, 2924 (2001).

[61] P. SCHAEFER, D. RICCARDI, and Q. CUI, *J. Chem. Phys.* **123**, 014905 1 (2005).

[62] B. ALBERTS, J. BRAY, D ABD LEWIS, M. RAFF, K. ROBERTS, and J. D. WATSON, *Molecular Biology of the Cell*, Garland Publishing, Inc, 1994.

[63] W. W. CLELAND, , and A. C. HENGGE, *Chem. Rev.* **106**, 3252 (2006).

[64] F. A. KIANI and S. FISCHER, *Proc. Natl. Acad. Sci. USA* **111**, E2947 (2014).

[65] G. LI and Q. CUI, *J. Phys. Chem. B* **108**, 3342 (2004).

[66] S. SCHWARZL, J. SMITH, and S. FISCHER, *Biochemistry* **45**, 5830 (2006).

[67] T. RUDACK, F. XIA, J. SCHLITTER, C. KOETTING, and K. GERWERT, *Proc. Natl. Acad. Sci. USA* **109**, 15295 (2012).

[68] M. MCCULLAGH, M. G. SAUNDERS, and G. A. VOTH, *J. Am. Chem. Soc.* **136**, 13053 (2014).

Chapter 2

Parameterization of DFTB3/3OB for Sulfur and Phosphorus for biological application[1]

2.1 Introduction

Phosphorus and sulfur are richly featured in chemistry and biology(1). Sulfur is part of amino acids Cys and Met and therefore involved in redox sensing; sulfur is the third most abundant mineral element in the human body. Sulfur is also part of many important biological co-factors such as iron-sulfur clusters, coenzyme A and several vitamins. As another example, sulfonation states on heparan sulfate chains are known to govern crucial signaling pathways and molecular-recognition event(2). Sulfur is also involved in many chemical and materials applications. For example, sulfonic acids are used in many detergents; Nafion(3), a sulfonated tetrafluoroethylene based fluoropolymer-copolymer, is an important materials used for proton exchange membrane in fuel cells; sulfur-containing heterocycles are broadly used in the field of organic electronics(4). Phosphorus is essential in biology because it is part of phospholipids, nucleic acids and many vital small molecules such as various phosphates (e.g., ATP/GTP) as well as bone (hydroxyapatite). Phosphoryl transfer reaction, for example, arguably represents the most important chemical transformation in biology(1, 5–7). Perturbations in phosphoryl transfer enzymes are involved in many serious human diseases such as cancer(8, 9). Protein kinases and phosphatases are among the most important drug targets(10–15); there are ∼2,000 protein kinases

[1]Reprinted (adapted) with permission from "Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications", M. Gaus, X. Lu and Q. Cui, *J. Chem. Theory Comput.* **10**, 1518 (2014). Copyright 2014 American Chemical Society.

and $\sim$1,000 phosphatases in the human genome, and these enzymes are essential to key cellular processes such as the control of cell cycles and division. In chemical industry, phosphorous compounds are predominantly consumed as fertilisers while organophosphorus compounds are also used in detergents, pesticides and nerve agents.

To describe the rich (bio)chemistry that involve phosphorus and sulfur in complex condensed phase environments, computational studies in the framework of QM/MM methods are essential(16). Although *ab initio* based QM/MM simulations have become increasingly powerful thanks to developments in both computational hardware and theoretical algorithms(17–20), they remain computationally demanding and therefore not ideally suited when multiple reaction mechanisms need to be analyzed. This is particularly the case when sampling is critical, such as in the study of intrinsically flexible systems (e.g., signaling proteins)(21, 22) or enzymes that feature rather solvent-accessible active sites(23–25); adequate sampling is also important to processes that occur at the solid/liquid interface(26, 27). Therefore, despite the tremendous progress in *ab initio* QM and QM/MM methods, it remains meaningful to explore further development of semi-empirical type of QM methods(28–31).

In this context, the Self-Consistent Charge Density Functional Tight-Binding (SCC-DFTB) method proves to be a promising approach(32–34). It is an approximate Density Functional Theory (DFT), and is derived by expanding the DFT total energy functional up to second order around a reference charge density. The resulting perturbative series is further approximated by applying a minimal basis LCAO expansion of the Kohn-Sham orbitals, by using a monopole approximation for the charge density fluctuations in the second order terms, and by approximating the zero-th order terms, which resemble the DFT double counting contributions for the reference density, with a sum of atomic pair potentials (32, 35).

The resulting approximate total energy terms have to be parametrized, and two classes of parameters can be distinguished: (i) the electronic parameters, which determine the atomic minimal basis set and the atomic reference densities as well as the chemical hardness values of the involved atoms. The determination of these parameters is quite straightforward. (ii) The repulsive energy parameters, which are necessary to determine the atomic pair potentials modeling the zero-th order contributions in the density expansion. Although these terms can in principle be computed based on DFT calculations, to achieve good general accuracy and partially compensate for approximations made in the other terms, an empirical fit to larger test sets is necessary

and therefore their determination is usually more involved.

SCC-DFTB has been parametrized for organic molecules containing O, N, C and H, for which extensive tests have been performed (36–38), as well for molecules containing S (39), Zn (40–42), P (43) and a few other elements(44–48). The resulting parameter set is referred to as the 'MIO' set and can be downloaded from the DFTB website: 'www.dftb.org'.

To improve the description of hydrogen-bonded complexes and proton affinities, which are important to biological applications, DFTB has been extended to include third-order contributions and a modified second-order Coulomb scaling law(33, 49, 50), leading to the DFTB3 approach. In its first implementation, the 'MIO' parameters, originally derived for the second-order method, were also used for DFTB3 (43, 50, 51). Recently, two of us have re-parametrized DFTB3 for O, N, C and H, resulting in a parameter set called '3OB' (52), emphasizing the application to organic and biologically relevant molecules. The main goal was to improve heats of formation and reaction energies, as well as non-bonded interactions. As benchmark calculations indicated, DFTB3-3OB generally gives results comparable to the most successful semi-empirical methods such as the OMx methods(30).

In this work, motivated by the importance of P and S in (bio)chemistry, we extend the '3OB' parametrization to these elements. This is a worthwhile effort because our parameterization includes *d* orbitals, which have been known to be essential to the proper description of chemical species involving P and S, especially hypervalent compounds(53) (see, for example, Refs.(54, 55)). By contrast, many popular semi-empirical methods such as PM3(56) and PDDG(28, 57) do not include *d* orbitals (PM6(58) does include *d* orbitals). We test the new parameters for P and S with a large set of biologically relevant molecules, focussing in particular on geometries, reaction energies, proton affinities and hydrogen-bonding interactions. These results show that the 3OB set represents a notable overall improvement over the 'MIO' set(39, 43) parameters for P/S containing compounds for both structural and energetic properties. For example, the deficiency concerning the S$\cdots$N interaction as reported in Ref.38 is removed. Moreover, due to major improvement in the atomization energies, reaction energies are generally better described with the 3OB parameters than the 'MIO' parameters. For the majority S/P-containing compounds tested here, DFTB3/3OB also out-performs PM6 and PDDG, especially for structural properties, vibrational frequencies, hydrogen bonding energies and proton affinities. Nevertheless, several limitations remain, such as too short hydrogen bonding distances for complexes that

involve charged S/P species; problems also remain for the energetics of phosphoryl transfer reactions that involve changes in the coordination number of the phosphorus. Some of these limitations are likely due to the intrinsic deficiencies of the underlying functional (PBE(59)), while others might be alleviated via continuing development of the DFTB3 framework, such as including three-center terms and multipoles for the charge fluctuation.

2.2 Theory

The DFTB3-parametrization of sulfur and phosphorus follows the protocol we outlined previously(52, 60) and used to parameterize for C, H, N, and O(52). The theory of DFTB3 has been described in detail in Ref.51; for a recent review, see Ref. 61. Here we only briefly discuss the parameters that enter the DFTB3 total energy:

$$\begin{aligned} E^{\mathrm{DFTB3}} &= E^{\mathrm{H0}} + E^{\gamma} + E^{\Gamma} + E^{\mathrm{rep}} \\ &= \sum_{iab} \sum_{\mu \in a} \sum_{\nu \in b} n_i c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{ab} \Delta q_a \Delta q_b \gamma_{ab} + \frac{1}{3} \sum_{ab} \Delta q_a^2 \Delta q_b \Gamma_{ab} + \frac{1}{2} \sum_{ab} V^{\mathrm{rep}}_{ab}. \end{aligned} \tag{2.1}$$

To determine the Hamilton matrix elements $H^0_{\mu\nu}$, the atomic orbital basis functions η_μ and the atomic reference densities ρ^0_a have to be determined. This is done by solving the Kohn-Sham equations for the atom in an additional harmonic potential with a confining radius r^{wf} for the atomic basis set and a different radius r^{dens} for the density. Since the d-orbitals of S and P are unoccupied, we use another value for the confinement radius, denoted by r^{wfd}. The off-diagonal matrix elements $H^0_{\mu\nu}$ are then precomputed and tabulated using an exchange-correlation functional (PBE(59)), the atomic basis set consisting of Slater functions and the initial atomic density. The diagonal elements $H^0_{\mu\mu}$ are equal to the atomic eigenvalues ε_x (x = s, p, or d-orbital). Here, no confinement of the orbitals is used to assure the proper dissociation limit. In the case of sulfur and phosphorus, however, we make an exception to this standard procedure and optimize ε_{d} as discussed in the following section. The Slater functions are defined by l_{max} as the highest angular momentum taken into account, n_{max}, another parameter that determines the size of the basis set, and α_0, α_1, ..., α_4, the exponents of the Slater functions.

For the second-order term of the energy E^{γ}, the atomic Hubbard parameters U_a are needed, which are calculated from DFT as the first derivative of the highest occupied orbital with respect to its occupation

number. [2] For the third-order term E^{Γ}, the derivative of the Hubbard parameter with respect to charge, U_a^{d} has to be determined. For CHNO it is calculated as the second derivative of the highest occupied orbital with respect to its occupation number. For P and S, as discussed below, we find a significant advantage in optimizing it.

The repulsive energy E^{rep} is described as pair potentials. Their parametrization is done by fitting to a selected set of reference atomization energies, molecular geometries, vibrational frequencies, and barrier or reaction energies. A careful benchmark of a resulting fit is the most time consuming part of the procedure, despite the progress of partially automatized fitting procedures (60, 63).

Specifically for the description of atomic energies and atomization energies (E^{at}) a further parameter was introduced, the spin-polarization energy E^{spin}. Because the total energy of an atom is described in a spin-unpolarized manner in standard DFTB, the total energy is overestimated for open-shell atoms. Therefore the energy difference between a restricted and an unrestricted spin DFT calculation, called E^{spin}, is subtracted to yield the atomic energies at the DFT level. [3]

---

[2] U_a describes the electron interaction on-site of an atom and enters the γ-function, which interpolates between the on-site and the long-range electron interaction of atom pairs. As a consequence of the form of the γ-function in the interpolating region, which is well within the covalent and non-covalent bond distances (e.g., hydrogen bonds), the Hubbard parameter determines the size of the atom in an inverse relation. While this was found to be reasonable within one period of the system of elements it is not for interactions between atoms of different periods. The most significant difference to the inverse relation was found for hydrogen(33). Therefore, the original γ-function of DFTB2(32) is modified for DFTB3(50, 51) (called γ^{h} in Ref. 51) for all pairs that include at least one hydrogen atom. Essentially one additional parameter (called ζ) was introduced to damp the influence of the Hubbard parameter such that γ is closer to Coulomb behavior at short distances (see also Figure 3 of Ref. 51) and therefore accounts for the small covalent radius of hydrogen. While the 3OB parameter set includes elements of the first (H) and second period (CNO) we now extend it to third period elements S and P. Naturally the question arises as to whether one can improve performance by applying a similar modification to γ. Since the Hubbard parameter for S and P are larger than they should be to fulfill the inverse relation of second period elements we would have to damp the γ function such that γ approximates the Coulomb behavior at larger distances. We have tested this idea but found insignificant improvement over the standard γ and therefore decided not to introduce additional complexity to our DFTB model. Note that an alternative to the γ-function has been suggested that intrinsically contains such effects, however, it has not been applied and tested yet (62).

[3] Note that a spin-polarization corrected DFTB (sDFTB) has been proposed(64). In principle, sDFTB could also be used to calculate atomic energies; however, for calculation of atomization energies it is not the standard. sDFTB uses parameters calculated for states with partial spin in the vicinity of spin-unpolarized atoms (as the Taylor-series suggests) and therefore is somewhat less accurate for single atoms than E^{spin}. In contrast, for radical molecules sDFTB is the method of choice.

2.3 Parametrization of Sulfur and Phosphorus

2.3.1 Parameters for General Application

This section summarizes the choices of all parameters discussed in the previous section. We follow the search protocol established in Refs. 60 and 52 and divide the parameters into two parts, electronic and repulsive parameters.

The electronic parameters are given in Table 3.1. For the sake of completeness we also include parameters that are calculated from DFT (U, E^{spin}, $\varepsilon_{\mathrm{s/p}}$) or are in line with the standard choice of the basis set (n_{max}, $\alpha_i, i = 0-4$); those parameters are not subject to optimization. We note that although l-dependent Hubbard parameters have been considered(65) and being explored by us for improving the description of transition metals (Gaus and Cui, unpublished), the Hubbard parameter is taken to be l-independent for P and S. For all DFT calculated values the PBE exchange correlation functional (59) has been used. The following parameters have been optimized to improve the overall performance:

- l_{max}: While a minimal basis has been used for H (s-orbital only) and CNO ($2s$ and $2p$-orbitals only), for S and P, we include 3s, 3p and 3d-orbitals, the latter being used as polarization functions, which are required to have a higher angular momentum than the valence orbitals. As already found previously for the 'MIO' parameters for S and P, d-orbitals are necessary to properly describe the diverse chemical environments of these elements.

- r^{wf}: In DFTB, the atomic orbitals used as a basis set are confined using a harmonic confining potential characterized by this parameter. It essentially determines the spatial extension of the basis functions. Therefore, this parameter has influence on bonding properties as well as on non-covalent interactions (Pauli-repulsion). For the H_2S dimer, for example (also see discussions below), the intermolecular distance is found to be too short with our parameterization; while this could be improved by using a larger r^{wf}, but only at the cost of less accurate binding energy and angle of the hydrogen bond. For phosphorus we have not tested dimer structures due to their rare appearance in biological systems, tests of other molecules indicate only a very subtle influence on the general performance of bond angles, bond lengths and atomization energies. With a larger r^{wf}, the angles improve but for bond distances and E^{at} slightly larger errors are found. With a smaller r^{wf}, the opposite effects are observed, description of

angles worsen, while bond distances and E^{at} slightly improve.

- r^{wfd}: Generally this parameter has only minor influence on the results. A larger r^{wfd} tends to improve bond distances and worsens bond angles. Note that the quality of computed bond distances is to a large degree dependent on the repulsive potentials; however, for a higher degree of transferability to different chemical environments, it is also helpful to tune r^{wfd}.

- r^{dens}: The density compression has a large impact on the total electronic energy; however, the effect can be compensated by the repulsive potential without introducing substantial errors (as long as r^{dens} is chosen within a reasonable range, e.g., r^{dens} ¿ 3.0 a_0). Therefore, r^{dens} is chosen such that the repulsive potentials smoothly interpolate to their cut-offs. Note that for hydrogen a very small density compression was chosen(52), causing the H–S and H–P potential to be negative in the binding region while other potentials like O–S, O–P, N–S, N–P are still significantly repulsive. Thus, the choice of r^{dens} for S and P is a trade-off between too negative and too repulsive potentials.

- ε_{d}: While all eigenvalues are usually calculated from atomic DFT calculations, an exception is made for the eigenvalues of the d-orbitals since they are unoccupied in the ground state for atoms S and P and therefore a change of ε_{d} does not affect the atomic total energy. Optimizing ε_{d} allows control of the d-orbital involvement in molecular bonding situations. This becomes particularly important for balancing energetics and geometries for the different oxides of S and P species. Generally, the calculated values of 0.02140766 a.u. for S and 0.02019087 for P are too low, i.e., d-orbitals are heavily involved in all types of bonding situation. Higher values reduce the excessive d-orbital participation. Empirically we found that for C–S and C–P bonds a larger amount of d-orbital involvement decreased errors, while a rather small amount of d-orbitals seemed more appropriate for O–S and O–P bonding situations.

- U^{d}: The Hubbard derivative appears in the third-order term E^{Γ} (Eq. 3.1) and is usually calculated from DFT. In the case of S and P, the values are –0.0695 and –0.0701 a.u., respectively. For the mean absolute deviation (MAD) of 8 proton affinities (PAs) for S containing species, however, we find a drop from 7.6 to 2.4 kcal/mol after optimizing U^{d} for sulfur. Similarly for P species, the MAD for 17 PAs is reduced from 9.1 to only 2.9 kcal/mol by optimizing U^{d} for P. The change of the Hubbard derivative only marginally affects other properties for systems with small charge fluctuations; however, for the

Table 2.1: Overview of Electronic Parameters (in Atomic Units if not Unitless)[a]

Parameter	S	P
$l_{\max}$	2	2
$n_{\max}$	2	2
α_0	0.50	0.50
α_1	1.19	1.17
α_2	2.83	2.74
α_3	6.73	6.41
α_4	16.00	15.00
r^{wf}	3.8	3.6
r^{wfd}	4.4	4.4
r^{dens}	9.0	9.0
ε_{s}	–0.63000872	–0.51063909
ε_{p}	–0.25802653	–0.20276532
ε_{d}	0.32140766	0.52019087
E^{spin}	–0.03121074	–0.06868820
U	0.3288	0.2894
U^{d}	–0.11	–0.14

[a] As described in the text, U, E^{spin}, $\varepsilon_{s,p}$, n_{max}, α_i are in line with the standard choices of DFTB parameterization and not subject to optimizations. By contrast, r^{wf}, r^{wfd}, r^{dens}, ε_{d} and U^{d} are adjusted based on properties of molecules in the fitting set.

highly oxidized phosphorus species, larger impacts are observed. Atomization energies become somewhat less accurate, P–P bond lengths are overestimated and the P–O–P angle in Diphosphoric acid is overestimated to be 141° instead of 116° for B3LYOPhydP/cc-pVTZ and 121° for DFTB3 using the calculated Hubbard derivative. Due to the importance of accurate PAs for many biochemistry application, we choose to use the fitted U^{d} and list the values in Table 3.1.

Note that the spin-polarization constants that enter the sDFTB formalism have also been calculated.

For the repulsive potentials, first a fit for sulfur is carried out (including all electronic parameters) and subsequently phosphorus related parameters are optimized. Table 3.2 provides an overview of all reference systems and values that lead to the repulsive potentials related to S and P. For the solution of the linear equation system set up to determine the repulsive potentials, division points have to be selected. Further, additional equations can be chosen to better represent second derivatives. These parameters are specified in Table 3.2. The geometries are taken from the B3LYP/cc-pVTZ level of theory with the exception of $[CH_3–COO–PO_3]^{2-}$, for which we use B3LYP/aug-cc-pVTZ. Atomization energies (E^{at}) are calculated from G3B3(66–68), barrier geometries and energetics (E^{bar}) from MP2/G3large. Additional equations for the fitting process (see Ref. 52)

Table 2.2: Parameters Defining the Repulsive Potentials [a]

Molecules (w^{eeq}, w^{feq} if different than 1.0) and E^{at} (kcal/mol)

SH_2 (10.0, 1.0)	181.9	$(CH_3)_3C$–SH (0.0, 1.0)	—	HP=CH_2	390.3
N_2S	248.2	P_2	116.6	H_3PO_4	760.0
H_2SO_4	592.5	PH_3 (2.0, 1.0)	140.2	P_4O_6 (0.1, 1.0)	1056.2
S_2	84.7	H_2P–PH_2	366.3	P_4O_{10} (0.1, 1.0)	1595.9
CH_2S	325.5	N≡P (0.1, 0.1)	147.8	CH_3S–P(=O)$(OH)_2$ (0.0, 0.1)	—
CH_3SH	472.5	$P(NH_2)_3$ (0.1, 0.1)	788.4	H_3PS_4 (0.05, 0.05)	527.2
CH_3S–SCH_3	828.4	CH_3PH_2	536.4	$[CH_3COO$–$PO_3]^{2-}$ (0.0, 1.0)	—

Proton transfer reactions	r^{XX} (Å)	E^{bar} (kcal/mol)
$[H_2S$–H–$SH_2]^+$	3.4, 3.5, 3.6, 3.7	0.6, 1.9, 3.7, 6.1
$[HS$–H–$SH]^-$	3.5, 3.6, 3.7, 3.8	2.5, 4.6, 7.2, 10.2

Potential	Division points (a.u.)	Additional equations (a.u.)
C–S	2.8, 3.4, 4.0, 4.8	$V''(3.044) = 0.49$
H–S	2.4, 2.9, 3.9, 4.3, 4.5	$V''(2.542) = 0.27$
N–S	2.8, 3.2, 4.3, 5.5	$V''(3.007) = 0.55$
O–S	2.6, 3.5, 4.7, 6.0, 6.2	$V''(3.986) = 0.00$, $V''(5.858) = 0.05$
S–S	3.5, 4.2, 4.8, 5.4, 5.8	$V''(3.942) = 0.13$
C–P	2.6, 3.0, 3.4, 3.8, 4.2, 4.6, 5.0	$V''(3.532) = 0.14$, $V'''(2.999) = -1.50$
H–P	2.2, 2.4, 2.6, 3.2, 3.4, 3.6, 3.8, 4.0	$V''(2.685) = 0.20$, $V'''(2.685) = -0.64$
N–P	2.7, 3.5, 4.0, 4.5, 5.0	$V''(2.812) = 0.84$
O–P	2.6, 3.2, 4.6, 5.0, 5.4	$V''(2.782) = 0.77$, $V''(3.024) = 0.40$
P–P	3.4, 4.0, 4.8, 5.0, 5.2	$V''(3.581) = 0.40$
P–S	3.4, 3.8, 4.6, 5.6, 6.2, 6.7	$V''(3.659) = 0.27$, $V''(4.033) = 0.10$

[a] w^{eeq} and w^{feq} are the weighting factors for energy and force equations in 1/a.u.; reaction equations and additional equations are weighted with 1/a.u., for details see Ref. 52; r^{XX} is the heavy atom distance.

are prepared using a few selected vibrational frequencies shown below as determined from BLYP/cc-pVTZ (unscaled) using the harmonic approximation.

2.3.2 Special Parametrization for Phosphate Hydrolysis Reactions

With the general parameter set outlined above, considerable errors are observed for phosphate hydrolysis reactions. The details are discussed in the benchmark sections. The bottom line is that the current DFTB3/3OB model still cannot properly describe the energetics of reactions where the phosphorus coordination changes from four to five/three oxygen atoms. This suggests that the current DFTB3 methodology has limited transferability for complex phosphorus chemistry with the required accuracy in energetics for mechanistic studies. As a temporary solution, we suggest a specific reaction parametrization(69) of the O–P

repulsive potential referred to as 'OPhyd'. The need of this specific fit suggests that additional improvements in the DFTB3 methodology, such as the inclusion of three-center terms and better electrostatic models (e.g., multipoles for charge fluctuations(62, 70)), are still needed and will be systematically analyzed in the future.

For 'OPhyd', we carry out the same fit for phosphorus as above with the inclusion of two reaction energies (the weight is 10/(kcal/mol)). The first one is a reaction of water with a Dimethyl hydrogen phosphate, leading to a penta-coordinated intermediate; the second reaction is the dissociation of methanol from the previous intermediate (n_int1 $\rightarrow$ n_com2). As reference energies MP2/G3large values are calculated at B3LYP/6-31+G(d,p) geometries. This fit shifts the O–P repulsive potential by about –10 kcal/mol, i.e., overbinding for O–P bonds is introduced.

2.4 Benchmarks and Discussion

In this section we discuss benchmark results for our 3OB parametrization for S and P in combination with the 3OB parameters(52) for O, N, C and H[4]. Comparisons are made to MP2, DFT, popular semi-empirical methods (PDDG(28, 57) and PM6(58)) and the previous DFTB3/MIO parametrization (using the parameters as defined in Ref. 51: $U^{\mathrm{d}} = -0.23, -0.16, -0.13, -0.19, -0.14, -0.09$ a.u. for C, H, N, O, P, S and $\zeta = 4.2$). For PM6, it is worth noting that the choice of heat of formation for H_2 and proton is essential to the computed hydrogenation reaction energies and proton affinities, respectively; we take the value of 0.0 and -54.0 kcal/mol for H_2 and proton, respectively, as recommended by the MOPAC program (http://openmopac.net/manual/pm6_accuracy.html). We have not carried out any comparison to OMx methods despite their impressive performance for typical organic molecules(30) because we are not aware of any systematic parameterization of OMx for P and S. Unless noted otherwise, energetics are given for geometries that are optimized at the respective level of theory. To compare directly the potential energy surfaces at different levels of theory, zero-point energies (ZPEs) are not included; the exception is for heat of formation calculations, in which vibrational contributions are included. The ZPEs calculated with DFTB3/3OB agree very well (within 1 kcal/mol) with higher-level calculations. For the calculation of non-covalent interactions, although we recognize the importance of including dispersion interactions in general, they are not included for most test cases here because these cases include small molecules and are dominated by hydrogen-bonding interactions; we explore explicitly the effect of dispersion to address a recently discussed caveat of DFTB in

[4]Note that in connection with 3OB as well as 'MIO', the hydrogen molecule is evaluated using the respective H–H-mod potential (71).

describing non-covalent interactions that involve sulfur(72). All DFTB calculations are carried out using our in-house DFTB code; for G3B3, MP2, B3LYP, BLYP, PBE, PM6, and PDDG/PM3, the Gaussian 03 and 09 software packages(73, 74) are used.

2.4.1 Sulfur

2.4.1.1 Atomization Energies and Geometries

The first test set consists of atomization energies and geometrical properties of 38 small neutral closed-shell molecules including sulfides, oxides, acids, thoils, sulfite, sulfate, sulfones, sulfoxides etc.. Table 2.3 summarizes the results.

DFTB3/3OB shows a very small MAD for atomization energies and clearly improves over DFTB3/MIO (4.7 vs. 68.7 kcal/mol). [5] The largest deviations are found for SO_3 and thiocyanates. For DFTB3 we have also carried out calculations for heats of formation following the standard approximations for the heat capacity corrections(75) and using unscaled vibrational frequencies calculated with DFTB3. Encouragingly, DFTB3/3OB yields a small MAD of only 4.8 kcal/mol; it is 68.1 kcal/mol for DFTB3/MIO.

The poor performance of MP2 and DFT for atomization energies is in line with large errors for heats of formation as has been reported earlier, e.g., in Ref. 76. A fit of atomic contributions reported in the same reference significantly reduces this error but has not been carried out in this study. PM6 and PDDG/PM3 are parametrized for directly yielding heats of formation and implicitly include corrections for heat capacity contributions. Therefore the calculation of atomization energies with these methods is not straightforward. However, heats of formation have been reported to be in very good agreement with experiments for large test sets including very challenging cases such as cations, anions and transition structures leading to a MAD of 6.5 and 6.4 kcal/mol for PM6(77) and PDDG(57), respectively. Note that those numbers were based on different test sets. Nevertheless, for our test set we find similar results, the MADs of 5.4 and 5.7 kcal/mol (see Table 2.3) indicate very similar performance of PM6 and PDDG. This is quite remarkable as PDDG does not use *d*-orbitals on sulfur atoms while PM6 does.

With a few exceptions geometries are described very well for all semi-empirical methods and particu-

[5] MIO reveals a strong over-binding that leads to large errors. Note that the over-bindings of different bond pairs are balanced such that they cancel out for reaction energies(52). For S, however, additional drawbacks have been found especially for the S–O bond (38). We further discuss the consequences of the unbalanced over-binding within MIO for reaction energies below.

Table 2.3: Mean and Maximum Absolute Deviations from G3B3 Atomization Energies and Heats of Formations and B3LYP/cc-pVTZ geometric properties for Our Sulfur Test Set

property[a]	N [b]	MP2[c]	B3LYP[d]	PBE[d]	PM6	PDDG	MIO	3OB
E^{at} (kcal/mol)	38	8.3	16.8	20.8			68.7	4.7
$E^{\mathrm{at}}_{\mathrm{max}}$ (kcal/mol)		20.5	56.0	100.7			221.7	34.4
ΔH^0_{f} (kcal/mol)	38	9.6[e]	18.4	21.4	5.4	5.7	68.1	4.8
$\Delta H^0_{\mathrm{f,max}}$ (kcal/mol)		24.1[e]	57.6	100.0	24.0	18.9	218.6	34.0
					(1.5/12.5)[f]	(3.8/15.7)[f]	(1.1/12.0)[f]	(0.4/1.8)[f]
r (Å)	124	0.008	0.007	0.014	0.015	0.027	0.014	0.008
r_{max} (Å)		0.037	0.019	0.041	0.079	0.121	0.224	0.042
a (°)	103	0.5	0.3	0.5	2.5	2.7	1.8	1.2
a_{max} (°)		3.7	1.6	3.2	31.6	14.8	11.4	9.6
d (°)	12	1.2	1.9	3.3	8.7	11.5	12.0	4.0
d_{max} (°)		2.9	17.4	27.0	72.6	53.6	86.2	10.8

[a] Bond lengths r, bond angles a, and dihedral angles d are compared to B3LYP/cc-pVTZ calculations; max stands for maximum absolute deviation. [b] Number of comparisons. [c] Basis set is cc-pVTZ. [d] Basis set is 6-31G(d). [e] Diphenylsulfoxide, -sulfone, and -sulfide have been excluded due to excessive computation time. [f] Single point G3B3 ΔH^0_{f} computed at the structures optimized by semi-empirical methods; the value before/after the slash is the mean/maximum absolute deviation from G3B3, which uses B3LYP/6-31G(d) structures.

larly well for DFTB3/3OB; e.g., the MAD in bond distances are 0.008, 0.014, 0.027 and 0.015 Å respectively, for DFTB3/3OB, DFTB3/MIO, PDDG and PM6. The largest bond distance error for PDDG is found for the singlet state of C=S with a difference of 0.12 Å relative to B3LYP/cc-PVTZ. Other deviations are substantially smaller. DFTB3/MIO overestimates the N=S bond within N_2S significantly by 0.22 Å. For the dihedral angle OSOH of Sulfonic acid large deviations to B3LYP/cc-pVTZ are found for almost all methods compared. Interestingly, this is not the case for *ab initio* methods using a triple zeta basis; MP2 and B3LYP with the cc-pVTZ basis set deviate only about 2°.

To demonstrate the superb performance of the semiempirical methods for structural properties we have carried out single point heat of formation calculations with G3B3 at geometries of the respective level. The MADs of PM6, PDDG, DFTB3/MIO, and DFTB3/3OB are as low as 1.5, 3.8, 1.1, and 0.4 kcal/mol, the MAX are 12.5, 15.7, 12.0, and 1.8 kcal/mol. These results again highlight the quality of DFTB3/3OB structures.

2.4.1.2 Vibrational Frequencies

Benchmark for vibrational frequencies is a cumbersome task since comparisons require matching the character of vibrational modes, which is not straightforward especially when comparing to experimental data

for fairly large molecules. A rigorous comparison is out of the scope of this work. In Table 2.4, a few unscaled harmonic vibrational frequencies as calculated from semi-empirical methods are compared to those from DFT calculations. For cases where an obvious assignment of the mode to experimental ones is possible, we have also listed the experimental values. Note that selected stretching frequencies are also used to determine the additional equations during the fitting (mainly for repulsive potentials), although the contributions from vibrational frequencies are rather limited.

Table 2.4 shows that the DFT methods are generally in good agreement with experiments, although BLYP tends to slightly underestimate the experimental values. DFTB3 with both parameter sets MIO and 3OB also performs quite well and the deviations from experiments are similar to the ones for the DFT methods. For larger test sets, we expect a more consistent performance of BLYP and B3LYP than DFTB3(78, 79). As to other semi-empirical methods, PM6 reveals very satisfying results,with only S–O stretch frequencies consistently underestimated. By contrast, PDDG shows large errors of several hundreds of wavenumbers, specifically for S-H and S–O stretch frequencies (in H_2S, H_2SO_4 and Dimethylsulfone).

2.4.1.3 Reaction Energies

We mentioned earlier the over-binding tendency of DFTB3 with the MIO parameters(52). We have shown recently that for a set of reactions (with small stoichiometric coefficients) that include CHNO containing species, the over-binding of different bonds cancel out and lead to relatively small errors for reaction energies with exception for some species (e.g., N–O bond) (52). During the MIO-parametrization of sulfur, however, the parameters could not be adjusted in such a way that the over-binding approximately cancel out even for simple reactions. Table 2.5 shows the large deviations for DFTB3/MIO for a small set of selected reaction energies.

By contrast, 3OB is parametrized to give reliable atomization energies (see discussions above). For reactions with small stoichiometric coefficients, it is expected that small errors in atomization energies lead to small errors for reaction energies as well. This is generally observed in Table 2.5. Together with an analysis of the atomization energies there are a few species that show large errors and spoil the performance for the respective reactions; those are molecules SO (singlet), SO_3, and O_2 (singlet) with large atomization energy errors of +11.8, +34.4, and +11.9 kcal/mol, respectively. Reactions that do not include these outliers are in very good agreement with G3B3 and therefore demonstrates the strength of the 3OB parametrization protocol.

Table 2.4: Selected Vibrational Frequencies in cm^{-1}

molecule (point group)	Irrep	description	exp	BLYP[a]	B3LYP[a]	PM6	PDDG	MIO	3OB
H_2S (C_{2v})	A_1	bending	1183	1176	1208	1055	1090	1042	1043
	A_1	stretch-sym	2615	2601	2684	2689	1542	2728	2611
	B_2	stretch-asy	2626	2615	2698	2698	1588	2778	2665
S_2^1 ($D_{\infty h}$)	Σ_g	S–S-stretch		657	706	718	712	704	707
HSSH (C_2)	A	S–S-stretch	515	464	498	542	388	409	530
Dimethyldisulfide (C_2)	A	S–S-stretch	509	459	492	525	362	469	495
Thioformaldehyde (C_{2v})	A_1	C=S-stretch	1059	1040	1087	1069	1076	1128	1052
Methanethiol (C_s)	A'	C–S-stretch	708	654	696	749	755	735	767
	A'	H–S-stretch	2597	2588	2674	2706	1556	2676	2595
Dimethylsulfide (C_{2v})	A_1	CSC-bend	282	250	258	257	256	255	255
	A_1	C–S-stretch-sym	695	640	680	742	741	705	735
	B_2	C–S-stretch-asy	743	688	733	776	754	734	774
Thiophene (C_{2v})	A_1	ring-sym	608	594	615	559	628	609	621
	B_2	ring-asy	751	716	751	678	692	783	776
	A_1	ring-sym	872	803	838	773	798	851	845
	B_2	ring-asy	903	850	879	853	907	912	903
2,5-Dihydrothiophene (C_{2v})	A_1	ring-sym	506	491	512	470	524	499	520
	B_2	ring-asy	665	592	632	628	639	642	675
	A_1	ring-sym	716	666	706	753	747	724	750
	B_2	ring-asy	824	797	826	806	848	851	860
Tetrahydrothiophene (C_2)	A	ring-sym	472	457	475	442	495	472	485
	B	ring-asy	684	631	669	724	729	691	723
	A	ring-sym	678	639	677	743	739	703	727
H_2SO_4 (C_2)	A	S–O-stretch-sym	831	706	794	773	723	700	805
	B	S–O-stretch-asy	882	766	851	775	860	735	809
	A	S=O-stretch-sym	1136	1118	1196	1036	874	1166	1165
	B	S=O-stretch-asy	1452	1372	1445	1250	902	1397	1389
Dimethylsulfoxide (C_2)	A'	CSO-bend-sym		264	278	269	359	266	262
	A''	CSO-bend-asy		294	313	296	355	324	313
	A'	S=O-stretch	1102	1047	1097	1054	794	1162	1122
Dimethylsulfone (C_{2v})	A_1	S=O-stretch-sym	1121	1081	1145	1059	735	1131	1110
	B_1	S=O-stretch-asy	1258	1258	1330	1170	816	1331	1312
MAD ($\delta\nu/\nu_{exp}\%$)				5.6	2.6	6.6	15.1	5.1	3.8
MAX ($\delta\nu/\nu_{exp}\%$)				15.0	8.5	13.9	41.0	20.6	11.8

[a] Basis set is cc-pVTZ.

Table 2.5: Deviation for Reaction Energies of Neutral Closed Shell Sulfur Containing Molecules Compared to G3B3[a]

Reaction			G3B3	B3LYP[b]	PBE[c]	PM6[d]	PDDG[d]	MIO	3OB
$S_2 + 2\ H_2$	$\rightarrow$	$2\ H_2S$	–59.3	–3.6	–0.1	+4.9	+11.1	+33.0	+1.5
$HSSH + H_2$	$\rightarrow$	$2\ H_2S$	–15.6	–0.2	+2.5	+10.5	+7.9	+19.7	–3.7
$CH_3SH + H_2$	$\rightarrow$	$H_2S + CH_4$	–19.2	–1.9	+2.1	+8.6	+9.3	+7.4	+2.4
$CH_3SH + H_2O$	$\rightarrow$	$H_2S + CH_3OH$	10.5	–2.8	–5.1	–2.9	–0.3	+11.5	+1.0
$CH_3SH + NH_3$	$\rightarrow$	$H_2S + CH_3NH_2$	7.0	–1.1	–1.8	–4.6	+3.4	+8.7	+2.2
$H_2C{=}S + H_2$	$\rightarrow$	CH_3SH	–37.1	–1.8	–2.3	–3.9	+3.7	–0.1	–3.7
$H_2C{=}S + H_2O$	$\rightarrow$	$H_2S + H_2C{=}O$	1.2	–5.5	–10.1	–6.9	–5.6	+14.8	–8.0
$H_2C{=}S + NH_3$	$\rightarrow$	$H_2S + H_2C{=}NH$	2.0	–3.2	–3.0	–11.1	–1.2	+11.1	–3.1
$H_2C{=}S + 2\ H_2$	$\rightarrow$	$H_2S + CH_4$	–56.4	–3.5	–0.1	+4.8	+13.2	+7.4	–1.2
$H_2SO_4 + 4\ H_2$	$\rightarrow$	$H_2S + 4\ H_2O$	–77.6	–10.3	+19.5	+26.7	+44.3	+49.2	–3.3
$O{=}S(OH)_2 + 3\ H_2$	$\rightarrow$	$H_2S + 3\ H_2O$	–65.1	–0.4	+22.3	+27.2	+41.8	+38.0	–10.2
$HS(=O)_2OH + 3\ H_2$	$\rightarrow$	$H_2S + 3\ H_2O$	–71.9	–9.7	+12.8	+34.0	+36.7	+45.3	–5.0
$SO_2 + 3\ H_2$	$\rightarrow$	$H_2S + 2\ H_2O$	–60.6	–5.8	+14.2	+13.3	+11.9	+35.4	+0.6
$2\ SO_2$	$\rightarrow$	$S_2^1 + 2\ O_2^1$	243.7	–11.9	–22.0	+16.4	–49.5	+50.7	–22.3
$2\ SO^1 + O_2^1$	$\rightarrow$	$2\ SO_2$	–213.0	+6.4	+5.0	+11.4	+75.9	–24.2	+31.3
$2\ SO_2 + O_2^1$	$\rightarrow$	$2\ SO_3$	–74.8	+6.2	+12.8	–28.0	+15.5	–74.3	–52.6
$SO_3 + H_2O$	$\rightarrow$	H_2SO_4	–22.0	+2.3	+1.0	+2.0	–24.5	+20.1	+35.7
$SO_2 + H_2O$	$\rightarrow$	$O{=}S(OH)_2$	4.5	–5.5	–8.1	–13.9	–29.8	–2.6	+10.8
$2\ O{=}S(OH)_2 + O_2^1$	$\rightarrow$	$2\ H_2SO_4$	–127.8	+21.7	+30.9	+3.7	+26.1	–28.8	–2.8
$O{=}S(OH)_2$	$\rightarrow$	$HS(=O)_2OH$	6.8	+9.3	+9.5	–6.8	+5.1	–7.3	–5.2
MAD				5.6	9.3	12.1	20.8	24.5	10.3
MAX				21.7	30.9	34.0	75.9	74.3	52.6

[a] Energies are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] Basis set is cc-pVTZ. [c] Basis set is 6-31G(d) [d] Heats of formation for H_2 are calculated as -26 and -22 kcal/mol for PM6 and PDDG. To correct for this exceptional error this value is set to 0.0 kcal/mol. See main text.

By comparison, PDDG and PM6 are featured with larger MADs; PDDG has a MAD of 20.8 kcal/mol. In fact, even DFT calculations, especially PBE, may have large errors; the MAD for B3LYP/cc-PVTZ and PBE/6-31G(d) is 5.6 and 9.3 kcal/mol, respectively. [6]

2.4.1.4 Proton Affinities

One goal of the further development of DFTB2 towards DFTB3 was to improve the description of proton affinities, which are critical to the proper description of many biochemical reactions(16, 34, 81–83). This property is evaluated for sulfur containing species in Table 2.6.

[6] Because the reactions have been chosen somewhat arbitrarily, the overall MAD given in Table 2.5 has to be considered with care. As pointed out by Fishtik(80), one can in principle compile a set of chemical reactions that yield arbitrarily high or low overall errors for any kind of method by linearly combining the reactions. One can, however, transform the problem and assign species errors that are independent of a particular choice of linearlly independent reactions(80). Of course, for a meaningful analysis, one would still have to compile a set of reactions that include a representative set of species. Therefore the test presented here serves as an illustration of the over-binding problematic for DFTB3/MIO only.

A difference between the 3OB and MIO parameters for CHNO is the use of calculated (3OB) and fitted (MIO) Hubbard derivative parameters. With the calculated Hubbard derivative for sulfur, the DFTB3/3OB proton affinities errors are quite large with a MAD of 9.7 kcal/mol (3OB/calc in Table 2.6). Therefore we optimize U^{d} for sulfur and observe a significant reduction of errors for many species (the final 3OB-set). The overall MAD remains somewhat large (5.8 kcal/mol) due mainly to the significant errors for sulfurous acid and its anions; for the geometry of the anion $[O_2S–OH]^-$, the S–OH bond length is overestimated by about 0.14 Å in comparison to B3LYP/aug-cc-pVTZ. In general, nevertheless, the improvement of DFTB3/3OB over PM6 and PDDG, is notable. It is also worth noting that a fit of U^{d}(S) for DFTB3/MIO does not substantially improve over the use of the calculated U^{d}(S) (MIO/calc), with a large MAD of 11.0 kcal/mol. Thus, the transferability of parameters has improved with the new 3OB parameter set. The comparison to PBE/6-31+G(d,p) shows that this set of proton affinities is well described by a standard density functional using a medium sized basis set; note that diffuse functions in the basis set are important for properly describing the anionic species. Finally, we note that G3B3 single points at DFTB3/3OB optimized structures lead to small errors in the proton affinities (a MAD of merely 1.2 kcal/mol) compared to the original G3B3, again highlighting the good quality of DFTB3/3OB structures.

2.4.1.5 Hydrogen Bonding Energies

Hydrogen bonding energies are generally well described by DFTB3 as shown in Table 2.7 for neutral, protonated, and deprotonated dimers, except for two cases (see below); by comparison, PM6 and PDDG give less satisfactory results and sometimes very large errors. However, somewhat surprisingly, considerable errors are found for geometries with all semi-empirical methods, including DFTB3 (Table 2.8). Besides the tremendous underestimation of the heavy atom distance, angles are also different than the counterpoise corrected MP2/G3large (MP2-CP) reference as shown for the most problematic cases in Fig. 2.1. Note for example the different positions of the shared hydrogen between the monomers for DFTB3 and MP2-CP within $[HS–H–SH]^-$ and $[H_2S–H–SH_2]^+$. This leads to the largest hydrogen bonding energy errors within the test set. Another obvious deficiency is that the shared hydrogen of $[H_2O–H–SH_2]^+$ is covalently bound to sulfur rather than to oxygen. We will discuss this issue below in connection with the proton transfer barriers.

With an attempt to understand the significant errors in the hydrogen-bonding structures, we look further into the DFT results. Geometries calculated with B3LYP/6-31G(d) (which are used in G3B3) in comparison to

Table 2.6: Proton Affinities for Sulfur Containing Species in kcal/mol: Deviations in Comparison to G3B3[a]

system	G3B3	B3LYP[b]	PBE[c]	PM6[d]	PDDG	MIO/calc	MIO	3OB/calc	3OB	G3B3//3OB
H_3S^+	174.1	+0.3	–0.3	–13.7	+8.6	–5.4	–7.5	+1.2	+0.8	–1.5
H_2S	355.7	–1.1	–3.6	–16.4	–12.1	+0.4	–2.1	+6.9	+0.5	–0.6
SH^-	498.4	–1.1	–4.8	–3.3	+12.2	+18.1	+15.2	+24.6	–2.8	+0.1
$CH_3SH_2^+$	190.4	+0.7	–1.1	+8.9	–6.3	–7.5	–10.0	–1.4	–1.7	–1.0
CH_3SH	362.5	–0.7	–3.4	–18.5	–21.8	–2.4	–5.2	+4.7	+0.7	–0.3
H_2SO_4	317.6	+0.7	–2.0	–4.1	+6.1	+19.9	+15.8	+7.8	+5.1	–1.0
HSO_4^-	454.3	–0.1	–2.8	–9.3	–2.9	+24.3	+17.2	+6.3	+2.2	–0.1
$O{=}S(OH)_2$	330.3	+2.0	–1.9	+3.8	+7.3	+17.5	+13.7	+18.4	+17.5	+3.7
$O{=}S(OH)O^-$	471.5	+0.6	–2.1	+7.4	+7.3	+35.5	+29.1	+28.9	+27.7	+2.6
$HS(=O)_2OH$	318.2	+2.4	+0.9	+11.2	+16.8	+19.7	+15.7	+7.7	+5.5	+1.2
$H_3CS(=O)_2OH$	323.5	+2.9	+1.4	+6.1	+8.3	+14.1	+10.0	+4.0	+2.0	+0.8
$[H_3CS(=O)(OH)_2]^+$	186.3	+4.0	+2.3	+13.0	+14.3	+5.2	+2.1	+4.7	+3.2	-2.0
MAD		1.4	2.2	9.6	10.3	14.2	11.0	9.7	5.8	1.2
MAX		4.0	4.8	18.5	21.8	35.5	28.9	28.9	27.7	3.7

[a] The molecules are given in the protonated form. The proton affinity is computed with the potential energies at 0 K without any zero-point energy correction (exception are PM6 and PDDG which calculate reaction enthalpies at 298 K). For the DFTB method the deviation is given as the difference to the G3B3 method ($E^{\text{method}} - E^{\text{G3B3}}$). [b] Basis set is aug-cc-pVTZ. [c] Basis set is 6-31+G(d,p). [d] The energy of the proton in PM6 is in error by -54 kcal/mol which has been accounted for by adding this number to the original result. See discussion at http://openmopac.net/manual/pm6_accuracy.html.

the ones from MP2-CP differ also with respect to the bond lengths (Table 2.8); calculated hydrogen-bonding angles are quite similar to MP2-CP though. For $[HS–H–SH]^-$ and $[H_2S–H–SH_2]^+$, similar to DFTB3/3OB, the shared hydrogen is half way between the sulfur atoms. When going to larger basis sets (e.g., G3large or aug-cc-pVTZ) the structure of $[H_2S–H–SH_2]^+$ becomes similar to the MP2-CP result shown in Fig. 2.1; however, for $[HS–H–SH]^-$, the shared hydrogen remains half way between the sulfur atoms. For PBE the position of the hydrogen remains always half way between the heavy atoms for both systems independent of the basis set. Therefore, these observations seem to suggest that the significant errors in the DFTB3/3OB structures are largely intrinsic to the DFT functional (PBE) used, although the errors in DFTB3/3OB are larger than those at the PBE level. Interestingly, despite the structural differences the binding energies are comparable for G3B3 and MP2-CP, with a small MAD of 0.4 kcal/mol. Also, B3LYP and PBE with the 6-31+G(d,p) basis set show MADs of 1.4 and 3.7 kcal/mol, respectively, the latter being in fact larger than that for DFTB3/3OB.

The modified basis set in the 3OB-parametrization for CHNO leads to an improved O–O distance in the water dimer(52), which was too short with DFTB3/MIO. The larger wave-function compression radius for

Table 2.7: Hydrogen Bonding Energies in kcal/mol: Deviations in Comparison to G3B3

			G3B3	MP2-CP[a]	MP2[a]	B3LYP[b]	PBE[b]	PM6	PDDG	MIO	3OB
2 H_2S	→	$(H_2S)_2$	–1.7	+0.1	–0.1	+0.3	–0.7	0.4	0.2	–0.2	+0.1
SH^- + H_2S	→	HS-H—SH^-	–11.7	–1.3	–2.3	–3.1	–8.4	–6.3	–18.6	–9.9	–11.1
H_3S^+ + H_2S	→	$[H_2S\text{-}H\text{—}SH_2]^+$	–13.6	–0.4	–1.5	–4.3	–9.4	4.0	–8.1	–7.5	–9.3
H_2O + H_2S	→	HS-H—OH_2	–2.4	–0.1	–0.6	–0.9	–1.8	–3.1	0.5	–1.6	–0.9
H_2O + H_2S	→	HO-H—SH_2	–3.0	+0.4	–0.1	+0.0	–0.9		0.1	+1.0	+1.2
H_2O + SH^-	→	HO-H—SH^-	–14.8	+0.5	–0.3	–0.6	–2.4	0.0	1.6	+0.2	+0.3
H_3O^+ + H_2S	→	$[H_2O\text{-}H\text{—}SH_2]^+$	–23.5	+0.4	–0.8	–2.9	–6.7	–12.0	–25.4	+4.1	–2.0
NH_3 + H_2S	→	HS-H—NH_3	–3.1	–0.1	–0.7	–1.4	–3.0	–0.7	2.5	–2.1	–0.2
NH_3 + SH^-	→	H_2N-H—SH^-	–8.2	+0.3	–0.2	+0.1	–1.4	–3.2	–2.2	+1.6	+1.0
NH_4^+ + H_2S	→	$[H_3N\text{-}H\text{—}SH_2]^+$	–12.8	+0.2	–0.4	–0.8	–2.8	2.4	2.9	+0.7	+1.3
MAD				0.4	0.7	1.4	3.7	3.6	6.2	2.9	2.7
MAX				1.3	2.3	4.3	9.4	12.0	25.4	9.9	11.1

[a] Basis set is G3large. [b] Basis set is 6-31+G(d,p), without counterpoise correction.

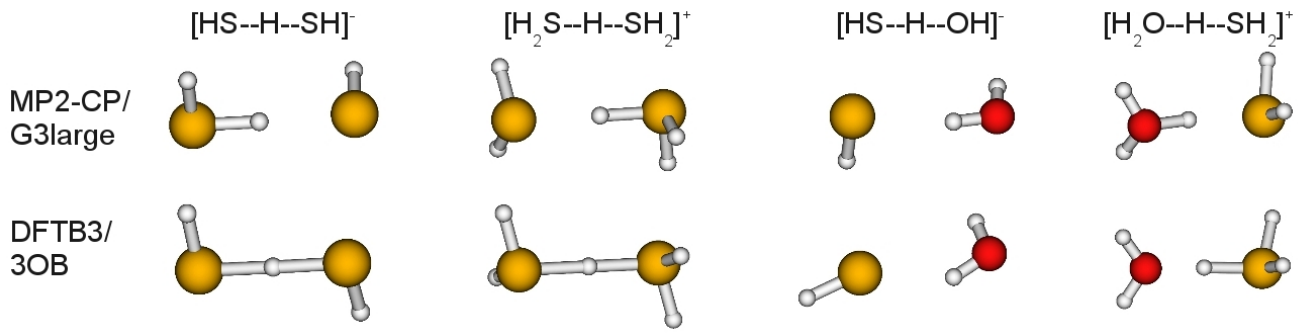


Figure 2.1: Problematic geometries for DFTB3/3OB with significant difference in angles or connectivity (i.e., the position of the hydrogen involved in hydrogen bonding) in comparison to counterpoise corrected MP2/G3large.

oxygen in 3OB increases the Pauli repulsion, such that the underestimation of the O–O distance is reduced. For a proper binding energy we then adjusted the parameter associated with the modified γ function for X-H pairs(33). In the case of sulfur, a larger r^{wf} leads also to larger intermolecular distances as expected, however at the cost of larger errors for bond angles, which obviously limits the optimization. The current set is a compromise within the DFTB3 framework that uses the PBE functional.

2.4.1.6 Proton Transfer Barriers

To study proton transfer in molecules including sulfur atoms, we consider several simple model systems: $[HS\text{–}H\text{–}SH]^-$, $[H_2S\text{–}H\text{–}SH_2]^+$, $[HS\text{–}H\text{–}OH]^-$, $[H_2S\text{–}H\text{–}OH_2]^+$, $[HS\text{–}H\text{–}NH_2]^-$, and $[H_2S\text{–}H\text{–}NH_3]^+$.

Table 2.8: Heavy Atom Distances in Å for Small Sulfur Containing Dimers: Deviations in Comparison to MP2-CP/G3large

Dimer	MP2-CP[a]	MP2[a]	B3LYP[b]	B3LYP[c]	PBE[c]	PM6	PDDG	MIO	3OB
$(H_2S)_2$	4.180	–0.066	+0.007	–0.004	–0.163	–0.098	–0.275	–0.372	–0.329
HS-H—SH^-	3.484	–0.056	–0.114	–0.050	–0.131	–0.255	–0.289	–0.301	–0.330
$[H_2S$-H—$SH_2]^+$	3.450	–0.044	–0.076	–0.084	–0.078	–0.234	–0.297	–0.205	–0.243
HS-H—OH_2	3.595	–0.069	–0.176	–0.103	–0.172	–0.866	–0.224	–0.394	–0.280
HO-H—SH_2	3.533	–0.054	+0.028	–0.036	–0.120		–0.043	–0.318	–0.085
HO-H—SH^-	3.245	–0.034	+0.007	–0.009	–0.063	–0.285	–0.060	–0.189	–0.117
$[H_2O$-H—$SH_2]^+$	2.929	–0.018	–0.007	–0.054	–0.047	–0.388	0.055	–0.126	+0.086
HS-H—NH_3	3.620	–0.057	–0.217	–0.129	–0.238	–0.223	0.253	–0.268	–0.248
H_2N-H—SH^-	3.533	–0.040	+0.007	+0.015	–0.060	–0.444	–0.379	–0.246	–0.227
$[H_3N$-H—$SH_2]^+$	3.290	–0.024	–0.007	–0.026	–0.081	–0.187	0.011	–0.178	–0.366
MAD		0.046	0.065	0.051	0.115	0.331	0.189	0.260	0.231
MAX		0.069	0.217	0.129	0.238	0.866	0.379	0.394	0.366

[a] Basis set is G3large. CP: counterpoise corrected calculation. [b] Basis set is 6-31G(d), this method is used within G3B3 for the geometry optimization. [c] Basis set is 6-31+G(d,p).

For comparison with higher level methods, we fix the heavy atoms at four different interatomic distances. The shared hydrogen atom is moved on a straight line from the first heavy atom to the second one and the total energy for each position is calculated after a relaxation of all other hydrogen atoms.

Fig. 2.2 shows the potential energy curves for all model systems for one fixed heavy atom distance. The energy is shown relative to the energy of two infinitely separated monomers, e.g., for $[HS–H–OH]^-$ the energy is shown relative to SH^- and H_2O. We compare DFTB3/3OB to MP2/G3large, B3LYP and PBE with the 6-31+G(d,p) basis set. While the relative energies vary by several kcal/mol between the methods due mainly to difference in the binding energies, the barriers are overall consistent. For DFTB3/3OB two major deficiencies are observed. First, the underestimation of the proton affinity of NH_3 leads to too low a relative energy for the system $[H_2S$-H—$NH_3]^+$. This drawback is already known and can be alleviated by applying a modified repulsive potential H–N-mod which improves the proton affinity (for details, see Ref.52). Second, for $[H_2S–H–OH_2]^+$, the total energy of the SH_2—H-OH_2^+ system is underestimated by almost 8 kcal/mol. The error does not stem from the differences in proton affinities for SH_2 and H_2O as their deviations from G3B3 are only +0.7 and +1.7 kcal/mol, respectively. However, the binding energy $H_2S + H_3O^+ \rightarrow H_2S–H_3O^+$ is underestimated by $\sim$8 kcal/mol, whereas the binding energy of $H_2O + SH_3^+ \rightarrow [SH_3^+–OH_2]$ is only slightly overestimated by 2 kcal/mol.

As a slightly larger example, we study a model proton transfer between a hydronium ion and sulfonic

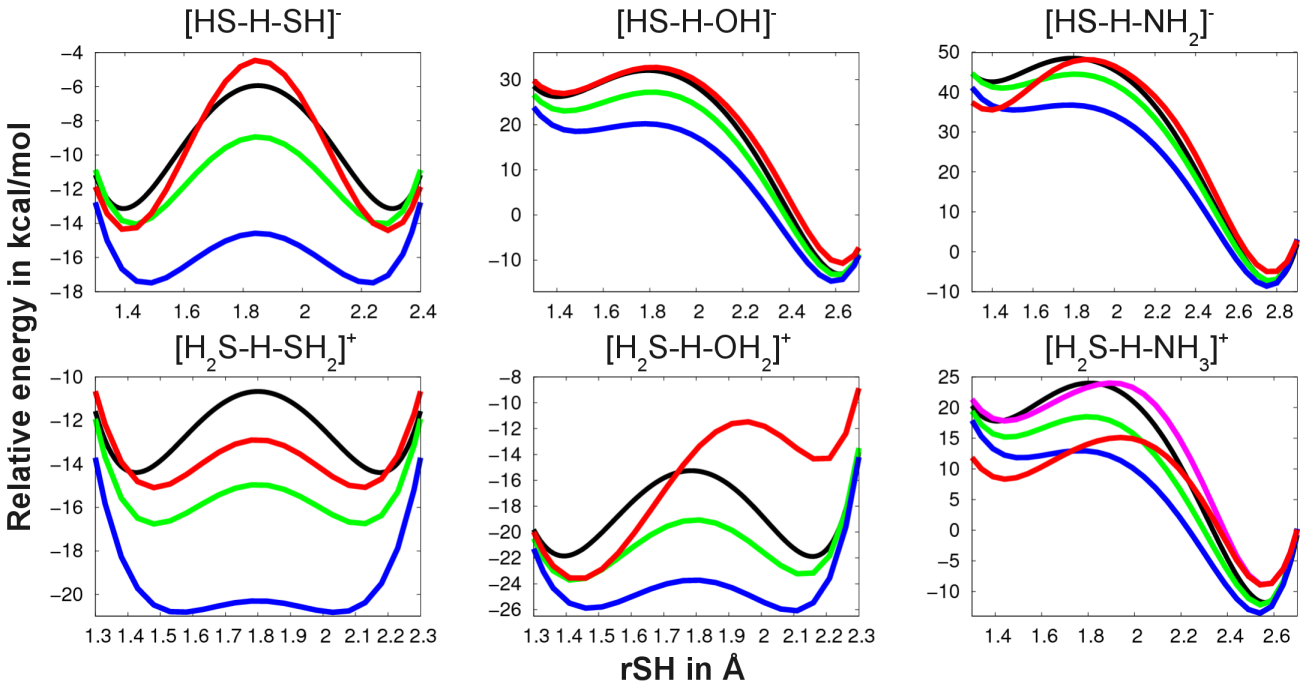


Figure 2.2: Proton transfer barriers. The energy is given relative to the energy of (SH^- + XH) for the upper row, and relative to (SH_2 + XH_2^+) for the lower row, where X=SH,OH,NH_2. The vertical axes shows the distance between sulfur and the shared hydrogen atom (rSH). The color code is: black=MP2/G3large, green=B3LYP/6-31+G(d,p), blue=PBE/6-31+G(d,p), red=DFTB3/3OB, light red=DFTB3/3OB/H–N-mod, the heavy atom distance is 3.7, 3.6, 3.8 Å for the upper row and 3.6, 3.2, 3.6 Å for the lower row.

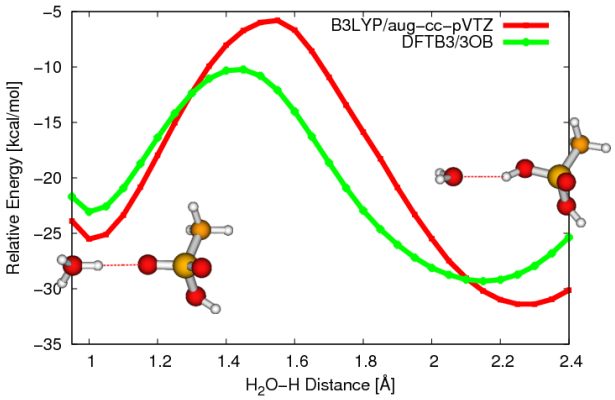


Figure 2.3: Potential energy curves for the proton transfer between a hydronium ion and a sulfonic acid calculated at DFTB3/3OB (red) and B3LYP/aug-cc-pVTZ (green) levels, respectively. The energies are relative to infinitely separated reactants.

acid. Proton transfers involving sulfonic acids are implicated in proton conducting membrane materials used in fuel cells(3). The hydronium oxygen is fixed at a set of distances from the sulfur in the sulfonic acid and the proton transfer is studied in a similar fashion as discussed above; the transferred proton is fixed for a set of distances along the S—O axis while the remaining atoms are relaxed using DFTB3/3OB. The result for the O-S distance of 4.5 Å is shown in Fig. 2.3 as an example. The DFTB3/3OB approach captures the exothermicity of the reactions rather well compared to B3LYP/aug-cc-pVTZ single point energies, while the barrier is substantially underestimated by almost 7 kcal/mol. The degree of underestimation is smaller at shorter O-O distances, but the result in Fig. 2.3 underlines the fact that DFTB3 is developed based on PBE and therefore may underestimate proton transfer barriers at large donor-acceptor distances. Therefore, in applications where rates or proton conductance is of interest, it is likely important to calibrate/correct the computed barrier height based on high-level calculations.

2.4.1.7 Non-covalent interactions

In a recent study(72), it was pointed out that DFTB3/MIO leads to artificially favorable interactions involving sulfur atoms thus spurious non-covalent binding for sulfur-containing compounds. Test calculations indicated that resolving the issue likely requires refitting the electronic parameters for sulfur. With the 3OB set of parameters, especially the new density compression radius (r^{dens}), we see that the artificial binding

discussed in Ref.72 is significantly reduced; rather than a few kcal/mol, now the over-binding is on the order of 0.2 kcal/mol for S-O/N interaction and 0.5 kcal/mol for S-S interaction (Fig. 2.4). We are unable to completely remove the attractive interactions within the framework of DFTB3 without affecting other properties discussed above, especially geometry and hydrogen-bonding interactions. For practical applications, however, we emphasize that when such weak over-binding effects need to be considered, explicit dispersion corrections is likely also important. In Fig. 2.5, we compare optimized structures for anti-parallel and parallel thiophene dimers, which were studied in Ref.72, at different levels of theory. Without explicit dispersion, both DFTB3/MIO and DFTB3/3OB lead to structures considerably different from B3LYP including dispersion (B3LYP-D3BJ(84)); the S-S distance is longer with DFTB3/3OB, reflecting the weaker spurious interaction between the sulfur atoms compared to DFTB3/MIO. When the empirical dispersion(85) is included, the DFTB3/3OB structures become in rather good agreement with B3LYP-D3BJ. The distance between the parallel thiophene monomers is underestimated; since the dispersion model used for DFTB3 here was developed in the framework of DFTB2(85), a more systematic refinement of the dispersion model together with DFTB3/3OB will likely further improve the result.

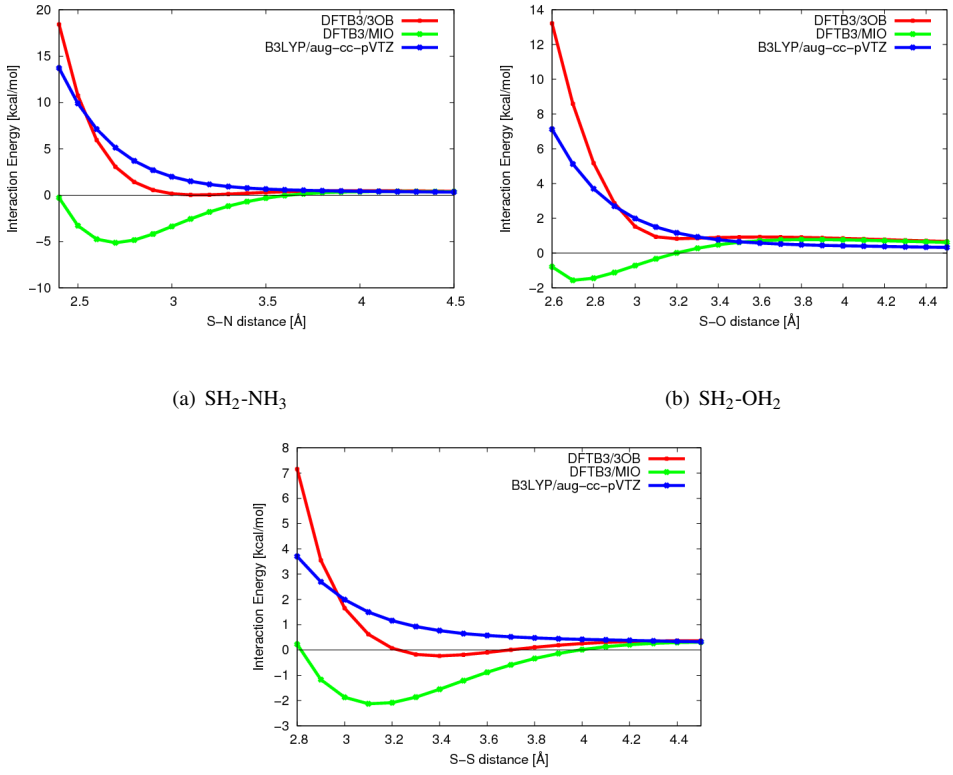
2.4.2 Phosphorus

2.4.2.1 Atomization Energies and Geometries

Similar to for sulfur we have compiled a test set of small neutral closed-shell phosphorus containing molecules. Among those are acids, oxides, phosphines, and thioesters with different oxidation states of phosphorus. Table 2.9 summarizes the results.

Even though the MAD is dependent on the particular choice of the test set, 3OB for phosphorus seems overall less accurate than for sulfur (7.3 vs. 4.7 kcal/mol). The largest deviations for atomization energies are found for the oxides P_4O_{10} and P_4O_6 with an error of –40.7 and –33.9 kcal/mol, respectively. Again, the 3OB set is a major improvement over the MIO set, which shows large errors (MAD of 82.9 kcal/mol) due to over-binding. Even MP2 and B3LYP, without fitting atomic contributions, show large errors as discussed above for sulfur. The trends in the heats of formation also follow those for the atomization energies.

For geometries 3OB also improves in comparison to MIO, although the magnitude of improvement is modest. However, some large errors remain, P–P bond lengths are overestimated leading to the largest devia-



(a) SH_2-NH_3

(b) SH_2-OH_2

(c) SH_2-SH_2

Figure 2.4: Potential energy curves for selected non-covalent interactions involving sulfur atoms studied in Ref.72; the format is similar to Fig. 2 of Ref.72. Although the artificially attractive interaction remains with DFTB3/3OB, the magnitude of interaction is significantly reduced compared to DFTB3/MIO.

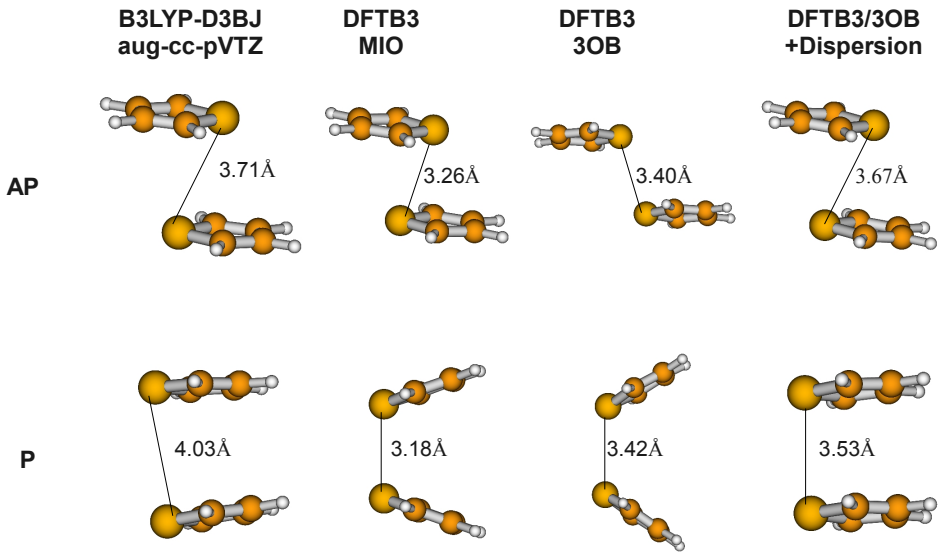


Figure 2.5: Optimized structures for anti-parallel (AP) and parallel (P) thiophene dimers at different levels of theory. The S-S distance is given in Å.

tions of over 0.1 Å. Furthermore, the P-O-P angle in Diphosphoric acid is overestimated by about 25°. Other large deviations for angles and dihedrals stem mainly from erroneous positions/orientations of a hydrogen atom. DFTB3/3OB also out-performs wavefunction based semi-empirical methods for geometries. PDDG generally underestimates H–P and P–P bond lengths in comparison to B3LYP/cc-pVTZ or aug-cc-pVTZ; an exceptional outlier is found for the bond distance within P_2 deviating by more than 0.4 Å. Considering that PDDG does not include *d*-orbitals on phosphorus the method performs remarkably well even though it is less accurate than PM6 and DFTB3.

For the purpose of representing also phosphate anions in our test sets we have compiled 9 model geometries that are important for many key biological processes (Table 2.10). The overall performance is similar to that for the neutral species. The largest deviations for bond lengths are found for P–O bonds in $[CH_3COO–PO_3]^{2-}$ and $[HPO_4]^{2-}$, which are underestimated by 0.127 and 0.044 Å, respectively. All other errors are smaller than 0.04 Å, demonstrating the excellent performance of DFTB3/3OB for bond lengths.

The special parametrization 3OB/OPhyd based on phosphate hydrolysis reactions reveals generally similar performance as 3OB (Table 2.9 and Table 2.10); computed atomization energy and heat of formation are substantially worse due to the shift of the P–O repulsive potential by about –10 kcal/mol.

Finally, to compare the structures from semi-empirical methods, we examine heats of formation using single point G3B3 energy calculations at structures determined by these methods. At DFTB3/MIO, DFTB3/3OB, and DFTB3/3OB/OPhyd geometries we find MADs of 3.4, 1.6, and 2.2 kcal/mol in comparison to the reference G3B3 values (which uses B3LYP/6-31G(d) geometries). By contrast, G3B3 single point energetics with PM6 structures gives an MAD of 8.1 kcal/mol. For PDDG, the considerable geometric inaccuracies cause the G3B3 single point energetics to deviate by as much as 27.6 kcal/mol on average. Clearly, the DFTB3 structures are considerably most reliable.

2.4.2.2 Vibrational Frequencies

A few selected unscaled harmonic vibrational frequencies calculated from DFT based and semi-empirical methods are compiled in Table 2.11. While stretching frequencies from B3LYP are consistently larger than those from BLYP, the differences are quite small (¡60 cm^{-1}). In general, DFTB3/3OB values are close to those predicted by DFT methods (only one exception for the N=P stretching frequency). PM6 shows more outliers; e.g., P–H stretching frequencies of PH_3 are overestimated by about 400 cm^{-1} in comparison to

Table 2.9: Mean and Maximum Absolute Deviations for Small Neutral Closed-Shell Phosphorus Containing Molecules[a]

property	N [b]	MP2[c]	B3LYP[d]	PBE[d]	PM6	PDDG	MIO[e]	3OB	3OB/OPhyd
E^{at} (kcal/mol)	32	12.9	33.2	8.8			82.9	7.3	21.2
$E^{\mathrm{at}}_{\mathrm{max}}$ (kcal/mol)		29.3	146.0	31.0			406.8	40.7	113.3
ΔH^0_{f} (kcal/mol)	32	14.3	34.3	8.8	15.3	15.0	82.4	7.2	21.5
$\Delta H^0_{\mathrm{f,max}}$ (kcal/mol)		32.7	146.8	30.8	57.3	43.1	405.3	40.5	114.3
					(8.1/33.6)[f]	(27.6/150.8)[f]	(3.4/13.9)[f]	(1.6/9.5)[f]	(2.2/11.1)[f]
r (Å)	130	0.007	0.007	0.018	0.029	0.069	0.020	0.012	0.012
r_{max} (Å)		0.040	0.022	0.034	0.109	0.423	0.142	0.129	0.112
a (°)	130	0.6	0.6	1.2	4.3	5.4	3.8	3.3	3.0
a_{max} (°)		2.5	2.0	4.7	32.8	18.2	33.1	25.2	21.3
d (°)	42	1.7	2.9	4.4	29.9	39.4	21.4	17.5	16.4
d_{max} (°)		6.5	17.5	25.5	139.0	152.3	70.0	101.4	119.5

[a] Atomization energies are compared to G3B3 results; bond lengths r, bond angles a, and dihedral angles d are compared to B3LYP/cc-pVTZ calculations; max stands for maximum absolute deviation. [b] Number of comparisons. [c] Basis set is cc-pVTZ. [d] Basis set is 6-31G(d) [e] Trimethylmethylenephosphorane does not converge and is excluded from the statistics. [f] Single point G3B3 ΔH^0_{f} computed at the structures optimized by semi-empirical methods; the value before/after the slash is the mean/maximum absolute deviations from G3B3, which uses B3LYP/6-31G(d) structures.

Table 2.10: Mean and Maximum Absolute Deviations for 9 Closed-Shell Phosphate Anions in Comparison to B3LYP/aug-cc-pVTZ Geometries

property[a]	N [b]	MP2[c]	B3LYP[d]	PBE[d]	PM6	PDDG[e]	MIO	3OB	3OB/OPhyd
r (Å)	53	0.004	0.007	0.019	0.024	0.031	0.017	0.013	0.013
r_{max} (Å)		0.027	0.025	0.056	0.180	0.111	0.173	0.127	0.116
a (°)	57	0.8	0.7	1.3	2.6	3.5	2.1	2.0	1.9
a_{max} (°)		4.5	6.3	10.2	8.1	14.5	11.8	10.3	9.5
d (°)	23	2.1	2.8	3.8	21.6	14.9	26.6	34.7	35.6
d_{max} (°)		6.4	9.3	10.6	98.5	101.7	67.3	79.2	82.3

[a] bond lengths r, bond angles a, and dihedral angles d; max stands for maximum absolute deviation. [b] Number of comparisons. [c] Basis set is cc-pVTZ. [d] Basis set is 6-31G(d). [e] $[CH_3COO–PO_3]^{2-}$ dissociates during geometry optimization and is therefore excluded from the statistics.

Table 2.11: Selected Vibrational Frequencies in cm^{-1}

molecule (point group)	Irrep	description	BLYP[a]	B3LYP[a]	PM6	PDDG	MIO	3OB	3OB/OPhyd
Phosphine (C_{3v})	A_1	sym. bending	998	1018	994	861	929	931	931
	E	asy. bending	1106	1038	1135	1046	1028	1033	1033
	A_1	sym. stretch	2305	2386	2760	2190	2407	2313	2313
	E	asy. stretch	2314	2394	2756	2302	2446	2354	2354
Diphosphorus ($D_{\infty h}$)	A_{1g}	P–P stretch	758	806	835	891	774	796	796
H_2P=PH_2 (C_2)	B	P–P stretch	388	419	648	477	412	412	412
Methylphosphine (C_s)	A'	C–P stretch	632	666	768	674	713	679	679
Methylenephosphine (C_s)	A'	C=P stretch	961	1004	1095	1192	1046	1065	1065
N≡P ($C_{\infty v}$)	A_1	N–P stretch	1322	1402	1496	1627	1453	1331	1331
H_2N=PH_2 (C_1)	A	N=P stretch	780	813	828	788	811	1027	1027
H_3PO_4 (C_3)	A	sym. P–O stretch	772	831	772	785	932	861	833
	E	asy. P–O stretch	863	919	834	812	1027	956	850
	A	sym. P=O stretch	1259	1318	1310	1348	1338	1289	1260
H_3PS_4 (C_3)	A	sym. P–S stretch	348	379	452	362	325	391	391
	E	asy. P–S stretch	447	489	586	415	377	461	461
	A	P=S stretch	647	673	638	536	567	655	655
MAD ($\delta\nu/\nu_{B3LYP}\%$)			5.4	–	12.2	9.4	7.0	4.8	5.0
MAX ($\delta\nu/\nu_{B3LYP}\%$)			8.6	–	54.7	20.4	22.9	26.3	26.3

[a] Basis set is cc-pVTZ.

B3LYP, whereas PDDG reproduces the overall trend quite well but also deviates substantially for the less typical bonding situations of N≡P and C=P.

2.4.2.3 Reaction Energies

Table 2.12 shows a few simple reactions that are compiled to evaluate the performance for certain bond breaking and forming processes. As this set of reactions is arbitrarily chosen, the MAD should be interpreted with caution. However, it is obvious that B3LYP performs quite well despite its poor description for atomization energies; the effect of basis set is also small. Within DFTB3/MIO the over-binding is not balanced between different bond types, thus large errors are observed (see also discussion for reactions containing sulfur above). DFTB3/3OB performs well overall but has outliers for particular species, HP=O, Hypodiphosphoric acid and also $P(OH)_5$. Note that the latter indicates already the underlying problem for hydrolysis reaction barriers where the transition state contains pentavalent phosphorus. Specifically the last reaction in Table 2.12 that changes the P coordination from 4 to 5 oxygen is reasonably well described by 3OB/OPhyd. By construction, all other reactions involving P–O bond forming or breaking deviate significantly for 3OB/OPhyd in comparison to the G3B3 method. The performance of PM6 and PDDG is less satisfying than DFTB3/3OB, with almost doubled MADs.

Table 2.12: Deviations for 16 Reaction Energies of Neutral Closed Shell Phosphorus Containing Molecules Compared to G3B3[a]

Reaction			G3B3	B3LYP[b]	B3LYP[c]	PBE[c]	PM6[d]	PDDG[d]	MIO	3OB	3OB/OPhyd
$PH_3 + H_2O$	→	$H_2P–OH + H_2$	9.4	-0.0	–2.4	–2.7	–7.5	–5.5	–27.2	–0.5	–10.9
$PH_3 + H_2O$	→	$HP{=}O + 2\ H_2$	40.1	+2.5	–2.4	–2.3	–2.3	–11.7	–39.6	–10.3	–18.5
$PH_3 + CH_4$	→	$H_3C–PH_2 + H_2$	13.7	+2.5	+2.2	+0.5	–11.6	–13.4	–8.7	–4.4	–4.4
$PH_3 + C_2H_6$	→	$H_3C–PH_2 + CH_4$	–4.6	+0.7	+1.2	+0.9	–2.1	–8.9	–7.5	–2.8	–2.8
$(CH_3)_2P{=}O + 2\ H_2O$	→	$HP(=O)(OH)_2 + 2\ CH_4$	–36.0	-4.0	–9.8	–7.2	+22.5	+12.1	–41.5	+3.3	–16.1
$PH_3 + 4\ H_2O$	→	$H_3PO_4 + 4\ H_2$	–31.5	+9.3	–4.5	–7.6	+4.3	–17.4	–92.7	–2.1	–39.0
$H_2P(=O)OH + H_2O$	→	$HP(=O)(OH)_2 + H_2$	–16.0	+1.9	–1.6	–2.2	–1.9	–0.7	–24.2	+2.7	–6.8
$HP(=O)(OH)_2 + H_2O$	→	$H_3PO_4 + H_2$	–12.1	+2.3	–1.1	–1.5	–3.2	+0.7	–32.3	–5.8	–15.1
$P(OH)_3 + H_2O$	→	$H_3PO_4 + H_2$	–24.1	+8.3	+3.1	+2.1	+14.6	+4.0	–18.7	–7.2	–13.2
$(HO)_2(O{=})P–P(=O)(OH)_2 + 2\ H_2O$	→	$2\ H_3PO_4 + H_2$	–29.5	-0.3	–3.7	+0.8	–6.3	+16.3	–60.8	–21.8	–39.7
$2\ PH_3$	→	$H_2P–PH_2 + H_2$	4.3	+2.0	+2.0	–0.3	–18.0	–37.9	–21.2	–8.1	–8.1
$PH_3 + NH_3$	→	$NP + 3\ H_2$	59.8	+8.7	+2.9	+4.0	–4.2	–35.8	–12.3	–6.4	–6.4
$PH_3 + NH_3$	→	$HN–PH + 2\ H_2$	48.3	+4.0	+1.7	+1.1	–8.4	–20.1	–2.5	–1.5	–1.5
$PH_3 + NH_3$	→	$H_2N–PH_2 + H_2$	11.4	+1.7	+0.6	–0.0	–14.0	–11.3	+3.9	+1.2	+1.2
$P(NH_2)_3 + 3\ H_2O$	→	$P(OH)_3 + 3\ NH_3$	–20.2	-6.5	–11.1	–9.3	+20.4	+4.2	–92.7	–8.6	–39.6
$H_3PO_4 + H_2O$	→	$P(OH)_5$	0.4	+0.7	–4.8	–8.8	–18.4	+16.1	–3.6	+19.3	+4.8
MAD				3.5	3.4	3.2	10.0	13.5	30.6	6.6	14.3
MAX				9.3	11.1	9.3	22.5	37.9	92.7	21.8	39.7

[a] Energies are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] Basis set is cc-pVTZ. [c] Basis set is 6-31G(d). [d] Heats of formation for H_2 are calculated as -26 and -22 kcal/mol for PM6 and PDDG. To correct for this exceptional error this value is set to 0.0 kcal/mol.

2.4.2.4 Proton Affinities and Hydrogen Bonding Energies

For a test on proton affinities we use one of our earlier compilations from Ref. 43. Similar to for sulfur, we find that when using the calculated Hubbard derivative proton affinities in the gas phase are overestimated (for both MIO/calc and 3OB/calc). The fit of the phosphorus Hubbard derivate eliminates this systematic deviation with both MIO and 3OB parameters, except for certain cases that still have error on the order of 5-10 kcal/mol. The special parametrization 3OB/OPhyd, which uses the same Hubbard parameter, also performs very well. This is not surprising as the binding situation of P–O bonds does not alter significantly for the calculation of proton affinities.

A few model systems for hydrogen bonding energies between water and phosphate esters in the gas phase are compiled in Table 2.14. In comparison to G3B3, DFTB3 slightly underestimates the binding; PM6 and PDDG generally have larger errors. Note that hydrogen bond lengths are underestimated with B3LYP/6-31G(d) (the geometry used for G3B3) in comparison to B3LYP/aug-cc-pVTZ by almost 0.04 Å on average. DFTB3/3OB predicts even shorter hydrogen bonds, the mean signed error (MSE) is –0.06 Å; for MIO it is –0.10 Å. PM6 often leads to very different structures (which is why only single point energies on PDDG structures are shown in Table 2.14), and PDDG is also featured with substantially larger errors with a MAD of 0.226 Å.

Table 2.13: 18 Proton Affinities for Phosphorus Containing Molecules in kcal/mol: Deviation of DFTB in Comparison to G3B3[a]

Molecule[b]	G3B3	B3LYP[c]	PBE[c]	PM6[d]	PDDG	MIO/calc	MIO	3OB/calc	3OB	3OB/OPhyd
H_3PO_4	334.0	–2.4	–4.3	–18.0	–2.9	+18.3	+5.5	+11.5	+4.2	+3.4
$H_2PO_4^-$	464.5	–3.3	–5.8	–17.8	+12.1	+17.2	–4.3	+10.4	–1.7	–2.0
DMPH[e]	336.3	–1.3	–4.4	–13.2	–6.4	+15.9	+4.8	+10.6	+4.5	+3.8
MMP[e]	336.7	–2.0	–4.5	–16.6	–6.0	+15.8	+3.8	+9.8	+3.1	+2.4
MMP^- [e]	460.5	–3.0	–6.0	–12.7	+11.0	+18.3	–1.2	+12.0	+1.3	+1.0
PH_3OH^+	201.6	+3.0	–0.1	+13.0	+13.6	+4.8	–0.0	+0.7	–1.5	–0.4
PH_2OHOH^+	201.6	+1.5	–0.7	+2.2	+7.3	+8.2	+2.1	+3.1	–0.1	+0.2
$PHOHOHOH^+$	200.8	–0.0	–2.5	–5.6	+4.0	+13.6	+6.2	+8.5	+3.7	+3.4
$PH_2(OH)$=O	336.6	+1.4	–0.9	+2.3	+21.6	+12.2	+3.3	+7.9	+3.3	+4.2
PH(OH)(OH)=O	334.7	–0.4	–2.6	–6.5	+12.8	+16.0	+5.3	+10.2	+4.4	+4.4
P(O)(OH)(–O–CH_2CH_2–O–)	336.3	–2.2	–5.0	–13.9	–7.3	+13.4	+2.4	+8.3	+2.1	+1.5
P(OH)(OH)(–O–CH_2CH_2–O–)(OH^*)	359.0	–2.8	–7.7	–18.3	–16.4	+12.9	+0.3	+5.9	–0.3	–1.0
P(OH^*)(OH)(–O–CH_2CH_2–O–)(OH)	350.4	–2.9	–6.7	–18.1	–17.2	+11.0	–0.4	+3.1	–2.9	–3.6
P(OH^*)(OH)(–O–CH_2CH_2–O–)(OCH_3)	351.2	–2.7	–6.5	–17.7	–20.9	+8.9	–1.4	—[f]	–3.6	–4.0
P(OH)(OCH_3)(–O–CH_2CH_2–O–)(OH^*)	359.6	–2.7	—[g]	–17.6	–18.9	+3.3	+0.1[h]	–4.1	–9.6	–10.2
P(OH^*)(OCH_3)(–O–CH_2CH_2–O–)(OH)	352.9	–2.4	–6.4	–19.1	–19.5	+10.1	–0.5	+2.6	–2.9	–3.5
P(OH)(OH)(OH)(OH^*)(OH)_ax	357.3	–2.2	–6.4	–18.8	–9.0	+14.2	–1.2	+6.8	–0.9	–1.6
P(OH)(OH)(OH)(OH^*)(OH)_eq	347.0	–3.5	–7.1	–17.1	–8.3	—[i]	–0.0	—[i]	—[i]	–3.1
MAD		2.2	4.3	13.8	12.0	12.6	2.4	7.2	2.9	3.0
MSE		–1.6	–4.3	–11.9	–2.8	+12.6	+1.4	6.7	0.2	–0.3
MAX		3.5	7.7	19.1	21.6	18.3	6.2	12.0	9.6	10.2

[a] The proton affinity is computed using potential energies at 0 K without any zero-point energy correction. [b] The molecules are given in the protonated form. [c] Basis set is 6-31+G(d,p). [d] The energy of the proton in PM6 is in error by -54 kcal/mol which has been accounted for by adding this number to the original result. See discussion at http://openmopac.net/manual/pm6_accuracy.html. [e] "DMPH" refers to dimethyl hydrogen phosphate, "MMP" to P(O)(OH)(OH)(OCH_3), and "MMP^-" to P(O)(O)(OH)(OCH_3)$^-$ [f] One ligand dissociates. [g] Converges to a slightly different minima and is excluded from the statistics. [h] This value is different to the one we reported in ref 51, where the molecule was erroneously optimized to a different local minimum. Here we stay as close to the conformation of the high-level optimized one as possible. [i] Molecule dissociates forming H_2O and has been excluded from the statistics. Depending on the basis set this dissociation also occurs for the DFT functionals PBE and B3LYP, e.g. dissociation for basis set 6-311G(2d,2p), no dissociation for basis set cc-pVTZ.

Table 2.14: Hydrogen Bonding Energies for Phosphates in kcal/mol: Deviations in Comparison to G3B3[a]

Molecule[b]	G3B3	MP2[c]	B3LYP[c]	B3LYP[d]	PBE[d]	PM6-SP[e]	PDDG	MIO	3OB	3OB/OPhyd
MMPH—OH_2	13.3	+0.2	–1.9	–0.4	+1.0	–5.4	–2.8	–2.0	–2.7	–2.7
MMP^{-1}—H_2O	17.9	–0.4	–2.7	–1.1	–0.3	–5.2	–3.1	–0.6	–1.1	–1.3
MMP^{-1}—OH_2	17.4	+0.1	–2.0	–1.0	+0.6	–4.8	–3.5	–1.3	–2.0	–2.1
MMP^{-2}—H_2O	32.3	–0.4	–3.2	–0.9	+0.5	–4.5	+2.4	–0.5	–0.6	–0.7
DMPH—OH_2	14.4	+0.0	–2.1	–0.5	+0.8	–5.3	–3.2	–2.9	–3.7	–3.8
DMP^{-1}—H_2O	18.1	–0.6	–2.9	–1.3	–0.6	–5.2	–3.0	–1.0	–1.4	–1.6
MAD		0.3	2.5	0.9	0.6	5.1	3.0	1.4	1.9	2.0
MAX		0.6	3.2	1.3	1.0	5.4	3.5	2.9	3.7	3.8

[a] The binding energy is computed with the potential energies at 0 K without any zero-point energy correction. [b] MMPH: Dihydrogenated monomethylphosphate, MMP^{-1}: monohydrogenated momomethylphosphate, MMP^{-2}: dehydrogenated monomethylphosphate, respectively for DMP: dimethylphosphate. [c] Basis set is aug-cc-pVTZ. [d] Basis set is 6-31+G(d,p). [e] During the geometry optimization using PM6 the hydrogen bonds relaxe to substantially different minima. Therefore single-point calculations have been used for comparison.

Table 2.15: Hydrogen Bonding Distances in Å for Small Phosphorus Containing Systems: Deviations in Comparison to B3LYP/aug-cc-pVTZ

System	H-bond	B3LYP[a]	MP2[a]	B3LYP[b]	B3LYP[c]	PBE[c]	PDDG[d]	MIO	3OB
DMPH—OH_2	HOH–O	1.892	–0.028	–0.004	+0.025	–0.029	–0.219	–0.062	–0.010
DMPH—OH_2	H–OH_2	1.776	–0.031	–0.055	–0.009	–0.081	–0.090	+0.018	+0.075
DMP^{-1}—H_2O	HOH–O1	2.052	–0.036	–0.028	–0.004	–0.031	–0.322	–0.170	–0.125
DMP^{-1}—H_2O	HOH–O2	2.093	–0.060	–0.031	+0.009	–0.023	–0.360	–0.191	–0.144
MMPH—OH_2	H–OH_2	1.773	–0.031	–0.054	–0.015	–0.087	–0.084	+0.020	+0.072
MMPH—OH_2	O–HOH	1.916	–0.034	–0.015	+0.024	–0.037	–0.239	–0.091	–0.034
MMP^{-1}—H_2O	HOH–O1	2.064	–0.046	–0.040	–0.002	–0.034	–0.329	–0.183	–0.140
MMP^{-1}—H_2O	HOH–O2	2.073	–0.043	–0.031	+0.008	–0.024	–0.335	–0.178	–0.127
MMP^{-1}—OH_2	H–OH_2	2.113	–0.063	–0.136	–0.019	–0.125	–0.346	–0.205	–0.102
MMP^{-1}—OH_2	O–HOH	1.638	–0.009	+0.093	+0.008	–0.030	0.003	+0.029	+0.001
MMP^{-2}—H_2O	HOH–O1	1.862	–0.040	–0.007	+0.009	–0.026	–0.193	–0.113	–0.115
MMP^{-2}—H_2O	HOH–O2	1.861	–0.040	–0.007	+0.009	–0.026	–0.192	–0.112	–0.114
MAD			0.038	0.042	0.012	0.046	0.226	0.114	0.088
MAX			0.063	0.136	0.025	0.125	0.360	0.205	0.142

[a] Basis set is aug-cc-pVTZ. [b] Basis set is 6-31G(d), this method is used within G3B3 for the geometry optimization. [c] Basis set is 6-31+G(d,p). [d] Optimization using PM6 often leads to very different geometries, thus only PDDG results are reported.

2.4.2.5 Hydrolysis reactions

Phosphorus parameters have been benchmarked and used to study phosphate hydrolysis reactions (21, 43, 86, 87). We have adopted the compilation of elementary steps of these reactions and show results for our new parametrization in Table 2.16.

In Ref. 43 transition and intermediate states, reactants and products have been optimized using B3LYP/6-31+G(d,p). First, we compare the single point energetics at those structures. DFTB3/3OB shows a rather poor performance with a MAD of 8.5 kcal/mol, although slightly better than PM6 and PDDG, which have MADs beyond 10 kcal/mol. A closer look reveals that the poor performance is systematically found for reactions where pentavalent phosphorus structures are involved, whereas the performance for reactions that include only tri- and tetravalent structures (i.e., all reactions that follow a dissociative pathway, abbreviated as “diss” in Table 2.16) are overall in reasonably good agreement with the MP2 reference. Along this line we have to distinguish even further. Some structures are trivalent, i.e., a PO_3 unit is (almost) trigonal-planar, however one water is still axially coordinated to the phosphorus. The P–water distance lies within the P–O repulsive potential, that is, the P–water distance is smaller than the cut-off of the P–O repulsive potential. Only for the geometry named diss_int, however, where no axial water is included, any P–water distance is beyond the cut-off of the repulsive potential; thus, in the reaction phosphorus reacts from a 4-coordinated to a 3-coordinated

complex. Specifically for that reaction a large error is found. In other words, the 3OB parameters seem to work well as long as P is coordinated by exactly four oxygen atoms. For coordinations with three and five oxygens large errors are obtained indicating a deficiency in properly describing different hybridization states. We have attempted tuning the *s*-orbital eigenvalue to achieve a more appropriate balance for different hybridized systems (similar to tuning the *s*-orbital eigenvalue of Nitrogen (88)), however by doing so we could not identify an overall well-performing parameter set. This indicates that the current DFTB3 methodology has limited transferability for complex phosphorus chemistry at the level of accuracy in energetics required for detailed mechanistic studies due to the small basis set, the lack of three-center integrals or multipole moments of the charge fluctuation (61). We emphasize that even for these challenging cases, as discussed below, the 3OB structures are rather reliable and MP2 single point energies at these structures have a rather small MAD of 2.5 kcal/mol compared to MP2 at B3LYP structures.

As a temporary and *ad hoc* solution to this problem we suggest the special parametrization OPhyd that reduces these errors as follows. In the first two reactions in Table 2.16 a fifth oxygen is bound to a phosphorus atom. Both the transition state (ts1) and the intermediate state (int1) are overestimated in energy by about 10 kcal/mol. Therefore, we shift the O–P repulsive potential by about 10 kcal/mol to be more attractive. While this introduces an over-binding for all P–O bonds, the energy difference between tetra- and pentavalent phosphorus is compensated. Note that for the reaction from a tetra- to tri-valent P (com1 $\rightarrow$ diss_int) a considerable error of about 10 kcal/mol remains. Due to the dominance of tetra- and penta-valent phosphorus models in our compilation, the MAD drops down to only 2.5 kcal/mol.

As second comparison we optimize all structures at the DFTB3 level. Transition states are found using the nudged elastic band technique (89). The results for DFTB3/3OB/OPhyd are impressive, with the MAD in comparison to MP2/G3large at B3LYP/6-31+G(d,p) geometries staying at 2.5 kcal/mol. However, a few structural problems appear. Hydrogen bonds are often too short by about 0.1 Å. P–O coordination of leaving groups are too long, most severely for the structures diss_prod and diss_prod2 with deviations of over 0.4 Å. The geometry n_w_com2 relaxes to a different minimum with a different type of hydrogen bonding network, therefore it is excluded from the statistics. For the standard 3OB parameters additionally the fifth ligand of int1_mmp dissociates, and this structure is therefore also excluded from the statistics.

In a third comparison we calculate MP2/G3large single-point energies at DFTB3 geometries. For

Table 2.16: Deviations of Exothermicities and Barrier Heights in Comparison to MP2/G3Large Single Point Calculations at B3LYP/6-31+G(d,p) relaxed geometries for 37 Elementary Steps in the Hydrolysis of MMP and DMP[a]

Method	MP2	B3LYP	PBE[b]	PM6	PDDG	MIO	3OB	3OB/OPhyd	3OB	3OB/OPhyd	MP2	MP2
at geometry optimized by	B3LYP	B3LYP	B3LYP	B3LYP	B3LYP	B3LYP	B3LYP	B3LYP	3OB	3OB/OPhyd	3OB	3OB/OPhyd
com1 → ts1 (MMP,B)	31.0	–1.7	–6.1	–14.5	+12.3	–7.2	+11.4	–0.8	—	–0.8	—	+0.5
com1 → int1 (MMP,E)	30.6	–1.4	–5.7	–16.4	+14.8	–7.0	+12.9	–0.8	—	–1.5	—	+0.6
com1 → ts1_2 (MMP,B)	41.5	–2.1	–7.4	–11.9	+9.4	–3.3	+9.7	+1.1	+9.8	+1.8	+1.8	+1.1
com1 → int1_2 (MMP,E)	31.0	–1.1	–5.1	–17.8	+15.6	–5.9	+15.4	+0.5	+14.5	–0.2	+0.8	+0.7
int1_2 → ts2_0 (MMP,B)	11.9	–2.0	–3.3	+7.4	–7.2	+2.7	–4.9	+1.6	–4.9	+0.3	–2.1	+3.0
int1_2 → ts2 (MMP,B)	3.6	+0.1	–0.4	+4.3	+0.4	–1.1	–2.7	–0.2	+2.3	+2.3	+6.6	+6.4
int1_2 → com2 (MMP,E)	–28.8	–0.9	+3.4	+20.4	–17.6	+4.4	–16.6	–1.8	–15.7	–0.9	–1.8	–1.8
com1 → diss_tsa (MMP,B)	36.8	–4.2	–10.2	–6.2	+11.8	+4.9	–2.8	–2.2	–7.8	–6.0	–2.0	+0.4
com1 → diss_int (MMP,E)	19.6	–6.5	–7.5	–4.6	–14.8	+0.3	–23.1	–10.5	–22.5	–10.0	–1.0	–1.4
com1_w2 → ts1_2_w2 (MMP,B)	39.9	–2.1	–8.5	–24.0	+10.0	–12.7	+5.3	–4.8	+6.2	–3.1	–1.5	+1.9
com1_w2 → int1_2a_w2 (MMP,E)	28.0	–0.5	–4.4	–21.6	+16.4	–7.9	+12.8	–2.1	+11.9	–2.8	+0.2	–0.1
int1_2a_w2 → int1_2_w2 (MMP,E)	0.4	+0.8	+0.5	+4.0	+1.3	+2.5	+2.9	+3.0	+2.7	+2.7	+1.1	+1.1
int1_2_w2 → ts2_0_w2 (MMP,B)	11.4	–1.8	–3.7	+1.4	–0.2	–5.4	–9.0	–3.5	–7.7	–2.1	–4.2	+0.0
com1_da → ts1_da (MMP,B)	55.0	–3.1	–6.8	–7.7	–11.1	–0.2	+2.0	–0.7	+4.2	–0.3	+7.4	+1.0
com1_da → int_da (MMP,E)	4.5	–2.0	–1.8	+1.4	+4.0	–2.0	–1.9	–1.9	–1.7	–1.6	–0.2	–0.4
com1 → ts1 (DMP,B)	38.6	–1.4	–5.9	–15.0	+5.8	-7.3	+9.4	–0.5	+9.9	–0.9	+0.6	+0.5
com1 → int1 (DMP,E)	35.4	–0.2	–4.4	–19.4	+11.8	–7.6	+13.8	–0.9	+12.9	–1.6	+1.8	+1.4
int1 → int1_2 (DMP,E)	1.3	–0.7	–1.3	+1.9	+1.6	+1.3	+1.3	+2.1	+0.7	+0.8	+0.1	+0.3
int1_2 → ts2 (DMP,B)	0.6	–0.5	–0.5	+3.1	–2.4	+1.2	–0.8	+1.5	–0.4	+3.4	–0.1	–0.4
int1_2 → com2 (DMP,E)	–35.2	–0.7	+4.3	+21.2	–15.2	+6.0	–15.0	–1.0	–13.5	+1.0	–1.8	–1.7
n_com1 → n_ts3 (DMP,B)	33.6	–1.4	–8.1	–11.0	+28.5	+0.2	+16.5	+4.4	+12.2	+0.8	+2.3	+1.2
n_com1 → n_int1 (DMP,E)	13.2	+0.4	–3.2	–15.6	+21.8	–4.9	+18.2	+3.6	+17.7	+3.1	+0.8	+1.1
n_int1 → n_ts4 (DMP,B)	22.9	–1.6	–5.4	+8.0	+13.4	+5.1	+0.6	+1.2	–3.6	–1.4	+0.2	+0.8
n_int1 → n_com2 (DMP,E)	–15.8	–1.9	+0.9	+23.2	–14.9	+3.7	–18.7	–4.2	–18.9	–4.2	–0.6	–1.0
DMP_P → diss_ts (DMP,B)	40.9	–2.9	–9.1	–2.3	+11.2	+8.8	+0.2	–0.7	–2.2	–3.2	–0.6	–1.5
DMP_P → diss_prod (DMP,E)	28.2	–3.8	–6.7	–12.6	–0.9	–0.9	–6.1	–5.2	–18.7	–7.3	+5.6	+4.0
diss_prod2 → diss_ts2 (DMP,B)	13.5	+0.7	–2.5	+12.1	+18.3	+10.4	+8.9	+5.1	+16.9	+4.7	–2.8	–1.2
diss_prod2 → MMP_P (DMP,E)	–29.8	+3.6	+6.7	+18.4	+12.3	+0.6	+8.1	+5.1	+19.6	+8.1	–2.2	–1.2
diss_w_reac → diss_w_ts (DMP,B)	20.9	–2.3	–7.4	–8.0	+10.8	+0.7	+1.3	+0.7	+2.1	+1.5	+1.0	+0.8
diss_w_reac → diss_w_prod (DMP,E)	18.4	–2.6	–5.6	–6.5	+1.0	+0.3	–2.5	–2.9	–12.9	–2.9	+12.0	+0.0
diss_w_prod2 → diss_w_ts2 (DMP,B)	1.9	+0.2	–1.6	+0.9	+10.9	+0.3	+3.3	+2.8	+16.6	+3.4	–8.2	–0.1
diss_w_prod2 → diss_w_reac2 (DMP,E)	–21.0	+2.8	+6.2	+14.5	+7.7	–0.7	+2.2	+2.3	+14.7	+2.0	–8.3	+0.1
n_w_com1 → n_w_ts3 (DMP,B)	28.2	–1.8	–10.1	–24.3	+32.4	–8.9	+13.7	+0.2	+10.6	–0.2	–0.2	+0.2
n_w_com1 → n_w_int1 (DMP,E)	13.1	+1.0	–3.1	–15.7	+22.2	–5.8	+17.5	+2.8	+17.3	+2.7	+0.6	+0.7
n_w_int1 → n_w_int2 (DMP,E)	–0.5	+0.5	+0.4	–0.0	+0.8	+0.6	+1.0	+1.0	+0.6	+0.6	+0.4	+0.4
n_w_int2 → n_w_ts4 (DMP,B)	15.1	–2.3	–6.2	–2.6	+17.2	–3.1	–2.9	–2.7	–0.9	–1.1	+4.2	+2.1
n_w_int2 → n_w_com2 (DMP,E)	–13.0	–2.0	+1.3	+23.8	–14.8	+4.1	–19.0	–4.4	—	—	—	—
MAD		1.8	4.7	11.5	11.4	4.1	8.5	2.5	9.8	2.5	2.5	1.1
MAX		6.5	10.3	24.3	32.4	12.7	23.1	10.5	22.5	10.0	12.0	6.4

[a] Compilation from ref 43; no zero-point corrections are included in either exothermicity or barrier heights. All quantities are given in kcal/mol. The processes are labeled in the notation of ref 43, "E" stands for "Exothermicity", "B" for "Barrier". [b] Basis set is 6-31+G(d,p).

DFTB3/3OB the MAD drops down to only 2.5 kcal/mol; for comparison, PBE/6-31+G(d,p) single points at B3LYP structures have a MAD of 4.7 kcal/mol (see Table 2.16). Thus, even though the energetics for tri- and penta-valent structures are inaccurate, DFTB3/3OB can be corrected using higher level methods to yield very good energetics. For the special parametrization 3OB/OPhyd results are excellent, the MAD is only 1.1 kcal/mol showing that the geometrical differences between B3LYP/6-31+G(d,p) and 3OB/OPhyd are of minor relevance, or from a different perspective, the potential energy surface seems very shallow and leads to similar energies.

2.5 Conclusion

Considering the importance of phosphorus and sulfur in (bio)chemistry, we extend the parameterization of the approximate density functional tight binding approach, DFTB3, to these elements. The parameterization is carried out in a framework consistent with the DFTB3/3OB set(52) for O, N, C and H, thus the resulting parameters can be used to describe a broad set of organic and biologically relevant P/S-containing molecules. The 3*d* orbitals, which are known to be polarization functions essential to the proper description of structure and energetics of P/S-containing molecules, are included in the parameterization, in contrast to a few popular wave function based semi-empirical methods such as PM3/PDDG. Compared to the previous parameterizations of DFTB2 and DFTB3 (the 'MIO' set), the electronic parameter of the current 3OB parameterization were chosen with the goal of minimizing errors in atomization energies. As a result, systematic over-binding of covalent interactions is avoided, leading to generally more reliable reaction energies.

The parameters are tested with a fairly diverse set of molecules of chemical and biological relevance. We focus on the geometries, reaction energies, proton affinities and hydrogen bonding interactions of these molecules; vibrational frequencies are also examined, although less systematically. These properties calculated at the DFTB3/3OB level are compared to results of DFT (B3LYP and PBE), *ab initio* (MP2, G3B3) and several popular semi-empirical methods (PM6 and PDDG), as well as predictions of DFTB3 with the older parameterization (the MIO set). In general, DFTB3/3OB is a major improvement over the previous parameterization (DFTB3/MIO), and for the majority cases tested here, it also outperforms PM6 and PDDG, especially for structural properties, vibrational frequencies, hydrogen bonding interactions and proton affinities. For reaction energies, DFTB3/3OB exhibits major improvement over DFTB3/MIO, due mainly to significant reduction of errors in atomization energies; compared to PM6 and PDDG, DFTB3/3OB also generally performs better although the magnitude of improvement is more modest. Compared to high-level calculations, DFTB3/3OB is most successful at predicting geometries; the energetics are largely on par with results of DFT (especially PBE, which is the functional used in DFTB3 for deriving the relevant matrix elements) using a medium basis set (e.g., 6-31G(d)), although larger errors are seen more often with DFTB3. We note that since we focus on molecules of biological interest, our test cases here do not feature a large number of radical species.

The extensive tests also indicate that there are remaining cases for which the current DFTB3 methodology has trouble properly describing. For example, SO (singlet) and SO_3 have large errors (11.8 and 34.4

kcal/mol, respectively) in atomization energies, thus reactions involving these species are generally not well described by DFTB3/3OB. Another example involves substantial underestimate of hydrogen-bonding distances for many (especially charged) sulfur containing species; comparison to DFT calculations using different functionals and basis sets suggest that the errors are likely due in part to the intrinsic limitations of the PBE functional and also to the minimal basis set used in DFTB3. It is worth noting that the computed hydrogen bonding energies, nevertheless, are generally fairly reliable and even slightly better than PBE/6-31+G(d,p). Finally, another remaining limitation is that DFTB3/3OB is still incapable of providing reliable energetics for phosphate hydrolysis reactions that involve a change in the coordination number of the phosphorus. Thus the current DFTB3 model has limited transferability for complex phosphorus chemistry at the level of accuracy in energetics required for detailed mechanistic investigations. As a temporary solution, we develop a specific parameterization, 3OB/OPhyd, where an overbinding is introduced to O-P interactions to empirically minimize the errors for a set of reference reactions. Although this is clearly not a satisfactory solution for the long-term, the 3OB/OPhyd can serve as a useful tool for exploring the energy landscape of phosphoryl transfer reactions in biological systems, especially for structural properties. Alternatively, the 3OB structures remain reliable for these challenging cases and one may still use the standard 3OB parameters to obtain structures and improve energetics with higher level methods.

Due to computational efficiency and general robustness for structural property predictions, QM/MM calculations using DFTB3/3OB as the QM level will be an effective computational approach for the analysis of condensed phase systems. The energetic properties from such calculations should be taken as semi-quantitative in nature and can be improved by combining with *ab initio* QM/MM calculations. For example, the fact that single point high-level calculations using DFTB3/3OB geometries often exhibit small errors suggest that, for example, DFTB3/3OB is likely effective as the low-level QM approach that drives the sampling for reliable free energy simulations in the framework of dual-level QM/MM calculations(20, 24, 90–93). Along this line, for phosphate hydrolysis reaction, the closer agreement between 3OB/OPhyd and higher level methods in energies makes 3OB/OPhyd better suited for dual-level QM/MM free energy calculations than the standard 3OB set. Since many reactions that involve phosphorus/sulfur chemistry implicate metal ions, a pressing need is to extend the DFTB3/3OB parameterization to these ions, such as the alkali metal ions, zinc and copper; this will be reported in separate work in the near future. Along this line, we note that although

there is currently an impressive effort to develop DFTB parameters for the entire periodic table in a largely automated fashion(94), developing parameters appropriate for many (bio)chemical applications would require careful calibration and refinement as we have done here for sulfur and phosphorous. Finally, the remaining limitations highlight that it is also worthwhile continuing to improve the DFTB3 methodology, such as by including multipole terms for the charge fluctuation and better description of polarization and short-range repulsions.

References

[1] Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. *Molecular biology of the cell*; Garland Publishing, Inc., 1994.

[2] Hanson, S. R.; Best, M. D.; Wong, C. H. Sulfatase: structure, mechanism, biological activity, inhibition and synthetic utility. *Angew. Chem. Int. Ed.* **2004**, *43*, 5736–5763.

[3] Mauritz, K. A.; Moore, R. B. State of understanding of Nafion. *Chem. Rev.* **2004**, *104*, 4535–4585.

[4] Anthony, J. E. Functionalized acenes and heteroacenes for organic electronics. *Chem. Rev.* **2006**, *106*, 5028–5048.

[5] Westheimer, F. H. Why nature chose phosphates. *Science* **1987**, *235*, 1173–1178.

[6] Knowles, J. R. Enzyme catalyzed phosphoryl transfer reactions. *Annu. Rev. Biochem.* **1980**, *49*, 877–919.

[7] Cleland, W. W.; Hengge, A. C. Enzymatic mechanisms of phosphate and sulfate transfer. *Chem. Rev.* **2006**, *106*, 3252–3278.

[8] Kiaris, H.; Spandidos, D. A. Mutations of Ras Genes in Human Tumors. *Int. J. Oncol.* **1995**, *7*, 413–421.

[9] Roberts, P. J.; Der, C. J. Targeting the Raf-MEK-ERK mitogen-activated protein kinase cascade for the treatment of cancer. *Oncogene* **2007**, *26*, 3291–3310.

[10] Cohen, P. Protein kinases - the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **2002**, *1*, 309–315.

[11] Davies, S. P.; Reddy, H.; Caivano, M.; Cohen, P. Specificity and mechanism of action of some commonly used protein kinase inhibitors. *Biochem. J.* **2000**, *351*, 95–105.

[12] Garber, K. Rapamycin's resurrection: A new way to target the cancer cell cycle. *J. Natl. Cancer Inst.* **2001**, *93*, 1517–1519.

[13] Collins, I.; Workman, P. Design and development of signal transduction inhibitors for cancer treatment: Experience and challenges with kinase targets. *Curr. Signal Transduct. Ther.* **2006**, *1*, 13–23.

[14] Robertson, J. G. Enzymes as a special class of therapeutic target: clinical drugs and modes of action. *Curr. Opin. Struct. Biol.* **2007**, *17*, 674–679.

[15] Zhang, Z. Y. Protein tyrosine phosphatases: Structure and function, substrate specificity, and inhibitor development. *Annu. Rev. Pharm. Toxi.* **2002**, *42*, 209–234.

[16] Kamerlin, S. C. L.; Sharma, P. K.; Prasad, R. B.; Warshel, A. Why nature really chose phosphate. *Q. Rev. Biophys.* **2013**, *46*, 1–132.

[17] Zhang, Y. Pseudobond ab initio QM/MM approach and its applications to enzyme reactions. *Theor. Chem. Acc.* **2006**, *116*, 43–50.

[18] Hu, H.; Yang, W. T. Free Energies of Chemical Reactions in Solution and in Enzymes with Ab Initio Quantum Mechanics/Molecular Mechanics Methods. *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.

[19] Senn, H. M.; Thiel, W. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.* **2009**, *48*, 1198–1229.

[20] Plotnikov, N. V.; Kamerlin, S. C. L.; Warshel, A. Paradynamics: An Effective and Reliable Model for Ab lnitio QM/MM Free-Energy Calculations and Related Tasks. *J. Phys. Chem. B* **2011**, *115*, 7950–7962.

[21] Yang, Y.; Yu, H.; Cui, Q. Extensive Conformational Transitions Are Required to Turn On ATP Hydrolysis in Myosin. *J. Mol. Biol.* **2008**, *381*, 1407–1420.

[22] Wong, K. Y.; Lee, T. S.; York, D. M. Active Participation of the Mg^{2+} Ion in the Reaction Coordinate of RNA Self-Cleavage Catalyzed by the Hammerhead Ribozyme. *J. Chem. Theo. Comp.* **2011**, *7*, 1–3.

[23] Hou, G. H.; Cui, Q. QM/MM analysis suggests that Alkaline Phosphatase (AP) and Nucleotide pyrophosphatase/phosphodiesterase (NPP) slightly tighten the transition state for phosphate diester hydrol-

ysis relative to solution: implication for catalytic promiscuity in the AP superfamily. *J. Am. Chem. Soc.* **2012**, *134*, 229–246.

[24] Hou, G. H.; Cui, Q. Stabilization of Different Types of Transition States in a Single Enzyme Active Site: QM/MM Analysis of Enzymes in the Alkaline Phosphatase Superfamily. *J. Am. Chem. Soc.* **2013**, *135*, 10457–10469.

[25] Duarte, F.; Amrein, B. A.; Kamerlin, S. C. L. Modeling catalytic promiscuity in the alkaline phosphatase superfamily. *Phys. Chem. Chem. Phys.* **2013**, *15*, 11160–11177.

[26] Zhao, Y. L.; Köppen, S.; T. Frauenheim, An SCC-DFTB/MD Study of the Adsorption of Zwitterionic Glycine on a Geminal Hydroxylated Silica Surface in an Explicit Water Environment. *J. Phys. Chem. C* **2011**, *115*, 9615–9621.

[27] Mori, T.; Hamers, R. J.; Pedersen, J. A.; Cui, Q. An Explicit Consideration of Desolvation is Critical to Binding Free Energy Calculations of Charged Molecules at Ionic Surfaces. *J. Chem. Theo. Comp.* **2013**, *9*, in press.

[28] Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. PDDG/PM3 and PDDG/MNDO: Improved semiempirical methods. *J. Comput. Chem.* **2002**, *23*, 1601–1622.

[29] Winget, P.; Selçuki, C.; Horn, A. H. C.; Martin, B.; Clark, T. Towards a "next generation" neglect of diatomic differential overlap based semiempirical molecular orbital technique. *Theor. Chem. Acc.* **2003**, *110*, 254–266.

[30] Korth, M.; Thiel, W. Benchmarking semiempirical methods for thermochemistry, kinetics, and noncovalent interactions: OMx methods are almost as accurate and robust as DFT-GGA methods for organic molecules. *J. Chem. Theo. Comp.* **2011**, *7*, 2929–2936.

[31] Giese, T. J.; York, D. M. Density-functional expansion methods: Grand challenges. *Theor. Chem. Acc.* **2012**, *131*, 1145.

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

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

[34] Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. Development of effective quantum mechanical/molecular mechanical (QM/MM) methods for complex biological processes. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.

[35] Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Phys. Rev. B* **1995**, *51*, 12947–12957.

[36] Krüger, T.; Elstner, M.; Schiffels, P.; Frauenheim, T. Validation of the density-functional based tight-binding approximation method for the calculation of reaction energies and other data. *J. Chem. Phys.* **2005**, *122*, 114110.

[37] Otte, N.; Scholten, M.; Thiel, W. Looking at self-consistent-charge density functional tight binding from a semiempirical perspective. *J. Phys. Chem. A* **2007**, *111*, 5751–5755.

[38] Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. L. Comparison of SCC-DFTB and NDDO-based semiempirical molecular orbital methods for organic molecules. *J. Phys. Chem. A* **2006**, *110*, 13551–13559.

[39] Niehaus, T. A.; Elstner, M.; Frauenheim, T.; Suhai, S. Application of an approximate density-functional method to sulfur containing compounds. *J. Mol. Struct.: THEOCHEM* **2001**, *541*, 185–194.

[40] Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. Modeling zinc in biomolecules with the self consistent charge-density functional tight binding (SCC-DFTB) method: Applications to structural and energetic analysis. *J. Comput. Chem.* **2003**, *24*, 565–581.

[41] Moreira, N. H.; Dolgonos, G.; Aradi, B.; A. L. da Rosa,; T. Frauenheim, Toward an Accurate Density-Functional Tight-Binding Description of Zinc-Containing Compounds. *J. Chem. Theo. Comp.* **2009**, *5*, 605–614.

[42] Saha, S.; Pal, S.; Sarkar, P.; Rosa, A. L.; T. Frauenheim, A Complete Set of Self-Consistent Charge

Density-Functional Tight-Binding Parametrization of Zinc Chalcogenides (ZnX; X=O, S, Se, and Te). *J. Comp. Chem.* **2012**, *33*, 1165–1178.

[43] Yang, Y.; Yu, H.; York, D.; Elstner, M.; Cui, Q. Description of phosphate hydrolysis reactions with the Self-Consistent-charge Density-Functional-Tight-Binding (SCC-DFTB) theory. 1. Parameterization. *J. Chem. Theo. Comp.* **2008**, *4*, 2067–2084.

[44] Cai, Z.; Lopez, P.; Reimers, J. R.; Cui, Q.; Elstner, M. Application of the computationally efficient self-consistent-charge density-functional tight-binding method to magnesium-containing molecules. *J. Phys. Chem. A* **2007**, *111*, 5743–5750.

[45] Kubar, T.; Bodrog, Z.; Gaus, M.; Aradi, B.; Frauenheim, T.; Elstner, M. Parametrization of the SCC-DFTB Method for Halogens. *J. Chem. Theo. Comp.* **2013**, *9*, 2939–2949.

[46] Zheng, G. S.; Witek, H. A.; Bobadova-Parvanova, P.; Irle, S.; Musaev, D. G.; Prabhakar, R.; Morokuma, K. Parameter calibration of transition-metal elements for the spin-polarized self-consistent-charge density-functional tight-binding (DFTB) method: Sc, Ti, Fe, Co, and Ni. *J. Chem. Theo. Comp.* **2007**, *3*, 1349–1367.

[47] Dolgonos, G.; Aradi, B.; Moreira, N. H.; Frauenheim, T. An Improved Self-Consistent-Charge Density-Functional Tight-Binding (SCC-DFTB) Set of Parameters for Simulation of Bulk and Molecular Systems Involving Titanium. *J. Chem. Theo. Comp.* **2010**, *6*, 266–278.

[48] Grundkotter-Stock, B.; Bezugly, V.; Kunstmann, J.; Cuniberti, G.; T. Frauenheim,; Niehaus, T. A. SCC-DFTB Parametrization for Boron and Boranes. *J. Chem. Theo. Comp.* **2012**, *8*, 1153–1163.

[49] Elstner, M. SCC-DFTB: What is the proper degree of self-consistency. *J. Phys. Chem. A* **2007**, *111*, 5614–5621.

[50] Yang, Y.; Yu, H.; York, D.; Cui, Q.; Elstner, M. Extension of the self-consistent-charge density-functional tight-binding method: Third-order expansion of the density functional theory total energy and introduction of a modified effective coulomb interaction. *J. Phys. Chem. A* **2007**, *111*, 10861–10873.

[51] Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB). *J. Chem. Theo. Comp.* **2011**, *7*, 931–948.

[52] Gaus, M.; Goez, A.; Elstner, M. Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theo. Comp.* **2013**, *9*, 338–354.

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

[54] Reed, A. E.; v. R. Schleyer, P. Chemical bonding in hypervalent molecules – the dominance of ionic bonding and negative hyperconjugation over d-orbital participation. *J. Am. Chem. Soc.* **1990**, *112*, 1434–1445.

[55] Dobado, J. A.; Martinez-Garcia, H.; Molina, J. M.; Sundberg, M. R. Chemical bonding in hypervalent molecules revised. Application of the Atoms in Molecules Theory to Y-3 X and Y-3 XZ (Y = H or CH3; X = N, P or As; Z = O or S) compounds. *J. Am. Chem. Soc.* **1998**, *120*, 8461–8471.

[56] Stewart, J. J. P. *Encyclopedia of Computational Chemistry*; Wiley, 1998; Chapter PM3.

[57] Tubert-Brohman, I.; Guimarães, C. R. W.; Jorgensen, W. L. Extension of the PDDG/PM3 Semiempirical Molecular Orbital Method to Sulfur, Silicon, and Phosphorus. *J. Chem. Theo. Comp.* **2005**, *1*, 817–823.

[58] Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

[59] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[60] Gaus, M.; Chou, C.-P.; Witek, H.; Elstner, M. Automatized parametrization of SCC-DFTB repulsive potentials: Application to hydrocarbons. *J. Phys. Chem. A* **2009**, *113*, 11866–11881.

[61] Gaus, M.; Cui, Q.; Elstner, M. Density Functional Tight Binding (DFTB): Application to organic an biological molecules. *WIREs Comput. Mol. Sci.* **2013**, In press.

[62] Bodrog, Z.; Aradi, B. Possible improvements to the self-consistent-charges density-functional tight-binding method within the second order. *Phys. Stat. Sol. B* **2012**, *249*, 259–269.

[63] Bodrog, Z.; Aradi, B.; Frauenheim, T. Automated repulsive parametrization for the DFTB method. *J. Chem. Theo. Comp.* **2011**, *7*, 2654–2664.

[64] Köhler, C.; Frauenheim, T.; Hourahine, B.; Seifert, G.; Sternberg, M. Treatment of collinear and non-collinear electron spin within an approximate density functional based method. *J. Phys. Chem. A* **2007**, *111*, 5622–5629.

[65] Kohler, C.; Seifert, G.; Frauenheim, T. Density functional based calculations for Fe-n (n $<=$ 32). *Chem. Phys.* **2005**, *309*, 23–31.

[66] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) theory for molecules containing first and second-row atoms. *J. Chem. Phys.* **1998**, *109*, 7764–7776.

[67] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-3 and density functional theories for a larger experimental test set. *J. Chem. Phys.* **2000**, *112*, 7374–7383.

[68] Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Assessment of Gaussian-3 and density-functional theories on the G3/05 test set of experimental energies. *J. Chem. Phys.* **2005**, *123*, 124107.

[69] Gonzalez-Lafont, A.; Truong, T. N.; Truhlar, D. G. Direct dynamics calculations with NDDO (neglect of diatomic differential overlap) molecular orbital theory with specific reaction parameters. *J. Phys. Chem.* **1991**, *95*, 4618–4627.

[70] Giese, T. J.; Chen, H. Y.; Dissanayake, T.; Giambasu, G. M.; Heldenbrand, H.; Huang, M.; Kuechler, E. R.; Lee, T. S.; Panteva, M. T.; Radak, B. K.; York, D. M. A Variational Linear-Scaling Framework to Build Practical, Efficient Next-Generation Orbital-Based Quantum Force Fields. *J. Chem. Theo. Comp.* **2013**, *9*, 1417–1427.

[71] http://www.dftb.org accessed on July 20th, 2012.

[72] Petraglia, R.; Corminboeuf, C. A Caveat on SCC-DFTB and Noncovalent Interactions Involving Sulfur Atoms. *J. Chem. Theo. Comp.* **2013**, *9*, 3020–3025.

[73] Frisch, M. J. et al. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[74] Frisch, M. J. et al. Gaussian 09 Revision A.1. Gaussian Inc. Wallingford CT 2009.

[75] Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063–1079.

[76] Tirado-Rives, J.; Jorgensen, W. L. Performance of B3LYP density functional methods for a large set of organic molecules. *J. Chem. Theo. Comp.* **2008**, *4*, 297–306.

[77] Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

[78] Witek, H. A.; Morokuma, K.; Stradomska, A. Modeling vibrational spectra using the self-consistent charge density-functional tight-binding method II. Infrared spectra. *J. Chem. Theo. Comp.* **2005**, *4*, 639–655.

[79] Małolepsza, E.; Witek, H. A.; Morokuma, K. Accurate vibrational frequencies using the self-consistent-charge density-functional tight-binding method. *Chem. Phys. Lett.* **2005**, *412*, 237–243.

[80] Fishtik, I. Unique Stoichiometric Representation for Computational Thermochemistry. *J. Phys. Chem. A* **2012**, *116*, 1854–1863.

[81] Range, K.; Riccardi, D.; Cui, Q.; Elstner, M.; York, D. M. Benchmark calculations of proton affinities and gas-phase basicities of molecules important in the study of biological phosphoryl transfer. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3070–3079.

[82] Riccardi, D.; Yang, S.; Cui, Q. Proton transfer function of carbonic anhydrase: Insights from QM/MM simulations. *Biochim. Biophys. Acta* **2010**, *1804*, 342–351.

[83] Goyal, P.; Ghosh, N.; Phatak, P.; Clemens, M.; Gaus, M.; M. Elstner,; Cui, Q. Proton storage site in bacteriorhodopsin: new insights from QM/MM simulations of microscopic pK_a and infrared spectra. *J. Am. Chem. Soc.* **2011**, *133*, 14981–14997.

[84] Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comp. Chem.* **2011**, *32*, 1456–1465.

[85] Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density functional-theory based treatment. *J. Chem. Phys.* **2001**, *114*, 5149–5155.

[86] Yang, Y.; Cui, Q. The hydrolysis activity of adenosine triphosphate in myosin: a theoretical analysis of anomeric effects and the nature of the transition state. *J. Phys. Chem. A* **2009**, *113*, 12439–12446.

[87] Yang, Y.; Cui, Q. Does water relay play an important role in phosphoryl transfer reactions? Insights from theoretical study of a model reaction in water and tert-butanol. *J. Phys. Chem. B* **2009**, *113*, 4930–4939.

[88] Gaus, M. Extension and Parametrization of an Approximate Density Functional Method for Organic and Biomolecules. Ph.D. thesis, Karlsruhe Institute of Technology, Karlsruhe, Germany, 2011; publicly available at www.bibliothek.kit.edu (accessed May 31, 2012).

[89] Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B., Cicotti, G., Coke, D., Eds.; World Scientific, 1998; p 385.

[90] Marti, S.; Moliner, V.; Tuñón, I. Improving the QM/MM description of chemical processes: A dual level strategy to explore the potential energy surface in very large systems. *J. Chem. Theo. Comp.* **2005**, *1*, 1008–1016.

[91] Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K. E.; Schutz, M.; Thiel, S.; Thiel, W.; Werner, H. J. High-accuracy computation of reaction barriers in enzymes. *Angew. Chem. Int. Ed.* **2006**, *45*, 6856–6859.

[92] Retegan, M.; Martins-Costa, M.; Ruiz-Lopez, M. F. Free energy calculations using dual-level Born–Oppenheimer molecular dynamics. *J. Chem. Phys.* **2010**, *133*, 064103.

[93] Polyak, I.; T. Benighaus,; Boulanger, E.; Thiel, W. Quantum mechanics/molecular mechanics dual Hamiltonian free energy perturbation. *J. Chem. Phys.* **2013**, *139*, 064105.

[94] Wahiduzzaman, M.; Oliveira, A. F.; Philipsen, P.; Zhechkov, L.; E. van Lenthe,; Witek, H. A.; Heine, T. DFTB Parameters for the Periodic Table: Part 1, Electronic Structure. *J. Chem. Theo. Comp.* **2013**, *9*, 4006–4017.

Chapter 3

Parameterization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications[1]

3.1 Introduction

Magnesium and zinc are two of the most common metal cofactors of numerous metalloenzymes and other biomolecules(1–3). The binding of Mg^{2+} to DNA and RNA helps to maintain their structures by stabilizing highly negatively charged phosphates. ATP, the main source of energy in cells, is bound to Mg^{2+} in its biologically active form(4). In addition, a large number of enzymes involved in the biochemistry of nucleic acids bind Mg^{2+} for catalytic activity; for example, Mg^{2+} is essential for DNA replication and phosphoryl transfers(5). Zinc is the second most abundant cation in the human body, so Zn^{2+}has important biological functions in numerous biomolecular systems(6, 7) such as zinc-finger class of transcription factors, signaling proteins, transport/storage proteins as well as enzymes. Zinc ion contributes to catalysis in more than 300 enzymes and serves structural role in many proteins and enzymes.(8–10)

To describe the structural properties of $Mg^{2+}$and Zn^{2+}in solution and biomolecules, molecular mechanical (MM) force field parameters have been developed(11–15). Some of the advanced force field models such as SIBFA(16), AMOEBA(17, 18) and Drude oscillator models(19) treat electronic polarization and charge transfer(20) terms explicitly and haven been demonstrated to describe potential energy surfaces in close

[1]Reprinted (adapted) with permission from "Parameterization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications", X. Lu, M. Gaus, M. Elstner and Q. Cui, *J. Phys. Chem. B* **119**, 1062 (2015). Copyright 2015 American Chemical Society.

agreement with quantum mechanics (QM) calculations.(21, 22) For chemical reactions that involve these metal ions, however, MM models are no longer appropriate. Even for structural properties, when the binding mode of ligands is flexible(23–25), it is challenging to describe the metal site with a MM model.

An appropriate modeling approach in this context is the QM/MM framework, where the metal site of interest is treated with QM while the rest enzyme/solution is described by an empirical force field. For relatively rigid metal sites, *ab initio* QM and QM/MM methods based on either reaction path or free energy simulations have been demonstrated to be powerful for mechanistic analysis(26–30). The cost of these calculations for flexible systems, however, remains very high. For instance, one key feature of zinc coordination is its flexibility:(23–25) a zinc ion can adopt multiple binding modes with the coordination number ranging from 4 to 6, underlining the importance of developing efficient QM(/MM) methods that complement *ab initio* based methods. The dynamical nature of the metal binding mode is particularly important to enzymes that feature solvent accessible active sites(31–33). For these applications, semi-empirical QM based QM/MM simulations can potentially strike the balance between computational accuracy and conformational sampling.

The traditional semi-empirical QM methods based on the Neglect-of-Diatomic-Differential-Overlap (NDDO) approximation, such as MNDO,(34, 35) MNDO/d,(36) AM1(37) and PM3,(38) however, give in general rather poor geometries for Mg/Zn-containing molecules. For example, bond lengths computed from AM1,(39, 40) PM3, MNDO and MNDO/d(40) deviate by roughly 0.1 Å on average in comparison to B3LYP/6-311++G** for small Mg-complexes.(41) Recently developed AM1/d parameters(42) for magnesium provide a clear improvement in accuracy compared to AM1 by incorporating d orbitals. According to a recent test,(43) the mean absolute deviations (MAD) of bond distances for AM1,(44) PM3, PM3(ZnB),(45) MNDO/d(46) and PM6(47) are more than 0.05 Å compared to CCSD(T) values for small zinc-containing compounds. The importance of a reliable treatment for zinc sites to the description of enzyme catalysis is illustrated by our recent study of alkaline phosphatase (AP)(48, 49), where we showed that previous AM1/d-PhoT/MM calculations(50, 51) likely led to incorrect transition states due to distortions of the bimetallic zinc motif.

The Self-Consistent-Charge Density-Functional Tight-Binding (SCC-DFTB) method provides a promising alternative to NDDO approaches.(52–54) It is derived by expanding the DFT total energy up to second order around a reference density, which is usually taken to be the sum of atomic charge densities. The computational speed of SCC-DFTB is comparable to NDDO, and it has been rather extensively benchmarked for

reaction energies, geometries and vibrational frequencies for a large number of small molecules(55–61). The fact that SCC-DFTB is formulated in a DFT framework suggests that it is potentially more reliable for metal ions than NDDO based methods, which are based on Hartree-Fock. Specifically for Mg(41) and Zn(62, 63), SCC-DFTB(/MM)(64–66) has been successfully applied to a rather diverse set of biological and chemical applications that include these ions(48, 49, 67–74).More recently, DFTB has been extended to include third-order contributions, termed as DFTB3.(75) Compared to the second-order formulation, DFTB3 considers the charge dependence of the atomic Hubbard parameter and therefore provides improved description of chemical properties such as proton affinity(53, 75, 76), which is important in many biological applications. With the most recent parameterization referred to as 3OB(77, 78), the electronic properties and therefore atomization energies have been substantially improved; as a result, the DFTB3 approach is shown to have comparable performance as DFT with double-zeta-plus-polarization basis sets for many systems of biological and chemical interest, although certain limitations remain(78, 79).

Considering the recent success of DFTB3/3OB, here we extend the 3OB parameterization of DFTB3 to magnesium and zinc, again focusing on chemical and biological applications. In the next section, we briefly summarize the DFTB3 methodology, the procedures for parameterization and simulations details for benchmark calculations. We then present data from multiple levels of test sets in the gas phase, which are followed by additional benchmark calculations in both solution and enzymes. Finally, we summarize this work with several conclusions.

3.2 Computational Methods

3.2.1 Theory

The DFTB3-parameterization of magnesium and zinc follows the protocol outlined in the previous publications(77, 80) and extends the 3OB parameter set that so far includes elements C, H, N, O(77), S and P(78). We first briefly review the formulation of DFTB3(54, 75).

The total energy of DFTB3 is given by

$$\begin{aligned} E^{\mathrm{DFTB3}} &= E^{\mathrm{H0}} + E^{\gamma} + E^{\Gamma} + E^{\mathrm{rep}} \\ &= \sum_{iab} \sum_{\mu \in a} \sum_{\nu \in b} n_i c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{ab} \Delta q_a \Delta q_b \gamma_{ab} + \frac{1}{3} \sum_{ab} \Delta q_a^2 \Delta q_b \Gamma_{ab} + \frac{1}{2} \sum_{ab} V^{\mathrm{rep}}_{ab}. \end{aligned} \tag{3.1}$$

Here $H^0_{\mu\nu}$ is the effective Kohn-Sham Hamiltonian evaluated at the atomic reference density ρ_0, which

is determined by solving the Kohn-Sham equations for the atom in the presence of an additional harmonic confining potential with a confining radius r^{wf} for the atomic basis set and a different radius r^{dens} for the density. Off-diagonal $H^0_{\mu\nu}$ integrals are computed in a two-center approximation together with the PBE(81) exchange-correlation functional and an atomic basis set consisting of Slater functions. The Slater functions are defined by l_{max} as the highest orbital angular momentum included, n_{max}, and α_0, α_1, ..., α_4, the exponents of the Slater functions. In principle, basis functions for each orbital angular momentum can be compressed differently; in this work, however, *s* and *p*-orbitals are confined with the same compression radius for Mg, and *s*, *p* and *d*-orbitals (compression radius for d-orbitals is defined by the parameter r^{wfd}) are confined with the same compression radius for Zn. The diagonal elements $H^0_{\mu\mu}$ are equal to the calculated atomic eigenvalues ε_x ($x = s$, p or d orbital). No confinement of the orbitals are applied here in order to yield the correct dissociation limit. We optimize ε_{p} as discussed in the following section.

The atomic Hubbard parameter U_a in the second order term of the energy E^{γ} describes the electron-electron interaction within one atom and enters the γ-function, which interpolates the on-site and the long-range electronic interaction of atom pairs. The Hubbard parameters are calculated from DFT as the first derivative of the highest occupied orbital with respect to its occupation number. For the third order term E^{Γ}, an additional parameter per element U^{d}, the derivative of the Hubbard parameter with respect to charge has to be determined. For CHNO it was calculated as the second derivative of the highest occupied orbital with respect to its occupation number(77); for Mg and Zn we find a significant advantage to optimize it.

3.2.2 Reference Systems and Parameter Fitting

The parameterization for Mg/Zn used herein follows that specified previously for the parameterization of CHNOPS(77, 78); two distinct groups of parameters have to be optimized, the electronic parameters appearing in the first three terms of the DFTB3 total energy (Eq. 3.1) and the repulsive potential.

The electronic parameters are summarized in Table 3.1. The following parameters are optimized to improve the overall performance:

- r^{wf} and r^{dens}: The general procedure was to start from the electronic parameters used in the MIO set(75, 77) (which are $r^{\mathrm{wf}}_{\mathrm{Mg}} = 5.0$, $r^{\mathrm{dens}}_{\mathrm{Mg}} = 12.0$, $r^{\mathrm{wf}}_{\mathrm{Zn}} = 4.9$, $r^{\mathrm{dens}}_{\mathrm{Zn}} = 10.0$). r^{wf} and r^{dens} are fitted in a "brute force" manner to achieve good performance on bond, angle, vibrational frequencies as well as reaction energies.
- ε_{p}: The eigenvalues are in principle computed directly from atomic DFT calculations and enter as

Table 3.1: Overview of Electronic Parameters (in Atomic Units if not Unitless) [a]

Parameter	Mg	Zn
$l_{\max}$	1	2
$n_{\max}$	2	2
α_0	0.50	0.50
α_1	1.11	1.39
α_2	2.45	3.87
α_3	5.42	10.78
α_4	12.00	30.00
r^{wf}	5.5	4.2
r^{wfd}	0.0	4.2
r^{dens}	14.0	9.0
ε_{s}	–0.17267741	–0.21408495
ε_{p}	–0.02	0.02
ε_{d}	–	–0.38831893
E^{spin}	0.0	0.0
U	0.2246	0.2663
U^{d}	–0.02	–0.03

[a] As described in the text, U, E^{spin}, $\varepsilon_{s,d}$, n_{max}, α_i are consistent with the standard choices of DFTB parameterization and not subject to optimizations. By contrast, r^{wf}, r^{wfd}, r^{dens}, ε_{p} and U^{d} are adjusted based on properties of molecules in the fitting set.

diagonal elements into the Hamiltonian matrix without the additional confining potential. Exceptions are made here to magnesium and zinc since the p orbitals are often not heavily involved in the coordination of $Mg^{2+}$and Zn^{2+}. ε_{p} is calculated to be -0.04914 au for Mg and -0.04261 au for Zn. ε_{p} are fitted in the way that p-orbitals are less involved in molecular bonding, so as to improve sequential ligation energies and proton affinities.

• U^{d}: The Hubbard derivative is generally computed from atomic DFT calculations. In the case of Mg and Zn, they are -0.0496 au and -0.0544 au, respectively. We found a significant improvement of proton affinities after optimizing U^{d} for Mg and Zn. It is important to note that fitting of the Hubbard derivatives does not significantly alter most properties of neutral molecules like equilibrium geometries, so with all the other parameters held fixed, U^{d} is tuned as the last step in the parameterization procedure.

Although the repulsive potential, which contains nuclear-nuclear repulsion and double-counting terms that depend on the reference density, can be directly computed, practical application to molecular systems with reasonable accuracy requires fitting the repulsive potential. The repulsive potential between atom type A and type B (V_{ab}^{rep} in Eq. 3.1) is described by a spline function in the covalent binding region. A cutoff is introduced at which the function and first three derivatives reach zero. The remaining degrees of freedom for the spline

function are fitted to high-level *ab initio* values of geometries, reaction energies and vibrational stretching frequencies. The overall quality of repulsive potential is sensitive to the chosen division points. Therefore, establishing reliable repulsive potential parameters remains a time-consuming step despite the progress of partially automatized fitting procedures.(80, 82)

The repulsive potentials are fitted to several properties as summarized in Table 3.2. Reference values are calculated from *ab initio* methods that are known to produce results close to experimental data. For magnesium, optimized geometries of small closed-shell molecules are calculated from B3LYP(83–85)/aug-cc-pVTZ, reaction energies are calculated from the G3B3(86–88) method. For zinc, both optimized geometries and reaction energies are calculated from B3LYP/aug-cc-pVTZ. Additional equations for the fitting process are prepared using a few vibrational stretching frequencies determined from BLYP/aug-cc-pVTZ (unscaled) using the harmonic approximation. B3LYP with good quality basis set has been shown to describe structures of Zn complexes in good agreement with X-ray(89) and CCSD(T)(43**?**) data. However, heats of formation (HOF) predicted from B3LYP/aug-cc-pVTZ deviates $\sim$10 kcal/mol on average from experimental data of eight small complexes even though CCSD(T)/aug-cc-pVTZ does not do much better with a mean unsigned error of $\sim$7 kcal/mol.(90) B97-1(91), B97-2(92) and B98(93, 94) have also been tested because they have been reported to be reliable for reproducing experimental data for small transition metal complexes;(95) geometries from the B97 family functionals and B3LYP are very similar and reaction energies differ by $\sim$2 kcal/mol. Despite shortcoming in HOF, we have chosen to use the B3LYP functional because it is computationally less intensive so that large sets of consistent reference data are accessible and it systematically gives reliable results for binding energies and proton affinities, which are of major importance to our work. Since the systems studied here are relatively small in size, dispersion interactions are not expected to impact the structural and energetic properties of interest. For a fairly large active site model for a zinc enzyme, we do discuss briefly the impact of dispersion. For even larger cases, we note that empirical dispersion models have been developed and tested for both SCC-DFTB(96, 97) and DFTB3(98).

3.2.3 Condensed Phase Benchmark Calculations

To supplement the gas phase calculations, simulations are carried out with DFTB3/MM to investigate if the parameterized model works in an explicit condensed-phase environment. They are important tests because

Table 3.2: Parameters Defining the Repulsive Potentials[a]

Reference geometries for force equations

MgH_2, $Mg(CH_3)_2$, MgO, $[Mg(H_2O)_2]^{2+}$, $[Mg(H_2O)_6]^{2+}$, $[Mg(NH_3)_4]^{2+}$, MgS,
$[Mg(SH_2)_3(SH)]^{+}$, $Mg(PH_2)_2$, $[Mg(PH_3)_4]^{2+}$, ZnH_2, $[Zn(CH_3)]^{+}$, $[Zn(CO)]^{2+}$
$[Zn(H_2O)_3]^{2+}$, $Zn(OH)_2$, $Zn(H)(NH_2)$, $[Zn(NH_3)_4]^{2+}$, $[Zn(SH)]^{+}$, $[Zn(SH_2)_3]^{2+}$
$[Zn(SH_2)_3(SH)]^{+}$, $[Zn(PH_2)]^{+}$, $[Zn(PH_3)_4]^{2+}$, $[Zn(PH_3)_3(PH_2)]^{+}$

Reference Potential energy differences in kcal/mol

$[Mg(H_2O)_2]^{2+}$	$\rightarrow$	$[Mg(H_2O)]^{2+} + H_2O$	73.3
$[Mg(H_2O)_3]^{2+}$	$\rightarrow$	$[Mg(H_2O)_2]^{2+} + H_2O$	59.2
$[Mg(CO)_4]^{2+}$	$\rightarrow$	$[Mg(CO)_3]^{2+} + CO$	30.9
$[Mg(NH_3)]^{2+}$	$\rightarrow$	$Mg^{2+} + NH_3$	97.7
$[Mg(NH_3)_2]^{2+}$	$\rightarrow$	$[Mg(NH_3)]^{2+} + NH_3$	83.9
$[Mg(NH_3)_3]^{2+}$	$\rightarrow$	$[Mg(NH_3)_2]^{2+} + NH_3$	63.1
$[Mg(NH_3)_4]^{2+}$	$\rightarrow$	$[Mg(NH_3)_3]^{2+} + NH_3$	50.1
$[Mg(NH_3)_5]^{2+}$	$\rightarrow$	$[Mg(NH_3)_4]^{2+} + NH_3$	28.6
$[Mg(NH_3)_6]^{2+}$	$\rightarrow$	$[Mg(NH_3)_5]^{2+} + NH_3$	26.3
$[Mg(SH_2)_2]^{2+}$	$\rightarrow$	$[Mg(SH_2)]^{2+} + SH_2$	62.4
$[Mg(SH_2)_3]^{2+}$	$\rightarrow$	$[Mg(SH_2)_2]^{2+} + SH_2$	43.9
$[Mg(SH_2)_4]^{2+}$	$\rightarrow$	$[Mg(SH_2)_3]^{2+} + SH_2$	34.5
$[Mg(SH_2)_5]^{2+}$	$\rightarrow$	$[Mg(SH_2)_4]^{2+} + SH_2$	20.0
$[Mg(PH_3)_2]^{2+}$	$\rightarrow$	$[Mg(PH_3)]^{2+} + PH_3$	66.9
$[Mg(PH_3)_4]^{2+}$	$\rightarrow$	$[Mg(PH_3)_3]^{2+} + PH_3$	34.8
$[Zn(H_2O)]^{2+}$	$\rightarrow$	$Zn^{2+} + H_2O$	104.3
$[Zn(H_2O)_4]^{2+}$	$\rightarrow$	$[Zn(H_2O)_3]^{2+} + H_2O$	43.0
$[Zn(CO)]^{2+}$	$\rightarrow$	$Zn^{2+} + CO$	81.6
$[Zn(CO)_2]^{2+}$	$\rightarrow$	$[Zn(CO)]^{2+} + 2\ CO$	149.8
$[Zn(CO)_3]^{2+}$	$\rightarrow$	$[Zn(CO)]^{2+} + 3\ CO$	190.1
$[Zn(SH_2)]^{2+}$	$\rightarrow$	$Zn^{2+} + SH_2$	126.1
$[Zn(SH_2)_2]^{2+}$	$\rightarrow$	$Zn^{2+} + 2\ SH_2$	210.5
$[Zn(PH_3)_4]^{2+}$	$\rightarrow$	$[Zn(PH_3)_3]^{2+} + PH_3$	30.5

Potential	Division points (a.u.)	Additional equations (a.u.)
C–Mg	3.6, 4.0, 4.5, 5.1	$V''(3.959) = 0.073$
H–Mg	2.8, 3.0, 3.3, 3.9	$V''(3.214) = 0.063$
N–Mg	3.2, 3.9, 4.2, 5.5	$V''(3.581) = 0.15$, $V''(3.906) = 0.065$
O–Mg	2.8, 3.8, 4.5, 5.1	$V''(3.330) = 0.23$, $V''(3.660) = 0.070$
P–Mg	4.4, 4.7, 5.3, 5.7	$V''(4.851) = 0.025$
S–Mg	3.8, 4.5, 5.3, 5.8	$V''(4.344) = 0.05$
C–Zn	3.5, 3.7, 3.9, 4.2	$V''(3.698) = 0.19$
H–Zn	2.4, 2.8, 3.0, 4.0	$V''(2.914) = 0.10$
N–Zn	3.0, 3.6, 4.4, 5.0	$V''(3.677) = 0.09$
O–Zn	2.8, 3.5, 3.8, 4.9	
P–Zn	4.0, 4.3, 4.7, 5.2	$V''(4.482) = 0.075$
S–Zn	3.5, 4.2, 4.7, 5.7	

[a] See text for the methods used as references for different properties.

the ultimate goal is to use DFTB3/3OB in enzyme simulations.

3.2.3.1 Mg^{2+} and Zn^{2+} Ion Solvation and Relative pK_a Calculations

Both MM and QM/MM simulations are performed using CHARMM(99). The generalized solvent boundary potential (GSBP)(100, 101) is used to treat long-range electrostatic interactions in MD simulations. The system is set up with an inner region of 20 Å. All molecules are located in the inner region, and the outer region is described with a constant dielectric continuum with $\varepsilon_w = 80$. To be consistent with the GSBP protocol, electrostatic interactions among inner region atoms are treated with the extended electrostatic model(102) in which interactions beyond 12 Å are modeled with multipolar expansions, including the dipolar and quadrupolar terms. The reaction field matrix **M** is evaluated using 400 spherical harmonics for GSBP simulations. The ion is fixed at the center of the sphere. Water molecules are described with the modified version of TIP3P(103). In the classical MM simulations, standard CHARMM force field(104) parameters are used for Mg^{2+} and Zn^{2+}. The QM/MM calculations employ the same water sphere as in MM simulations. The QM region includes the metal ion together with the nearest 6 water molecules (referred to as small QM region) or the nearest 25 water molecules (referred to as large QM region) which include the first and the second solvation shells. All the remaining water molecules are represented by the TIP3P water model. In the large QM region simulations, to ensure the QM water molecules are always closest to the metal ion, Flexible Inner Region Ensemble Separator (FIRES(105)) restraining potential is imposed to any outer sphere MM water molecules that are closer to the ion than the most distant inner sphere QM water molecule. The FIRES potential takes the form $E_{\text{FIRES}} = \frac{1}{2}\sum k_{\text{FIRES}}(r_i - R_{\text{inner}})^2$ where R_{inner} is defined as the distance between the metal ion and the most distant inner QM water molecule from it. k_{FIRES} is set to be 100 kcal/(mol·Å^2). FIRES restrains are not necessary for the small QM region simulations because water molecules in the first solvation shell are strongly bound to the metal ion. All bonds involving hydrogen are constrained with the SHAKE algorithm,(106) allowing a 1 fs time step for MD propagation. Simulations are equilibrated for 500 ps followed by a production run of 500 ps.

To further evaluate the performance of DFTB3/3OB for Mg and Zn from the energetics perspective, we perform calculations of relative pK_a of Mg^{2+}/Zn^{2+}-bound water in a water sphere using the thermodynamic integration (TI)(107, 108) approach within the dual topology single coordinate (DTSC) scheme.(66, 109) In this approach, the dominant contribution to the total free energy of deprotonation is from the electrostatic free

energy change ($\Delta F_{M^{2+}\cdot(H_2O)_5(HOH/D)}$) associated with converting the acidic proton to a dummy atom (D), i.e., transforming $M^{2+}\cdot(H_2O)_6$ to $M^{2+}\cdot(H_2O)_5(HOD)^-$, where M represents the metal ion, Mg^{2+} or Zn^{2+}. The corresponding free energy derivative is given by,

$$\frac{\partial \Delta F_{M^{2+}\cdot(H_2O)_5(HOH/D)}}{\partial \lambda} = \langle U_{\text{elec}}^{M^{2+}\cdot(H_2O)_5(HOD)^-} - U_{\text{elec}}^{M^{2+}\cdot(H_2O)_6} \rangle_\lambda, \tag{3.2}$$

which represents the QM/MM energy difference averaged for a specific coupling parameter λ using the same set of coordinates for both protonation states. The total electrostatic energy contribution ($\Delta F_{M^{2+}\cdot(H_2O)_5(HOH/D)}$) is determined by integrating the free energy derivatives (∂F/∂λ) over λ from 0 to 1. Estimation of the absolute pK_a is in principle possible with the DTSC-TI protocol, but the solvation free energy of a proton has to be taken into account. To avoid the difficulty of getting accurate experimental/calculated value for this term, we compute the relative pK_a between $Mg^{2+}\cdot(H_2O)_6$ and $Zn^{2+}\cdot(H_2O)_6$ in water so that other contributions are canceled out to a great extent, such as the zero-point energy difference between the protonated and deprotonated states and van der Waals interactions involving the acidic proton(66, 109). Since $\Delta F_{M^{2+}\cdot(H_2O)_5(HOH/D)}$ is electrostatic in nature, we expect that the free energy derivative depends largely linearly on the coupling parameter λ in this simple system; this is supported by actual calculations (see below). We carry out pK_a calculations with both small QM region and large QM region as defined above. To avoid dissociation of hydroxide from the metal ion in the fully deprotonated state (λ = 1 window $M^{2+}\cdot(H_2O)_5(HOD)^-$), a NOE potential is added to the distance between M^{2+} and OH^-. The NOE potential takes the form

$$\begin{aligned} E_r &= 0.0 \qquad\qquad r_{\min} < r < r_{\max} \\ &= 0.5 \cdot k_{\max} \cdot (r - r_{\max})^2 \qquad r_{\max} < r \end{aligned} \tag{3.3}$$

in which $r_{\min}$ and $r_{\max}$ set the interval between which the restraining potential is zero; they are taken to be 1.8 and 2.5 Å, respectively. $k_{\max}$ is set to be 1000 kcal/(mol·Å^2). To prevent the dummy atom D from being too close to the metal site in the fully deprotonated state, a weak angular constrain is applied to the M-O-D angle to keep it no smaller than 120°.

3.2.3.2 Myosin-II Enzyme Setup: test for a Mg site

The enzyme model is built from a high-resolution X-ray structure for the Mg·ADP·vanadata bound myosin(110) (PDB code 1VOM), which corresponds to the hydrolyzing state of the *D. discoideum* myosin-II motor domain.(111, 112) As described in our earlier work(68, 113, 114), starting from the PDB structure, the

ADP·vanadata is replaced by ATP or ADP·Pi. The QM region includes the tri-phosphate and part of the ribose group of ATP, three water molecules (include the lytic water) in the active site, Ser 181, Ser 236, the Mg^{2+} ion and all its ligands (Thr 186 and Ser 237 and two water molecules). Only side chains of protein residues are included in the QM region and link atoms are introduced to saturate the valence of the QM boundary atoms. ATP is assumed to be fully deprotonated and the titratable groups are kept in their normal protonation states (i.e., Lys, Arg are protonated, Asp, Glu are deprotonated) which are consistent with simple estimate of pK_a using the Poisson-Boltzmann (PB) approach.(115) GSBP is used as the simulation protocol. The system is divided into a 24 Å spherical inner region centered at the Mg ion. Newtonian equations-of-motion are solved for the MD region (within 20 Å), and Langevin equations-of-motion are solved for the buffer region (20-24 Å) with a temperature bath of 300 K. Protein atoms in the buffer region are harmonically constrained with force constants determined from the crystallographic B-factors.(116) Protein atoms in the MM region are described by the all-atom CHARMM22 force field(104) and water molecules are described with the TIP3P model. All bonds involving hydrogen are constrained using SHAKE, and time step is set to 1 fs. All the water molecules in the inner region are subject to a weak GEO type of restraining potential to keep them inside the inner sphere with the MMFP module in CHARMM. The static field due to outer-region atoms, $\phi_s^{(io)}$, is evaluated with the linear PB equations. The reaction field matrix **M** is evaluated with 625 spherical harmonics. In the PB calculations, the protein dielectric constant is set as 1, the water dielectric constant is set as 80, and 0.0 M salt concentration is used. Our previous(117) and recent(118) analyses indicated that the GSBP protocol is reliable for a site well shielded from the bulk solvent.

To study the structural flexibility in the active site of myosin before and after ATP hydrolysis, potential of mean force (PMF) simulations have been carried out. The reaction coordinate is defined as the distance between Mg^{2+} and atom O in Ser 237 (see Fig. 3.6). The umbrella sampling approach(119) is used to restrain the system along the reaction coordinate with a force constant of 150 kcal/mol·$Å^{-2}$. 32 windows are used for PMF and 500 ps simulation is performed for each window. Convergence of PMF is monitored by examining the overlap of reaction coordinate distributions sampled in different windows. Weighted histogram analysis (WHAM)(120) is used to post-analyze the probability distribution to obtain the PMF along the reaction coordinate. Errors are calculated from block average.

3.2.3.3 Alkaline Phosphatase Enzyme setup: test for a bimetallic zinc site

The enzyme model is constructed based on the X-ray structures for the *E. coli* Alkaline Phosphatase (AP) mutant R166S with bound inorganic phosphate at 2.05 Å resolution (PDB code 3CMR(121)). The substrate methyl *p*-nitrophenyl phosphate ($MpNPP^-$) is first "mutated" to the orientation with the -OMe group oriented towards the magnesium ion (denoted as α orientation) starting from PDB structure. All basic and acidic amino acids are kept in their physiological protonation states except for Ser102 in AP, which is accepted to be the nucleophile and deprotonated in the reactive complex. Water molecules are added following the standard protocol of superimposing the system with a water droplet of 25 Å radius centered at $Zn1^{2+}$ (see Fig. 3.7 for atomic labels) and removing water molecules within 2.8 Å from any heavy atoms resolved in the crystal structure. GSBP is used to treat long-range electrostatic interactions in MD simulations. In the QM/MM simulations, as described in details in our previous work,(48, 49) the QM region includes the two zinc ions and their 6 ligands (Asp51, Asp369, His370, Asp327, His412, His331), Ser102 and the substrate $MpNPP^-$. Only side chains of protein residues are included in the QM region and link atoms are added between C_α and C_β atoms. A NOE potential is added to the C-O bonds in Asp51 in AP to prevent artificial polarization(48). The integration time step is 1 fs, and all bonds involving hydrogens are constrained using SHAKE. The DFTB3/MM results are also compared to MM simulations using a conventional nonbonded zinc model(14) (referred to as Coulomb scheme) or short-long effective functions (SLEF1)(25) model developed by Zhang and co-workers. Protein atoms in the MM region are described by the all-atom CHARMM22 force field and water molecules are described with the TIP3P model.

To further benchmark the applicability of DFTB3/3OB to the reaction of interest, we also study an active site model that includes all QM atoms in the QM/MM enzyme model. The C_β carbons are fixed at their positions in the crystal structure during geometry optimization; the positions of the link atoms used to saturate the valence of the C_β atoms in the active site model are also fixed. The reactant (Michaelis) complex and transition state are located for $MpNPP^-$ (Methyl *p*-nitrophenyl phosphate), $MmNPP^-$ (Methyl *m*-nitrophenyl phosphate) and MPP^- (Methyl phenyl phosphate) using DFTB3 and B3LYP with the 6-31+G(d,p) basis set. The Minimum Energy Path (MEP) calculations are carried out by one dimensional adiabatic mapping at the DFTB3 level; the reaction coordinate is the antisymmetric stretch involving the breaking and forming P-O bonds. Subsequently, the saddle point is further refined by conjugated peak refinement (CPR(122)) to obtain

precise transition state structure. Single point energy calculations at DFTB3 and B3LYP geometries are then carried out using B3LYP, M06,(123) PBE and MP2 methods using a larger basis set of 6-311++G(d,p). The D3(124) dispersion correction are added for B3LYP, M06 and PBE methods.

3.3 Results and Discussions

In the following subsections, we present benchmark calculations for DFTB3/3OB regarding geometries, bond dissociation energies and proton affinities for Mg/Zn containing molecules and compare the results with commonly used *ab initio* methods in the gas phase. We also report DFTB3 results when the old set of MIO parameters are used(75); DFTB3/MIO use parameters as defined in Ref.75: U^d = -0.23, -0.16, -0.13, -0.19, -0.14, -0.09 au for C, H, N, O, P and S and ζ=4.2; U^d for Mg/Zn is fitted to -0.1 to achieve the best sequential bond dissociation energies and proton affinities. Unless noted otherwise, energies are given for geometries that are optimized at the respective level of theory. For PM6, we note that the choice of heat of formation for protons is essential to proton affinities; we take -54.0 kcal/mol for protons as suggested by the MOPAC program. To compare the potential energy surfaces between different computation levels, zero-point energies (ZPE) are not included. For the calculations of non-covalent interactions, dispersion interactions are not included because most test molecules discussed here are small although we are aware of the importance of dispersion in general. We explore explicitly the effects of dispersion to a non-trivial active site model of AP that involves zinc. For DFTB3 calculations, we have used our in-house DFTB code, all PM6, DFT and MP2 calculations are performed using the Gaussian09 software package.(125) Following the gas phase benchmarks, we show results on Mg^{2+}/Zn^{2+}-water structural properties and relative pK_a in solution. Finally we apply DFTB3/3OB to enzyme systems to further test the new parameters.

3.3.1 Magnesium

3.3.1.1 Geometries

The "small" magnesium test set is composed of 55 molecules, which are simplified biologically relevant ligated magnesium species of the form $Mg^{2+}[X]_n$ with X = H_2O, NH_3, SH_2, PH_3, CO for n = 1-4/6 and X = OH^-, SH^-, NH_2^- and PH_2^- for n = 1-2. Table 3.3 summarizes the results. Good agreement is seen between MP2 and B3LYP with small or large basis set. Of the semiempirical methods, geometries are described very well by DFTB3 models and particular well by DFTB3/3OB. The MADs in bond distances are 0.013, 0.017

Table 3.3: Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Magnesium Test Set

property[a]	N [b]	MP2[c]	B3LYP[d]	PM6	DFTB3/MIO	DFTB3/3OB
r (Å)	120	0.009	0.009	0.050	0.017	0.013
r_{max} (Å)		0.051	0.024	0.284	0.141	0.072
a (°)	117	0.9	0.5	3.4	3.6	2.9
a_{max} (°)		23.7	10.2	38.7	41.3	45.3

[a] max stands for maximum absolute deviation. [b] number of comparisons. [c] basis set aug-cc-pVTZ. [d] basis set 6-31+G(d,p).

and 0.050 Å for DFTB3/3OB, DFTB/MIO and PM6, respectively. 3OB improves over MIO, although the magnitude of improvement is modest. The largest bond distance error of 0.072 Å from DFTB3/3OB is found in Mg-S bond in $[Mg(SH_2)]^{2+}$, which is likely due to the minimal basis nature of DFTB3 for a polarizable ligand. Other deviations in bond distances are substantially smaller in DFTB3/3OB. Molecules containing S or P are described poorly by PM6 within our test case(78); PM6 systematically underestimates Mg-S bond distance by 0.2 Å and Mg-P bond distance by 0.1 Å. For angles, DFTB3/3OB differs from the reference data by a MAD of 2.9°, which is of the same order with other semiempirical methods. The largest angle derivation of 45.3° is for angle Mg-O-H in $[Mg(OH)]^{+}$, for which DFTB3/3OB predicts a linear rather than bent structure; interestingly, MP2 also predicts a structure almost linear, while PM6 is close to the B3LYP result. Future development of DFTB3 to include multipoles may improve the geometry in this case. It is worth pointing out that the Mg-S-H angle is generally overestimated by $\sim$13° in DFTB3/3OB.

3.3.1.2 Sequential Bond Dissociation Energies

We define sequential bond dissociation energies (sBDE) corresponding to the reaction

$$M^{2+}L_n \rightarrow M^{2+}L_{n-1} + L \tag{3.4}$$

where M is Mg^{2+}/Zn^{2+} and L is a neutral or charged ligand.

Table 3.4 shows the sequential bond dissociation energies for compounds containing Mg^{2+}. Values obtained for DFTB3/3OB and other methods are compared to those obtained using G3B3. The differences between B3LYP, MP2 and G3B3 are relatively small; however, complexes with charged ligands give several deviations greater than 5 kcal/mol. DFTB3/3OB shows a MAD of 2.4 kcal/mol for complexes with neutral ligands and 13.8 kcal/mol for complexes with charged ligands. Compared to DFTB3/MIO, DFTB3/3OB improve substantially for complexes involving P, S elements and charged ligands. For the other semiempirical

Table 3.4: Sequential Bond Dissociation Energies of Magnesium Containing Molecules Compared to G3B3[a]

Molecule[b]	G3B3	MP2[c]	B3LYP[c]	PM6	DFTB3/MIO	DFTB3/3OB	G3B3//DFTB3[d]
$[Mg(H_2O)_2]^{2+}$	73.3	-3.5	-0.1	-10.9	-3.4	+0.5	+0.6
$[Mg(H_2O)_3]^{2+}$	59.2	-2.0	-1.1	-11.7	-1.5	+1.4	-0.3
$[Mg(H_2O)_4]^{2+}$	49.1	-2.3	-3.0	-11.0	-3.3	-1.9	+1.2
$[Mg(H_2O)_5]^{2+}$	34.7	-2.2	-5.0	-7.4	-6.7	-6.4	-0.2
$[Mg(H_2O)_6]^{2+}$	32.9	-2.5	-5.9	-11.0	-6.8	-6.4	+0.4
$[Mg(CO)_2]^{2+}$	46.3	-0.9	-0.3	+10.0	-2.9	-3.8	-0.6
$[Mg(CO)_3]^{2+}$	36.1	+0.5	-0.9	+1.3	-0.3	-2.3	-0.7
$[Mg(CO)_4]^{2+}$	30.9	-0.1	-2.8	-1.5	-0.1	-2.6	-0.9
$[Mg(NH_3)_2]^{2+}$	83.9	-3.4	-0.3	-4.4	-13.8	-6.4	+0.5
$[Mg(NH_3)_3]^{2+}$	63.1	-1.8	-2.0	-9.2	-7.5	-1.4	+0.2
$[Mg(NH_3)_4]^{2+}$	50.1	-1.8	-3.7	-8.8	-3.9	+1.2	+0.2
$[Mg(NH_3)_5]^{2+}$	28.6	-1.0	-5.3	-2.9	-5.8	-3.5	-0.2
$[Mg(NH_3)_6]^{2+}$	26.3	-1.5	-6.2	-5.5	-5.6	-2.5	-0.2
$[Mg(SH_2)_2]^{2+}$	62.4	-3.0	-0.3	-7.0	-7.9	-2.9	-1.2
$[Mg(SH_2)_3]^{2+}$	43.9	-1.1	-2.6	-12.1	-4.5	+0.1	-0.6
$[Mg(SH_2)_4]^{2+}$	34.5	-0.8	-4.4	-13.3	-2.8	+1.2	+1.7
$[Mg(SH_2)_5]^{2+}$	20.0	+0.2	-6.1	-12.6	+0.0	+1.1	-0.8
$[Mg(SH_2)_6]^{2+}$	20.7	-0.5	-7.6	-14.0	-1.8	-0.1	+0.5
$[Mg(PH_3)_2]^{2+}$	66.9	-2.3	-0.5	-2.4	-8.6	-0.1	+0.7
$[Mg(PH_3)_3]^{2+}$	44.5	-0.8	-2.9	-18.6	-5.4	+2.3	+0.7
$[Mg(PH_3)_4]^{2+}$	34.8	-0.6	-4.5	-17.9	-3.8	+3.2	+0.6
$[Mg(OH)_2]$	243.1	-7.2	-4.6	-7.1	-21.0	-14.0	+0.1
$[Mg(NH_2)_2]$	237.9	-7.3	-6.6	-4.6	-9.4	-12.0	+0.9
$[Mg(SH)_2]$	201.2	-3.1	-5.0	-30.3	-16.1	-15.8	+0.1
$[Mg(PH_2)_2]$	188.8	-1.8	-5.2	+11.0	-20.9	-13.3	+0.6
MAD		2.1	3.5	9.9	6.6	4.3	0.6
MAX		7.3	7.6	30.3	21.0	15.8	1.7

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] The non-dissociated is listed. [c] Basis set aug-cc-pVTZ. [d] G3B3 single point calculations on top of DFTB3/3OB geometries.

method PM6, the deviation is 9.9 kcal/mol compared to G3B3. The large error is anticipated because geometries containing S and P are often poor. Note that we compare potential energy differences for all the methods except PM6 which intrinsically considers heats of formation. We have also carried out single-point G3B3 calculations at DFTB3/3OB geometries, and the MAD for Mg^{2+} compounds with neutral and charged ligands is less than 1 kcal/mol. These results highlight the similarities between DFTB3 and B3LYP optimized structures, suggesting that DFTB3/3OB is of particular value for the study of biologically relevant magnesium-containing molecules.

Table 3.5: Gas Phase Proton Affinities of Magnesium Containing Molecules Compared to G3B3[a]

Molecule[b]	G3B3	MP2[c]	B3LYP[c]	PM6	DFTB3/MIO	DFTB/3OB	G3B3//DFTB3[d]
$[Mg(H_2O)_1]^{2+}$	95.6	+0.6	-3.5	+0.1	-5.5	-13.9	-0.8
$[Mg(H_2O)_2]^{2+}$	115.6	+0.0	-1.4	-5.4	-2.2	-9.0	-0.9
$[Mg(H_2O)_3]^{2+}$	133.1	+0.6	+0.2	-7.1	+3.0	-1.2	-0.4
$[Mg(H_2O)_4]^{2+}$	149.2	-0.6	-0.1	-11.4	+4.8	+2.3	+0.8
$[Mg(H_2O)_5]^{2+}$	160.3	-2.2	-1.9	-15.7	+3.2	+2.0	+1.1
$[Mg(H_2O)_6]^{2+}$	170.5	-3.0	-2.2	-24.3	+0.2	-0.2	+0.8
$[Mg(NH_3)_1]^{2+}$	120.2	+1.5	-3.5	+2.6	-10.5	-15.7	+0.0
$[Mg(NH_3)_2]^{2+}$	144.1	+0.8	-0.4	+1.5	-10.2	-14.1	+0.2
$[Mg(NH_3)_3]^{2+}$	163.9	+0.5	+0.3	+1.9	-10.6	-13.6	+0.5
$[Mg(NH_3)_4]^{2+}$	181.3	-0.1	+0.6	+1.0	-6.4	-7.5	+1.0
$[Mg(NH_3)_5]^{2+}$	190.5	-0.4	+0.9	+4.9	-5.2	-6.1	+1.4
$[Mg(NH_3)_6]^{2+}$	200.7	+1.8	-0.5	+5.0	-6.2	-5.4	+2.3
$[Mg(SH_2)_1]^{2+}$	83.4	+2.1	-1.8	+2.4	-6.7	-0.7	-3.6
$[Mg(SH_2)_2]^{2+}$	107.0	+0.8	+0.5	+4.6	-6.2	+0.8	-2.9
$[Mg(SH_2)_3]^{2+}$	123.1	+0.4	+1.5	+3.3	-6.2	+2.1	-3.5
$[Mg(SH_2)_4]^{2+}$	135.4	+0.0	+2.0	+1.0	-5.7	+4.2	-1.7
$[Mg(SH_2)_5]^{2+}$	143.1	-0.2	+2.0	-5.7	-6.9	+3.9	-3.9
$[Mg(SH_2)_6]^{2+}$	151.1	-0.6	+1.8	-12.4	-8.7	+3.5	-3.0
$[Mg(PH_3)_1]^{2+}$	96.8	+4.1	-1.8	+12.6	-12.8	-1.1	+0.6
$[Mg(PH_3)_2]^{2+}$	129.4	+1.7	+0.4	+10.2	-15.3	-0.4	+1.4
$[Mg(PH_3)_3]^{2+}$	150.1	+1.0	+1.2	+8.2	-17.6	-1.7	+1.4
$[Mg(PH_3)_4]^{2+}$	164.2	+0.7	+1.7	+3.0	-18.1	-0.7	+1.2
MAD		1.1	1.4	6.6	7.8	5.0	1.5
MAX		4.1	3.5	24.3	18.1	15.7	3.9

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] The protonated form is listed. [c] Basis set aug-cc-pVTZ. [d] G3B3 single point calculations on top of DFTB3/3OB geometries.

3.3.1.3 Proton Affinities

Table 3.5 lists the proton affinities for 22 biological relevant molecules $Mg^{2+}(XH_m)_n$ where X = O, N, S, P obtained from G3B3, DFTB3/3OB and other methods. MADs for MP2 and B3LYP are generally small ($\sim$1 kcal/mol). The discrepancy between DFTB3/3OB and G3B3 is 5 kcal/mol, whereas DFTB3/MIO and PM6 yield slightly larger MADs of 7.8 kcal/mol and 6.6 kcal/mol, respectively. We note that G3B3 single points at DFTB3/3OB optimized structures lead to small MAD of 1.5 kcal/mol and MAX of 3.9 kcal/mol (due to the unsatisfactory deprotonated $[Mg(SH_2)_n]^{2+}$ structures) compared to the original G3B3, again highlighting the good quality of DFTB3/3OB structures. Future improvements should focus on the nitrogen-containing species, an issue we noted in previous studies(75, 77).

3.3.1.4 Large Test Set

To further test DFTB3/3OB parameters we have compiled a "large test set" that models biologically relevant ligands of magnesium with larger molecules than those discussed above. The set mainly consists of model magnesium compounds that are simplified active sites of crystal structures collected in a survey(126). Although $Mg^{2+}$is generally 6-coordinated, complexes with coordination number of 4 and 5 are also investigated. All complexes are constructed in the way that all ligands are coordinated to the central magnesium ion rather than via hydrogen bonds. CH_3COO^-/$HCOO^-$ is used to model Asp/Glu side chain, CH_3OH to represent Ser/Thr side chain and $OCHNH_2$ to represent carbonyls. In total, the set contains 58 magnesium compounds.

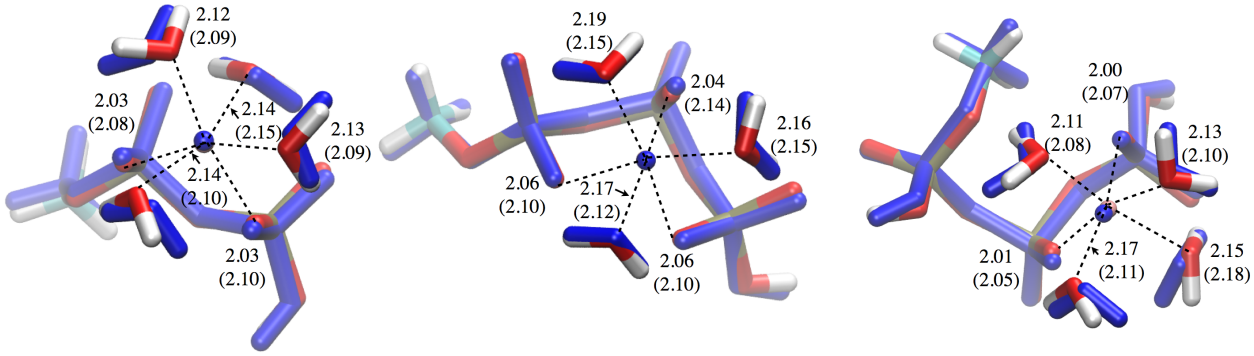
Bond distances calculated with DFTB3/3OB deviate only by 0.018 Å on average and 0.125 Å at most. The case with the largest error is the Mg-S bond in $[Mg(SH_2)_4(SH)]^-$. Besides this outlier, all other bond length errors are substantially smaller within 0.1 Å. DFTB3/3OB outperforms DFTB3/MIO in the large test set. B3LYP with the small basis set 6-31+G(d,p) deviates on average by 0.011 Å, which is comparable with DFTB3/3OB, and the maximum deviation is 0.033 Å, which is much better than 3OB. PM6 strongly underestimates Mg-O bond length by 0.04 Å. For the angle, the MADs are 3.1, 4.5, 0.7 and 11.2° respectively for DFTB3/3OB, DFTB3/MIO, B3LYP/6-31+G(d,p) and PM6. In the O-Mg-O angles calculated with DFTB3/3OB, about 6% out of 505 angles deviate more than 10°; the large errors stem mainly from erroneous orientation of hydroxide.

ATP and Mg complexes in the aqueous solution have been studied extensively by spectroscopic methods, but it is unclear from experiments which of the three phosphate groups (α, β and γ) are involved in the Mg^{2+} coordination. NMR studies have proposed several models for the Mg·ATP complex: β monodentate,(127) β,γ-bidentate,(128) α,β,γ-tridentate(129, 130) and a mixture of α,β-, β,γ- and α,γ- bidentate.(131) Raman and infrared spectra suggests a mixture of β,γ- bidentate and α,β,γ- tridentate.(132) Due to the relatively high affinity of ATP/ADP to Mg^{2+} and the importance of this metal chelates in ATP-binding enzyme, we have constructed model molecules involving Mg and phosphate/model AT(D)P as an additional test (Fig. 3.1). Two stable binding modes are tested for Mg·ATP·water complex, one with Mg coordinating β- and γ- phosphates and the other one with Mg coordinating α-, β- and γ phosphates(133). Geometries are described very well by DFTB3/3OB. The Mg-O bond lengths agree on average within 0.04 Å with B3LYP/aug-cc-pVTZ calculations.

Table 3.6: Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Magnesium Large Test Set

property[a]	N [b]	B3LYP[c]	PM6	DFTB3/MIO	DFTB3/3OB
r (Å)	329	0.011	0.063	0.041	0.018
r_{max} (Å)		0.033	1.414	0.481	0.125
a (°)	658	0.7	11.2	4.5	3.1
a_{max} (°)		17.4	64.0	37.5	30.8

[a] max stands for maximum absolute deviation. [b] number of comparisons. [c] basis set 6-31+G(d,p).



(a) $Mg(ADP)(H_2O)_4$ (b) $Mg(ATP)(H_2O)_3^-$ (c) $Mg(ATP)(H_2O)_4$

Figure 3.1: Optimized structures of Mg·AT(D)P·water model molecules at different QM levels. Distance are given in angstroms. The numbers without parentheses were obtained at the B3LYP aug-cc-pVTZ level; those in the parentheses are DFTB3 values. The B3LYP optimized structures are depicted in colors whereas DFTB3 optimized structures are colored in blue.

The O-Mg-O, O-P-O, and P-O-P angles differ on average 5, 1, and 7° respectively.

Bond dissociation energies are compared in Table 3.7. One neutral oxygen ligand is removed from the magnesium complexes selected from large test set. Because the reactions are chosen somewhat arbitrarily, the statistics given in Table 3.7 and Table 3.8 have to be considered with care. MADs are reasonably small for all semiempirical methods; the MAD for DFTB3/3OB is 4.6 kcal/mol, even smaller than B3LYP/aug-cc-pVTZ with 5.3 kcal/mol. Interestingly, DFTB3/3OB and B3LYP turn out to systematically underestimate the dissociation energy of oxygen ligand; the errors seem to inherit from the sBDE reactions where one water molecule is taken away from $[Mg(H_2O)_5]^{2+}$ or $[Mg(H_2O)_6]^{2+}$. Further fitting U^d or ε_p cannot improve these reaction energies significantly. To demonstrate the good performance of DFTB3 for structural properties, we

Table 3.7: Bond Dissociation Energies of Magnesium Containing Large Molecules Compared to G3B3[a]

Reaction			G3B3	B3LYP[b]	PM6	DFTB3/MIO	DFTB3/3OB	G3B3//DFTB3[c]
$[Mg(H_2O)_3meoh]^{2+}$	→	$[Mg(H_2O)_3]^{2+}$ + meoh	56.2	-3.8	-10.0	-2.6	-3.2	+0.7
$[Mg(H_2O)_4meoh]^{2+}$	→	$[Mg(H_2O)_4]^{2+}$ + meoh	40.5	-6.3	-5.4	-5.9	-7.7	-0.7
$[Mg(H_2O)_5meoh]^{2+}$	→	$[Mg(H_2O)_5]^{2+}$ + meoh	37.2	-7.1	-7.6	-5.5	-7.2	+0.0
$[Mg(H_2O)_3(OCHNH_2)]^{2+}$	→	$[Mg(H_2O)_3]^{2+}$ + $OCHNH_2$	76.8	-3.3	-13.0	-4.7	-2.9	+0.3
$[Mg(H_2O)_4(OCHNH_2)]^{2+}$	→	$[Mg(H_2O)_4]^{2+}$ + $OCHNH_2$	57.7	-5.5	-8.1	-6.6	-6.6	-1.2
$[Mg(H_2O)_5(OCHNH_2)]^{2+}$	→	$[Mg(H_2O)_5]^{2+}$ + $OCHNH_2$	50.7	-6.2	-6.4	-5.3	-5.8	-0.2
$[Mg(H_2O)_5ach]^{2+}$	→	$[Mg(H_2O)_5]^{2+}$ + ach	40.3	-5.9	-4.7	-3.7	-3.6	-0.3
$[Mg(H_2O)_4meoh]^{2+}$	→	$[Mg(H_2O)_3meoh]^{2+}$ + H_2O	33.4	-5.4	-6.4	-6.6	-6.4	-0.2
$[Mg(H_2O)_5meoh]^{2+}$	→	$[Mg(H_2O)_4meoh]^{2+}$ + H_2O	31.5	-6.0	-9.7	-6.4	-6.0	+0.3
$[Mg(H_2O)_4(OH)meoh]^{+}$	→	$[Mg(H_2O)_3(OH)(meoh)]^{+}$ + H_2O	18.9	-5.2	-0.5	-1.3	-1.4	+1.8
$[Mg(H_2O)_4(OCHNH_2)]^{2+}$	→	$[Mg(H_2O)_3(OCHNH_2)]^{2+}$ + H_2O	29.9	-5.1	-6.0	-5.2	-5.5	-0.2
$[Mg(H_2O)_5(OCHNH_2)]^{2+}$	→	$[Mg(H_2O)_4(OCHNH_2)]^{2+}$ + H_2O	27.8	-5.8	-5.8	-5.4	-5.7	+0.7
$[Mg(H_2O)_4(OH)(OCHNH_2)]^{+}$	→	$[Mg(H_2O)_3(OH)(OCHNH_2)]^{+}$ + H_2O	20.5	-6.1	-2.8	-3.0	-5.1	-0.5
$[Mg(H_2O)_2(meoh)(OCHNH_2)]^{2+}$	→	$[Mg(H_2O)(meoh)(OCHNH_2)]^{2+}$ + H_2O	41.9	-3.3	-10.3	-2.9	-3.0	+0.0
$[Mg(H_2O)(OH)(meoh)(OCHNH_2)]^{+}$	→	$[Mg(OH)(meoh)(OCHNH_2)]^{+}$ + H_2O	29.5	-3.3	-7.5	-3.8	-4.4	+0.5
$[Mg(H_2O)_4(meoh)(OCHNH_2)]^{2+}$	→	$[Mg(H_2O)_3(meoh)(OCHNH_2)]^{2+}$ + H_2O	26.2	-6.0	-7.2	-4.9	-5.2	+0.6
$[Mg(H_2O)_3(OH)(meoh)(OCHNH_2)]^{+}$	→	$[Mg(H_2O)_2(OH)(meoh)(OCHNH_2)]^{+}$ + H_2O	20.7	-6.8	-3.9	-4.5	-6.5	-2.0
$[Mg(H_2O)_4(OH)meoh]^{+}$	→	$[Mg(H_2O)_4(OH)]^{+}$ + meoh	22.5	-6.8	+1.8	+0.7	-1.9	+2.2
$[Mg(H_2O)_3(OH)(meoh)(OCHNH_2)]^{+}$	→	$[Mg(H_2O)_3(OH)(OCHNH_2)]^{+}$ + meoh	23.6	-7.9	-2.4	-2.0	-6.4	-2.1
$[Mg(H_2O)_4(OH)(OCHNH_2)]^{+}$	→	$[Mg(H_2O)_4(OH)]^{+}$ + $OCHNH_2$	33.5	-6.9	-0.9	-1.9	-4.8	-0.1
$[Mg(H_2O)_3(OH)(meoh)(OCHNH_2)]^{+}$	→	$[Mg(H_2O)_3(OH)(meoh)]^{+}$ + $OCHNH_2$	32.9	-6.9	-2.6	-2.8	-5.5	-2.1
$[Mg(H_2O)_4ac]^{+}$	→	$[Mg(H_2O)_3ac]^{+}$ + H_2O	21.9	-5.5	-2.4	-1.1	-4.1	+0.3
$[Mg(H_2O)_5ac]^{+}$	→	$[Mg(H_2O)_4ac]^{+}$ + H_2O	17.7	-2.9	-1.5	+2.0	-2.6	-0.2
$[Mg(H_2O)_4(OH)ac]$	→	$[Mg(H_2O)_3(OH)ac]$ + H_2O	17.3	-3.9	-5.0	+2.0	-3.0	-0.7
$[Mg(H_2O)_2(ac)_2]$	→	$[Mg(H_2O)(ac)_2]$ + H_2O	14.3	-5.4	+0.2	-0.8	-1.7	+2.1
$[Mg(H_2O)_4(ac)_2]$	→	$[Mg(H_2O)_3(ac)_2]$ + H_2O	13.2	-1.5	-0.1	+0.4	-4.1	+0.6
MAD				5.3	5.1	3.5	4.6	0.8
MAX				7.9	13.0	6.6	7.7	2.2

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. meoh stands for CH_3OH, ac stands for CH_3COO^- and ach represents for CH_3COOH. [b] Basis set aug-cc-pVTZ. [c] G3B3 single point calculations on top of DFTB3/3OB geometries.

have carried out single point G3B3 energies at DFTB3 geometries, the MAD is only 0.8 kca/mol.

Proton affinities are compiled in Table 3.8. The MAD for DFTB3/3OB is 2.6 kcal/mol, similar to B3LYP/aug-cc-pVTZ with 1.9 kcal/mol. Despite the structural difference for the deprotonated structures mentioned above, the G3B3 single-point calculations at DFTB3/3OB geometries show excellent statistics for proton affinities, demonstrating the strength of DFTB3/3OB in reproducing B3LYP/aug-cc-pVTZ geometries or at least reproducing relative energies for multiple points on potential energy surface. Similar to findings from our previous studies(78), PM6 appear to be less reliable for proton affinities with large errors.

3.3.2 Zinc

3.3.2.1 Geometries

Typical biological ligands for Zn^{2+} include water, cysteine, histidine and glutamic/aspartic acids. Therefore, we performed calculations for a series of complex that involve the zinc ion and small molecule

Table 3.8: Gas Phase Proton Affinities of Magnesium Containing Large Molecules Compared to G3B3[a]

Reaction			G3B3	B3LYP[b]	PM6	DFTB3/MIO	DFTB3/3OB	G3B3//DFTB3[d]
$[Mg(H_2O)_4meoh]^{2+}$	$\rightarrow$	$[Mg(H_2O)_3(OH)meoh]^+ + H^+$	162.3	-1.3	-15.7	+2.1	+1.4	+0.5
$[Mg(H_2O)_5meoh]^{2+}$	$\rightarrow$	$[Mg(H_2O)_4(OH)meoh]^+ + H^+$	175.0	-2.2	-25.0	-3.1	-3.3	-1.1
$[Mg(H_2O)_4ac]^+$	$\rightarrow$	$[Mg(H_2O)_3(OH)ac] + H^+$	251.3	-2.2	-14.4	+1.1	+1.7	-1.7
$[Mg(H_2O)_4ach]^{2+}$	$\rightarrow$	$[Mg(H_2O)_4ac]^+ + H^+$	145.2	+4.8	-15.2	+7.5	+6.1	+0.1
$[Mg(H_2O)_5ac]^+$	$\rightarrow$	$[Mg(H_2O)_4(OH)ac] + H^+$	251.7	-1.2	-11.0	+1.2	+2.1	-1.2
$[Mg(H_2O)_5ach]^{2+}$	$\rightarrow$	$[Mg(H_2O)_5ac]^+ + H^+$	155.7	+1.7	-20.0	-0.2	+2.8	+0.7
$[Mg(H_2O)_4(OCHNH_2)]^{2+}$	$\rightarrow$	$[Mg(H_2O)_3(OH)(OCHNH_2)]^+ + H^+$	170.2	-1.5	-18.2	+2.2	+1.7	-0.1
$[Mg(H_2O)_5(OCHNH_2)]^{2+}$	$\rightarrow$	$[Mg(H_2O)_4(OH)(OCHNH_2)]^+ + H^+$	177.5	-1.3	-21.2	-0.2	+1.1	+1.1
$[Mg(H_2O)_2(ac)(ach)]^+$	$\rightarrow$	$[Mg(H_2O)_2(ac)_2] + H^+$	238.5	+3.8	-17.0	+8.0	+5.2	-1.7
$[Mg(H_2O)_2(meoh)(OCHNH_2)]^{2+}$	$\rightarrow$	$[Mg(H_2O)(OH)(meoh)(OCHNH_2)]^+ + H^+$	164.0	+0.0	-15.0	+2.2	+0.4	+0.0
$[Mg(H_2O)_3(meoh)(OCHNH_2)]^{2+}$	$\rightarrow$	$[Mg(H_2O)_2(OH)(meoh)(OCHNH_2)]^+ + H^+$	172.5	-1.6	-17.5	+0.7	+0.4	-0.3
$[Mg(H_2O)_4(meoh)(OCHNH_2)]^{2+}$	$\rightarrow$	$[Mg(H_2O)_3(OH)(meoh)(OCHNH_2)]^+ + H^+$	178.0	-0.8	-20.9	+0.3	+1.7	+2.3
$[Mg(H_2O)_5(phos)]^{2+}$	$\rightarrow$	$[Mg(H_2O)_5(phos)]^+ + H^+$ [c]	160.3	+2.1	-12.7	+6.4	+5.7	+0.7
MAD				1.9	17.2	2.7	2.6	0.9
MAX				4.8	25.0	8.0	6.1	2.3

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. meoh stands for CH_3OH, ac stands for CH_3COO^-, ach represents for CH_3COOH, and phos represents for phosphate. [b] Basis set aug-cc-pVTZ. [c] $[Mg(H_2O)_5(phos)]^{2+}$ has three oxygens (bonded to phosphorous) protonated and $[Mg(H_2O)_5(phos)]^+$ has two oxygens protonated. [d] G3B3 single point calculations on top of DFTB3/3OB geometries.

models for these ligands. Similar to what have been done for magnesium, we have compiled a test set in the form of $Zn^{2+}[X]_n$ with X = H_2O, NH_3, SH_2, PH_3, CO, NH=CH_2 and SH-CH_3 for n = 1-4/6 and X = OH^-, SH^-, NH_2^-, PH_2^-, NCH_2^- and SCH_3^- for n = 1-2. Here two extra ligands NH=CH_2 and SH-CH_3 (and their deprotonated form) are added to model histidine and cysteine side chains. The structural comparisons are given in Table 3.9.

The DFTB3/3OB results are tested against B3LYP/aug-cc-pVTZ for geometries and energetics. Overall good geometries are obtained from DFTB3/3OB with bond distance MAD of 0.012 Å and angle MAD of 2.3°, which are notable improvements over DFTB3/MIO. The largest deviation 0.091 Å of bond length in DFTB3/3OB is for bond in $[Zn(N=CH_2)]^+$. Distance between Zn and N is underestimated by $\sim$0.1 Å if r(ZnN) is shorter than 1.95 Å due to the imperfect Zn-N repulsive potential at the short range. Similar to cases for magnesium, angle Zn-S-H are uniformly shifted up by $\sim$10° in complexes containing ligand SH_2 or SH-CH_3 in DFTB3/3OB optimized structures. B3LYP/6-31+G(d,p) is structurally very similar to B3LYP/aug-cc-pVTZ with a distance MAD of merely 0.004 Å. Nevertheless, MP2/aug-cc-pVTZ is rather different from B3LYP/aug-cc-pVTZ with a MAD of 0.017 Å; bond distances between zinc and C, N, O, P, S are uniformly underestimated by 0.02-0.03 Å. Molecular geometries are described rather poorly by PM6 in our test set; for example, the bond lengths between zinc and phosphorous deviate on average 0.3 Å.

Table 3.9: Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Zinc Test Set

property[a]	N [b]	MP2[c]	B3LYP[d]	PM6	DFTB3/MIO[e]	DFTB3/3OB
r (Å)	163	0.017	0.004	0.065	0.027	0.012
r_{max} (Å)		0.077	0.011	0.553	0.139	0.091
a (°)	180	0.6	0.2	4.2	3.0	2.3
a_{max} (°)		4.3	4.2	64.0	16.2	14.6

[a] max stands for maximum absolute deviation. [b] number of comparisons. [c] basis set aug-cc-pVTZ. [d] basis set 6-31+G(d,p). [e] Molecules containing PH_2^- or H^- don't converge and are excluded from statistics.

3.3.2.2 Sequential Bond Dissociation Energies

Sequential bond dissociation energies for zinc species with neutral and negatively charged ligands are summarized in Table 3.10. DFTB3/3OB shows a MAD of 3.6 kcal/mol compared to B3LYP/aug-cc-pVTZ. Again, deviations close to or greater than 10 kcal/mol are found for charged ligands. The difference between small and large basis set in B3LYP is rather small, even charged ligands produce errors no larger than 2.7 kcal/mol, indicating B3LYP with small basis set may be useful in *ab initio* QM/MM calculations. There is a slightly large discrepancy between MP2 and B3LYP; it may be attributed to the differences in structures between these two methods as discussed above. PM6 shows large deviations for sBDE with a MAD of 16.8 kcal/mol. With the unphysical Zn-P geometries removed from statistics, the MAD for PM6 drops to 9.1 kcal/mol, still substantially larger than that for DFTB3. We have also performed single-point calculations of B3LYP/aug-cc-pVTZ on top of DFTB3 geometries. The results are remarkably good with a MAD of only 0.5 kcal/mol.

3.3.2.3 Proton affinities

Table 3.11 shows the protonation energies calculated for 24 biological relevant ligand model molecules obtained using B3LYP, MP2 and semiempirical methods. As expected, B3LYP/6-31+G(d,p) performs very well compared to B3LYP with large basis set. The MAD for DFTB3 is 4.2 kcal/mol, which is substantially better than DFTB3/MIO and PM6; the errors are small enough to render this approach useful. High-level single-points correction at DFTB3/3OB geometries once again delivers excellent proton affinities except the ones involving SH^- ligand whose geometries are less well described as discussed above.

Table 3.10: Sequential Bond Dissociation Energies of Zinc Containing Molecules Compared to B3LYP/aug-cc-pVTZ[a]

Molecule[b]	B3LYP[c]	MP2[c]	B3LYP[d]	PM6	DFTB3/MIO	DFTB3/3OB	B3LYP//DFTB3[e]
$[Zn(H_2O)_2]^{2+}$	89.7	-2.9	+2.0	-15.7	-5.2	+4.3	+0.2
$[Zn(H_2O)_3]^{2+}$	56.4	+1.2	+1.5	+1.4	-1.1	-2.2	-0.1
$[Zn(H_2O)_4]^{2+}$	43.0	+2.2	+1.9	+3.2	-2.0	-0.9	+0.8
$[Zn(H_2O)_5]^{2+}$	25.3	+3.2	+2.6	+1.8	-3.8	-1.7	-0.4
$[Zn(H_2O)_6]^{2+}$	23.4	+3.7	+2.3	-3.2	-3.9	-1.5	+0.3
$[Zn(CO)_2]^{2+}$	68.2	+0.4	-0.4	+28.0	-5.8	-3.7	+0.4
$[Zn(CO)_3]^{2+}$	40.3	+3.6	+0.1	+15.3	+1.2	-7.3	+0.3
$[Zn(CO)_4]^{2+}$	30.6	+4.5	+0.3	+16.9	+3.9	-4.6	-0.2
$[Zn(NH_3)_2]^{2+}$	108.3	-0.6	+1.9	-8.1	-7.8	-4.3	+0.4
$[Zn(NH_3)_3]^{2+}$	59.7	+2.9	+2.4	+8.0	+0.2	-2.6	+0.3
$[Zn(NH_3)_4]^{2+}$	43.8	+3.6	+2.0	+9.4	+4.1	+2.8	+0.1
$[Zn(NH_3)_5]^{2+}$	15.3	+3.5	+2.1	+1.7	-2.7	+3.2	-0.2
$[Zn(NH_3)_6]^{2+}$	13.1	+4.3	+1.8	+1.4	-2.3	+4.1	-0.4
$[Zn(SH_2)_2]^{2+}$	84.4	+0.0	+0.5	-5.8	-2.0	+1.0	-1.4
$[Zn(SH_2)_3]^{2+}$	42.9	+3.9	+0.5	-1.6	+3.5	+1.8	-0.1
$[Zn(SH_2)_4]^{2+}$	29.1	+5.2	+1.2	+0.4	+4.8	+4.9	+2.0
$[Zn(PH_3)_2]^{2+}$	89.4	+1.9	+1.1	+55.7	-7.1	+1.2	+0.7
$[Zn(PH_3)_3]^{2+}$	43.0	+4.9	+0.8	+51.9	+0.6	+3.7	+0.6
$[Zn(PH_3)_4]^{2+}$	30.5	+7.0	+0.9	+65.9	+3.0	+5.5	+0.5
$[Zn(NH{=}CH_2)_2]^{2+}$	111.9	+0.5	+0.9	-2.2	-1.7	+0.1	+0.1
$[Zn(NH{=}CH_2)_3]^{2+}$	60.1	+4.4	+1.2	+10.5	+1.3	-1.5	-0.4
$[Zn(NH{=}CH_2)_4]^{2+}$	43.5	+5.9	+1.1	+9.8	+3.7	+2.8	+0.4
$[Zn(SH\text{-}CH_3)_2]^{2+}$	92.2	+2.1	+0.1	-8.2	-5.8	-2.8	-0.3
$[Zn(SH\text{-}CH_3)_3]^{2+}$	46.0	+6.4	+0.5	-1.2	+0.5	-0.4	+0.4
$[Zn(SH\text{-}CH_3)_4]^{2+}$	31.1	+8.1	+1.1	-0.1	+2.3	+2.9	+0.3
$[Zn(OH)_2]$	253.8	+1.9	+2.7	+12.0	-15.1	-3.4	+0.5
$[Zn(NH_2)_2]$	248.6	+4.9	+2.6	+15.9	+11.5	-4.5	+3.1
$[Zn(SH)_2]$	213.2	+6.1	+0.2	-22.1	-6.7	-9.1	-0.2
$[Zn(PH_2)_2]$	199.3	+7.1	+0.8	+108.8	-[f]	-11.7	+0.9
$[Zn(OCH_3)_2]$	235.4	+7.6	+2.2	+6.9	-10.0	-0.5	+0.2
$[Zn(NCH_2)_2]$	220.6	+10.9	+0.6	+25.7	+13.8	+8.0	+1.1
$[Zn(SCH_3)_2]$	207.4	+8.2	+0.5	-18.7	-7.8	-7.5	+0.0
MAD		4.2	1.3	16.8	4.7	3.6	0.5
MAX		10.9	2.7	108.8	15.1	11.7	3.1

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] The non-dissociated is listed. [c] Basis set aug-cc-pVTZ. [d] Basis set 6-31+G(d,p). [e] B3LYP/aug-cc-pVTZ single point calculations on top of DFTB3/3OB geometries. [f] Molecule doesn't converge and is excluded from the statistics.

Table 3.11: Gas Phase Proton Affinities of Zinc Containing Molecules Compared to B3LYP/aug-cc-pVTZ[a]

Molecule[b]	B3LYP[c]	MP2[c]	B3LYP[d]	PM6	DFTB3/MIO	DFTB/3OB	B3LYP//DFTB3[e]
$[Zn(H_2O)_1]^{2+}$	61.7	+5.4	-0.6	+7.8	-1.2	-7.7	+0.8
$[Zn(H_2O)_2]^{2+}$	94.2	+0.6	-0.3	-2.5	-1.1	-2.8	+0.5
$[Zn(H_2O)_3]^{2+}$	121.3	-0.7	-0.5	-9.6	+1.9	-1.2	+0.1
$[Zn(H_2O)_4]^{2+}$	141.5	-1.6	-0.7	-16.9	+2.3	-1.3	+0.7
$[Zn(NH_3)_1]^{2+}$	86.9	+4.1	+0.1	+6.5	-13.6	-10.8	+1.8
$[Zn(NH_3)_2]^{2+}$	127.3	-0.8	+0.7	-2.9	-18.5	-10.7	+0.6
$[Zn(NH_3)_3]^{2+}$	155.0	-1.7	+1.0	-9.7	-18.0	-10.6	+0.2
$[Zn(NH_3)_4]^{2+}$	178.1	-2.2	+1.0	-4.8	-10.3	-8.4	+0.7
$[Zn(SH_2)_1]^{2+}$	56.6	+1.6	-1.7	+16.1	-13.0	-5.4	-4.0
$[Zn(SH_2)_2]^{2+}$	93.2	-2.0	-1.5	+12.5	-10.6	-0.2	-4.0
$[Zn(SH_2)_3]^{2+}$	115.0	-2.6	-2.0	+7.7	-9.8	-0.5	-4.0
$[Zn(SH_2)_4]^{2+}$	131.1	-3.0	-2.2	+6.0	-9.1	-1.0	-2.1
$[Zn(PH_3)_1]^{2+}$	66.9	+3.2	-0.7	+21.8	-[f]	-1.5	+0.3
$[Zn(PH_3)_2]^{2+}$	115.6	-0.2	-0.3	-5.7	-[f]	-0.9	+0.8
$[Zn(PH_3)_3]^{2+}$	140.8	-1.0	-0.1	-30.0	-[f]	-2.7	+0.7
$[Zn(PH_3)_4]^{2+}$	158.0	-1.6	+0.1	+14.3	-18.1	-3.2	+1.0
$[Zn(NH{=}CH_2)_1]^{2+}$	89.1	+6.1	-0.9	+2.3	-16.2	-2.8	+1.1
$[Zn(NH{=}CH_2)_2]^{2+}$	140.9	-0.4	-0.2	-12.6	-14.1	-2.1	+0.4
$[Zn(NH{=}CH_2)_3]^{2+}$	168.6	-1.5	+0.1	-18.1	-14.2	-2.2	+0.0
$[Zn(NH{=}CH_2)_4]^{2+}$	188.1	-2.1	-0.1	-14.3	-14.8	-1.6	+2.4
$[Zn(SH\text{-}CH_3)_1]^{2+}$	74.2	+0.3	-2.5	+12.1	-4.6	-8.3	-2.5
$[Zn(SH\text{-}CH_3)_2]^{2+}$	114.9	-3.0	-2.3	+5.6	-10.0	-4.2	-2.1
$[Zn(SH\text{-}CH_3)_3]^{2+}$	136.7	-4.1	-2.4	+1.1	-10.5	-4.8	-1.7
$[Zn(SH\text{-}CH_3)_4]^{2+}$	152.6	-6.0	-2.2	+0.3	-2.6	-5.9	-1.4
MAD		2.3	1.0	10.1	10.2	4.2	1.4
MAX		6.1	2.5	30.0	18.5	10.8	4.0

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] The protonated form is listed. [c] Basis set aug-cc-pVTZ. [d] Basis set 6-31+G(d,p). [e] B3LYP/aug-cc-pVTZ single point calculations on top of DFTB3/3OB geometries. [f] Molecule doesn't converge and is excluded from the statistics.

Table 3.12: Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Zinc Mixed Ligand Test Set

property[a]	N [b]	B3LYP[c]	PM6	DFTB3/MIO[d]	DFTB3/3OB
r (Å)	161	0.005	0.118	0.053	0.019
r_{max} (Å)		0.027	0.514	0.179	0.070
a (°)	231	0.9	6.7	3.0	2.6
a_{max} (°)		22.2	57.2	21.6	24.7

[a] max stands for maximum absolute deviation. [b] number of comparisons. [c] basis set 6-31+G(d,p). [d] Molecules containing PH_2^- don't converge and are excluded from statistics.

3.3.2.4 Mixed Ligands

One goal to develop DFTB3/3OB is to describe enzyme active sites. Zn-containing structures are mainly tetrahedrally bound in metalloenzymes. The most abundant primary ones are CCCC, zinc ion with four coordinating cysteines, followed by CCCH, CCHH, CHHH, HHHH, HHHO, HHOO, HOOO, HHHD and HHDD, where O stands for water and the remaining are 1-letter amino acids codes.(12) In this subsection, we compare geometries and energetics for 43 zinc complexes with mixed types of simple ligands (including the deprotonated form) to model representative active sites in proteins. Here, H_2O is used to represent Asp/Glu side chain, NH_3 to represent His side chain, and H_2S to represent Cys side chain. More complicated/realistic ligands are discussed in the next subsection on "large test set". Table 3.12 summarizes the geometrical properties. The MADs in comparison to B3LYP/aug-cc-pVTZ for 161 bond lengths are 0.005, 0.019, 0.053, 0.118 Å for B3LYP/6-31+G(d,p), DFTB3/3OB, DFTB3/MIO and PM6, respectively. Bond angles on average deviate by 0.9, 2.6, 3.0 and 6.7°. Again, the 3OB set is a major improvement over the MIO set in bond lengths. PM6 is not able to give reliable geometries due to a large number of unreasonable bond distances between zinc and other main group elements.

DFTB3/3OB energetics are predicted with similar accuracy as for complexes with the same type of ligand as shown in Tables 3.13 and 3.14. Once again, the reactions are collected somewhat arbitrarily so the error analysis needs to be treated with care. B3LYP/6-31+G(d,p) gives small error compared to B3LYP with large basis set, demonstrating basis set superposition error is not severe here. DFTB3 gives rather good results in comparison with the reference (MAD of ∼3 kcal/mol), especially after single-point energies using B3LYP/aug-cc-pVTZ (MAD of 0.9 kcal/mol), demonstrating that the 3OB set is transferable from one type ligand to complicated coordinating environment with multiple types of ligands.

Table 3.13: Bond Dissociation Energies of Zinc Containing Mixed Ligand Molecules Compared to B3LYP/aug-cc-pVTZ[a]

Reaction			B3LYP[b]	B3LYP[c]	PM6	DFTB3/MIO	DFTB3/3OB	B3LYP//DFTB3[d]
$[Zn(NH_3)_2(SH_2)_2]^{2+}$	→	$[Zn(NH_3)(SH_2)_2]^{2+} + NH_3$	45.6	+2.6	+0.7	+0.7	+0.2	+1.6
$[Zn(NH_3)_2(SH_2)_2]^{2+}$	→	$[Zn(NH_3)_2(SH_2)]^{2+} + SH_2$	27.9	+0.9	+9.8	+6.8	+6.2	-0.7
$[Zn(NH_3)(H_2O)_2(SH_2)]^{2+}$	→	$[Zn(H_2O)_2(SH_2)]^{2+} + NH_3$	53.7	+2.7	+1.8	-4.1	-5.1	+1.6
$[Zn(NH_3)(H_2O)_2(SH_2)]^{2+}$	→	$[Zn(NH_3)(H_2O)(SH_2)]^{2+} + H_2O$	35.1	+1.9	+0.1	+3.2	+4.5	+1.2
$[Zn(NH_3)(H_2O)_2(SH_2)]^{2+}$	→	$[Zn(NH_3)(H_2O)_2]^{2+} + SH_2$	36.0	+1.2	+29.1	+2.5	+0.9	-1.5
$[Zn(NH_3)_4(H_2O)]^{2+}$	→	$[Zn(NH_3)_3(H_2O)]^{2+} + NH_3$	31.6	+2.4	+2.6	-1.7	-5.0	+0.1
$[Zn(NH_3)_4(H_2O)]^{2+}$	→	$[Zn(NH_3)_4]^{2+} + H_2O$	19.3	+2.1	-2.5	+0.9	-0.3	+0.1
$[Zn(NH_3)_2(H_2O)_4]^{2+}$	→	$[Zn(NH_3)(H_2O)_4]^{2+} + NH_3$	28.9	+2.6	+5.8	-3.9	-3.2	+1.7
$[Zn(NH_3)_2(H_2O)_4]^{2+}$	→	$[Zn(NH_3)_2(H_2O)_3]^{2+} + H_2O$	17.5	-0.8	-2.7	-5.2	-1.4	-2.6
$[Zn(NH_3)_2(H_2O)(SH_2)]^{2+}$	→	$[Zn(NH_3)(H_2O)(SH_2)]^{2+} + NH_3$	48.6	+2.1	+2.0	-0.6	-1.5	+1.4
$[Zn(NH_3)_2(H_2O)(SH_2)]^{2+}$	→	$[Zn(NH_3)_2(SH_2)]^{2+} + H_2O$	31.9	+1.9	+1.9	+5.4	+6.5	+1.1
$[Zn(NH_3)_2(H_2O)(SH_2)]^{2+}$	→	$[Zn(NH_3)_2(H_2O)]^{2+} + SH_2$	30.7	+0.8	+22.1	+5.8	+4.6	-1.3
$[Zn(NH_3)(H_2O)(OH)(SH_2)]^{+}$	→	$[Zn(H_2O)(OH)(SH_2)]^{+} + NH_3$	31.8	+2.8	-1.4	-4.6	-3.7	+0.0
$[Zn(NH_3)(H_2O)(OH)(SH_2)]^{+}$	→	$[Zn(NH_3)(OH)(SH_2)]^{+} + H_2O$	19.8	+2.5	-2.4	-0.5	+2.1	+0.3
$[Zn(NH_3)(H_2O)(OH)(SH_2)]^{+}$	→	$[Zn(NH_3)(H_2O)(OH)]^{+} + SH_2$	14.8	+1.4	+31.6	+1.8	+1.6	-0.8
$[Zn(NH_3)(H_2O)_2(SH)]^{+}$	→	$[Zn(H_2O)_2(SH)]^{+} + NH_3$	31.1	+2.4	+4.3	-2.8	-3.0	+0.1
$[Zn(NH_3)(H_2O)_2(SH)]^{+}$	→	$[Zn(NH_3)(H_2O)(SH)]^{+} + H_2O$	18.0	+1.9	+3.9	+1.5	+3.2	+0.0
$[Zn(NH_3)_2(OH)(SH_2)]^{+}$	→	$[Zn(NH_3)_2(OH)]^{+} + SH_2$	12.2	+1.5	+23.3	+3.5	+4.0	-1.0
MAD				1.9	8.2	3.1	3.2	0.9
MAX				2.8	31.6	6.8	6.5	2.6

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] Basis set aug-cc-pVTZ. [c] Basis set 6-31+G(d,p). [d] B3LYP/aug-cc-pVTZ single point calculations on top of DFTB3/3OB geometries.

Table 3.14: Gas Phase Proton Affinities of Zinc Containing Mixed Ligand Molecules Compared to B3LYP/aug-cc-pVTZ[a]

Reaction			B3LYP[b]	B3LYP[c]	PM6	DFTB3/MIO	DFTB3/3OB	B3LYP//DFTB3[d]
$[Zn(NH_3)_2(H_2O)_2]^{2+}$	→	$[Zn(NH_3)_2(H_2O)(OH)]^{+} + H^{+}$	150.5	-0.4	-16.6	-0.3	-4.8	+1.4
$[Zn(NH_3)_2(SH_2)_2]^{2+}$	→	$[Zn(NH_3)_2(SH_2)(SH)]^{+} + H^{+}$	137.0	-1.4	+1.9	-11.0	-3.5	-1.4
$[Zn(NH_3)_3(H_2O)]^{2+}$	→	$[Zn(NH_3)_3(OH)]^{+} + H^{+}$	156.6	+0.3	-16.5	-1.5	-6.9	+0.6
$[Zn(NH_3)_3(SH_2)]^{2+}$	→	$[Zn(NH_3)_3(SH)]^{+} + H^{+}$	140.9	-1.0	-2.4	-11.3	-4.0	-0.7
$[Zn(NH_3)(H_2O)(SH_2)_2]^{2+}$	→	$[Zn(NH_3)(OH)(SH_2)_2]^{+} + H^{+}$	149.7	-0.2	-21.7	-1.7	-6.4	-0.6
$[Zn(NH_3)(H_2O)(SH_2)_2]^{2+}$	→	$[Zn(NH_3)(H_2O)(SH_2)(SH)]^{+} + H^{+}$	131.7	-1.7	+1.8	-8.7	-0.6	-1.1
$[Zn(NH_3)(H_2O)_2(SH_2)]^{2+}$	→	$[Zn(NH_3)(H_2O)(OH)(SH_2)]^{+} + H^{+}$	147.2	-0.6	-17.4	+0.0	-4.6	+0.0
$[Zn(NH_3)(H_2O)_2(SH_2)]^{2+}$	→	$[Zn(NH_3)(H_2O)_2(SH)]^{+} + H^{+}$	130.9	-1.4	-4.3	-8.0	+0.8	-0.8
$[Zn(NH_3)(SH_2)_3]^{2+}$	→	$[Zn(NH_3)(SH_2)_2(SH)]^{+} + H^{+}$	134.0	-1.9	+3.8	-10.7	-2.9	-2.4
$[Zn(NH_3)(H_2O)_3]^{2+}$	→	$[Zn(NH_3)(H_2O)_2(OH)]^{+} + H^{+}$	145.3	-0.5	-17.0	+1.5	-2.6	+1.1
$[Zn(NH_3)_2(H_2O)(SH_2)]^{2+}$	→	$[Zn(NH_3)_2(OH)(SH_2)]^{+} + H^{+}$	152.7	-0.6	-18.5	-2.0	-7.1	+0.1
$[Zn(NH_3)_2(H_2O)(SH_2)]^{2+}$	→	$[Zn(NH_3)_2(H_2O)(SH)]^{+} + H^{+}$	134.9	-1.3	-3.0	-9.5	-1.3	+0.9
MAD				0.9	10.4	5.5	3.8	0.9
MAX				1.9	21.7	11.3	7.1	2.4

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. [b] Basis set aug-cc-pVTZ. [c] Basis set 6-31+G(d,p) [d] B3LYP/aug-cc-pVTZ single point calculations on top of DFTB3/3OB geometries.

Table 3.15: Mean and Maximum Absolute Deviation of Bond Lengths and Bond Angles in Comparison to B3LYP/aug-cc-pVTZ for Zinc Large Test Set

property[a]	N [b]	B3LYP[c]	PM6	DFTB3/MIO	DFTB3/3OB
r (Å)	173	0.005	0.128	0.053	0.024
r_{max} (Å)		0.063	0.955	0.177	0.141
a (°)	236	0.6	5.3	4.0	3.3
a_{max} (°)		5.1	63.3	25.8	25.8

[a] max stands for maximum absolute deviation. [b] number of comparisons. [c] basis set 6-31+G(d,p).

3.3.2.5 Large Test Set

To further test the 3OB parameter set, we have chosen larger molecules to model biologically relevant zinc coordinating environment. In this "large test set", CH_3COO-/$HCOO^-$ is used to model Asp/Glu acids side chain, imidazole/NCH_2CH_3 to represent His and $SHCH_3$ to model Cys side chain. Cys can adopt both protonated or deprotonated form in the active site so that a random mixture of $SHCH_3$ and SCH_3^- are tested in modeling Cys containing active site. In total we have compiled 51 zinc molecules. Geometries are compared to B3LYP/aug-cc-pVTZ optimized structures. B3LYP with the small basis set deviates on average by only 0.005 Å and the maximal deviation is only 0.063 Å. The MAD for bond length from DFTB3/3OB optimized structures is 0.024 Å. MAD of bond lengths Zn-O is 0.033 Å and Zn-S is 0.027 Å; they appear to be worse than small test set but are good enough for modeling enzyme active sites. The largest deviations for bond lengths are found in Zn-O bonds in $[Zn(Im)(H_2O)(SCH_3)_2]$ and $[Zn(H_2O)(SHCH_3)(SCH_3)_2]$, which are underestimated by 0.141 Å and 0.103 Å, respectively. All others errors are smaller than 0.1 Å. The largest bond angle deviation 25.8° occurs in angle Zn-N-C in $[Zn(C_3N_2H_3)]^+$ (i.e., Zn^{2+}bound to a single deprotonated imidazole), which is not encountered often in biological systems.

Bond dissociation energies are compiled in Table 3.16. DFTB3 shows that removal of neutral imidazole may cause an error of more than 10 kcal/mol, indicating interactions between Zn-sp_2 hybridized N is problematic. Nevertheless, when calculating single-point energies using B3LYP/aug-cc-pVTZ the errors turn out to be small with the MAD of merely 0.5 kcal/mol. Proton affinities are compared in Table 3.17. DFTB3/3OB shows excellent results with a MAD of 3.3 kcal/mol, which is fairly remarkable in this set of rather complicated molecules; single point B3LYP calculations lead to an even smaller MAD of 0.6 kcal/mol.

Table 3.16: Bond Dissociation Energies of Zinc Containing Large Molecules Compared to B3LYP/aug-cc-pVTZ[a]

Reaction			B3LYP[b]	B3LYP[c]	PM6	DFTB3/MIO	DFTB3/3OB	B3LYP//DFTB3[d]
$[Zn(S(CH_3)_2)_2]^{2+}$	→	$[Zn(S(CH_3)_2)]^{2+} + S(CH_3)_2$	96.1	+0.2	-9.0	-9.3	-5.0	+0.4
$[Zn(S(CH_3)_2)_3]^{2+}$	→	$[Zn(S(CH_3)_2)_2]^{2+} + S(CH_3)_2$	46.0	+0.7	+0.3	-0.7	-0.7	+0.4
$[Zn(N(=CH_2)(CH_3))_2]^{2+}$	→	$[Zn(N(=CH_2)(CH_3))]^{2+} + N(=CH_2)(CH_3)$	111.7	+0.6	+0.2	-0.9	+1.7	+0.2
$[Zn(N(=CH_2)(CH_3))_3]^{2+}$	→	$[Zn(N(=CH_2)(CH_3))_2]^{2+} + N(=CH_2)(CH_3)$	56.9	+1.0	+12.5	+2.9	+1.1	-0.2
$[Zn(N(=CH_2)(CH_3))_4]^{2+}$	→	$[Zn(N(=CH_2)(CH_3))_3]^{2+} + N(=CH_2)(CH_3)$	39.6	+0.9	+10.4	+5.6	+5.2	-0.3
$[Zn(Im)_2]^{2+}$	→	$[Zn(Im)]^{2+} + Im$	128.3	+0.9	-2.6	+4.3	+5.8	+0.2
$[Zn(Im)_3]^{2+}$	→	$[Zn(Im)_2]^{2+} + Im$	64.4	+1.1	+11.0	+4.4	+1.4	+0.3
$[Zn(Im)(SH\text{-}CH_3)_3]^{2+}$	→	$[Zn(SH\text{-}CH_3)_3]^{2+} + Im$	58.2	+1.9	+1.8	+14.4	+13.8	+1.3
$[Zn(Im)(SH\text{-}CH_3)_3]^{2+}$	→	$[Zn(Im)(SH\text{-}CH_3)_2]^{2+} + SH\text{-}CH_3$	27.8	+1.0	+2.8	+1.9	+2.1	+0.6
$[Zn(Im)(H_2O)_3]^{2+}$	→	$[Zn(H_2O)_3]^{2+} + Im$	83.5	+1.3	+20.2	+5.8	+4.8	+0.2
$[Zn(Im)(H_2O)_3]^{2+}$	→	$[Zn(Im)(H_2O)_2]^{2+} + H_2O$	33.3	+1.9	+2.7	-1.0	+0.3	+0.4
$[Zn(Im)_2(SH\text{-}CH_3)_2]^{2+}$	→	$[Zn(Im)_2(SH\text{-}CH_3)]^{2+} + SH\text{-}CH_3$	24.6	+0.7	+7.2	+0.6	+1.2	+0.7
$[Zn(Im)_2(H_2O)_2]^{2+}$	→	$[Zn(Im)_2(H_2O)]^{2+} + H_2O$	25.6	+1.8	+3.0	-0.2	+1.0	+0.4
$[Zn(Im)_3(SH\text{-}CH_3)]^{2+}$	→	$[Zn(Im)_3]^{2+} + SH\text{-}CH_3$	20.7	+0.7	+13.2	-1.1	-0.5	-0.6
$[Zn(Im)_3(H_2O)]^{2+}$	→	$[Zn(Im)_3]^{2+} + H_2O$	20.2	+1.8	+3.2	+0.4	+2.3	+0.8
$[Zn(Im)(H_2O)(SH\text{-}CH_3)_2]^{2+}$	→	$[Zn(H_2O)(SH\text{-}CH_3)_2]^{2+} + Im$	64.4	+1.5	+1.2	+12.4	+11.8	+0.8
$[Zn(Im)(H_2O)(SH\text{-}CH_3)_2]^{2+}$	→	$[Zn(Im)(SH\text{-}CH_3)_2]^{2+} + H_2O$	25.0	+2.4	+3.3	+3.0	+4.4	+0.7
$[Zn(Im)(H_2O)(SH\text{-}CH_3)_2]^{2+}$	→	$[Zn(Im)(H_2O)(SH\text{-}CH_3)]^{2+} + SH\text{-}CH_3$	32.7	+1.0	+9.3	-0.4	-0.4	+0.0
$[Zn(H_2O)(SH\text{-}CH_3)_3]^{2+}$	→	$[Zn(SH\text{-}CH_3)_3]^{2+} + H_2O$	27.0	+2.6	+3.6	+4.5	+5.7	+0.8
$[Zn(H_2O)(SH\text{-}CH_3)_3]^{2+}$	→	$[Zn(H_2O)(SH\text{-}CH_3)_2]^{2+} + SH\text{-}CH_3$	36.0	+0.8	+2.5	+1.3	+1.4	+0.2
$[Zn(Im)_2(H_2O)(ac)]^{+}$	→	$[Zn(Im)_2(ac)]^{+} + H_2O$	11.4	+2.2	+2.9	-0.4	+0.7	-0.4
MAD				1.3	5.9	3.6	3.4	0.5
MAX				2.6	20.2	14.4	13.8	1.3

[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. Im stands for imidazole, ac stands for CH_3COO^- and ach represents for CH_3COOH. [b] Basis set aug-cc-pVTZ. [c] Basis set 6-31+G(d,p). [d] B3LYP/aug-cc-pVTZ single point calculations on top of DFTB3/3OB geometries.

3.3.3 Di-Nuclear Mg/Zn Active Sites

An important catalytic motif in metalloenzymes is the bi-metallic (or di-nuclear) Mg/Zn site(134, 135). In these bi-metallic motif, metal ions are commonly coordinated through a monoatomic bridging ligand like water/hydroxide or polyatomic group like carboxylate in a bidentate fashion. Two-metal-ion mechanism(5) was proposed for this type of motif, where one metal ion actively participates in facilitating the dissociation of the leaving group and stabilizing the pentavalent intermediate, while the second metal ion holds an important structural role during catalysis. The advantage of bi-metallic centers is the charge delocalization so as to lower the barrier of binding large substrates and generating a nucleophile.(136) To assess DFTB3/3OB for these important catalytic motifs, we have constructed four model compounds based on bi-metallic Mg/Zn centers in enzymes (Fig. 3.2). In our models, certain ligands are protonated to prevent a high net charge in the gas phase.

Fig. 3.2 (a) is a simplified active site structure taken from pyruvate kinase crystal structure (PDB ID 1A49(137)). Pyruvate kinase is an enzyme of the glycolytic pathway and phosphorylates ADP to generate ATP and pyruvate. This enzyme can also utilize other divalent metal ions like Mn^{2+}, Ni^{2+} and Zn^{2+}. Fig.

Table 3.17: Gas Phase Proton Affinities of Zinc Containing Large Molecules Compared to B3LYP/aug-cc-pVTZ[a]

Reaction			B3LYP[b]	B3LYP[c]	PM6	DFTB3/MIO	DFTB3/3OB	B3LYP//DFTB3[d]
$[Zn(Im)(SH\text{-}CH_3)(S\text{-}CH_3)_2]$	→	$[Zn(Im)(S\text{-}CH_3)_3]^- + H^+$	331.4	-1.0	-11.2	-2.0	+5.2	+0.0
$[Zn(Im)(SH\text{-}CH_3)_2(S\text{-}CH_3)]^+$	→	$[Zn(Im)(SH\text{-}CH_3)(S\text{-}CH_3)_2] + H^+$	243.8	-1.6	-4.9	-8.1	+0.8	-0.6
$[Zn(Im)(SH\text{-}CH_3)_3]^{2+}$	→	$[Zn(Im)(SH\text{-}CH_3)_2(S\text{-}CH_3)]^+ + H^+$	160.6	-2.1	-1.7	-10.7	-2.5	+0.5
$[Zn(Im)_2(SH\text{-}CH_3)(S\text{-}CH_3)]^+$	→	$[Zn(Im)_2(S\text{-}CH_3)_2] + H^+$	255.7	-0.9	-10.9	-3.9	+3.7	-0.7
$[Zn(Im)_2(SH\text{-}CH_3)_2]^{2+}$	→	$[Zn(Im)_2(SH\text{-}CH_3)(S\text{-}CH_3)]^+ + H^+$	169.7	-2.0	-3.8	-8.1	+0.4	+0.7
$[Zn(Im)_2(H_2O)_2]^{2+}$	→	$[Zn(Im)_2(H_2O)(OH)]^+ + H^+$	170.8	-0.8	-14.3	+6.0	+2.9	+0.4
$[Zn(Im)_3(SH\text{-}CH_3)]^{2+}$	→	$[Zn(Im)_3(S\text{-}CH_3)]^+ + H^+$	180.3	-1.4	-5.2	-7.0	+1.4	-0.1
$[Zn(Im)_3(H_2O)]^{2+}$	→	$[Zn(Im)_3(OH)]^+ + H^+$	185.5	-0.4	-14.3	+6.1	+3.5	+0.9
$[Zn(Im)(H_2O)(SH\text{-}CH_3)(S\text{-}CH_3)]^+$	→	$[Zn(Im)(H_2O)(S\text{-}CH_3)_2] + H^+$	240.5	-1.3	-8.1	-6.3	+2.4	-2.6
$[Zn(Im)(H_2O)(SH\text{-}CH_3)_2]^{2+}$	→	$[Zn(Im)(H_2O)(SH\text{-}CH_3)(S\text{-}CH_3)]^+ + H^+$	159.1	-1.7	-2.5	-10.1	-1.7	+0.5
$[Zn(Im)(H_2O)(SH\text{-}CH_3)_2]^{2+}$	→	$[Zn(Im)(OH)(SH\text{-}CH_3)_2]^+ + H^+$	167.5	-0.4	-19.3	+0.1	-3.9	-0.3
$[Zn(Im)(OH)(SH\text{-}CH_3)_2]^+$	→	$[Zn(Im)(OH)(SH\text{-}CH_3)(S\text{-}CH_3)] + H^+$	243.4	-1.8	+4.5	-5.4	+3.3	-0.3
$[Zn(Im)(H_2O)(SH\text{-}CH_3)(S\text{-}CH_3)]^+$	→	$[Zn(Im)(OH)(SH\text{-}CH_3)(S\text{-}CH_3)] + H^+$	251.8	-0.5	-12.3	+4.7	+1.1	-1.1
$[Zn(Im)_2(H_2O)(ac)]^+$	→	$[Zn(Im)_2(OH)(ac)] + H^+$	262.7	-0.1	-9.4	+5.2	+6.9	-0.5
$[Zn(H_2O)(SH\text{-}CH_3)_2(S\text{-}CH_3)]^+$	→	$[Zn(H_2O)(SH\text{-}CH_3)(S\text{-}CH_3)_2] + H^+$	232.4	-1.6	-5.0	-8.1	-0.1	+1.0
$[Zn(H_2O)(SH\text{-}CH_3)_3]^{2+}$	→	$[Zn(H_2O)(SH\text{-}CH_3)_2(S\text{-}CH_3)]^+ + H^+$	149.2	-2.1	-0.9	-12.7	-3.9	-1.1
$[Zn(Im)_2]^{2+}$	→	$[Zn(Im)(C_3N_2H_3)]^+ + H^+$	180.9	-0.6	-11.7	-2.1	+4.9	-0.2
$[Zn(Im)_3]^{2+}$	→	$[Zn(Im)_2(C_3N_2H_3)]^+ + H^+$	201.0	-0.4	-13.7	-0.8	+4.8	-0.3
$[Zn(Im)(SH\text{-}CH_3)_3]^{2+}$	→	$[Zn(C_3N_2H_3)(SH\text{-}CH_3)_3]^+ + H^+$	195.0	-0.8	-13.1	-2.4	+3.4	-0.1
$[Zn(Im)(H_2O)_3]^{2+}$	→	$[Zn(C_3N_2H_3)(H_2O)_3]^+ + H^+$	188.8	-0.7	-16.7	-0.8	+4.8	+0.2
$[Zn(Im)_2(H_2O)_2]^{2+}$	→	$[Zn(Im)(C_3N_2H_3)(H_2O)_2]^+ + H^+$	198.4	-0.7	-15.9	-1.0	+4.8	-0.2
$[Zn(Im)_2(ac)(ach)]^+$	→	$[Zn(Im)_2(ac)_2] + H^+$	249.3	-1.1	-17.1	+1.1	+6.4	+0.0
$[Zn(Im)_3(ach)]^{2+}$	→	$[Zn(Im)_3(ac)]^+ + H^+$	178.8	-1.7	-15.3	-0.1	+2.7	-0.4
MAD				1.1	10.1	4.9	3.3	0.6
MAX				2.1	19.3	12.7	6.9	2.6

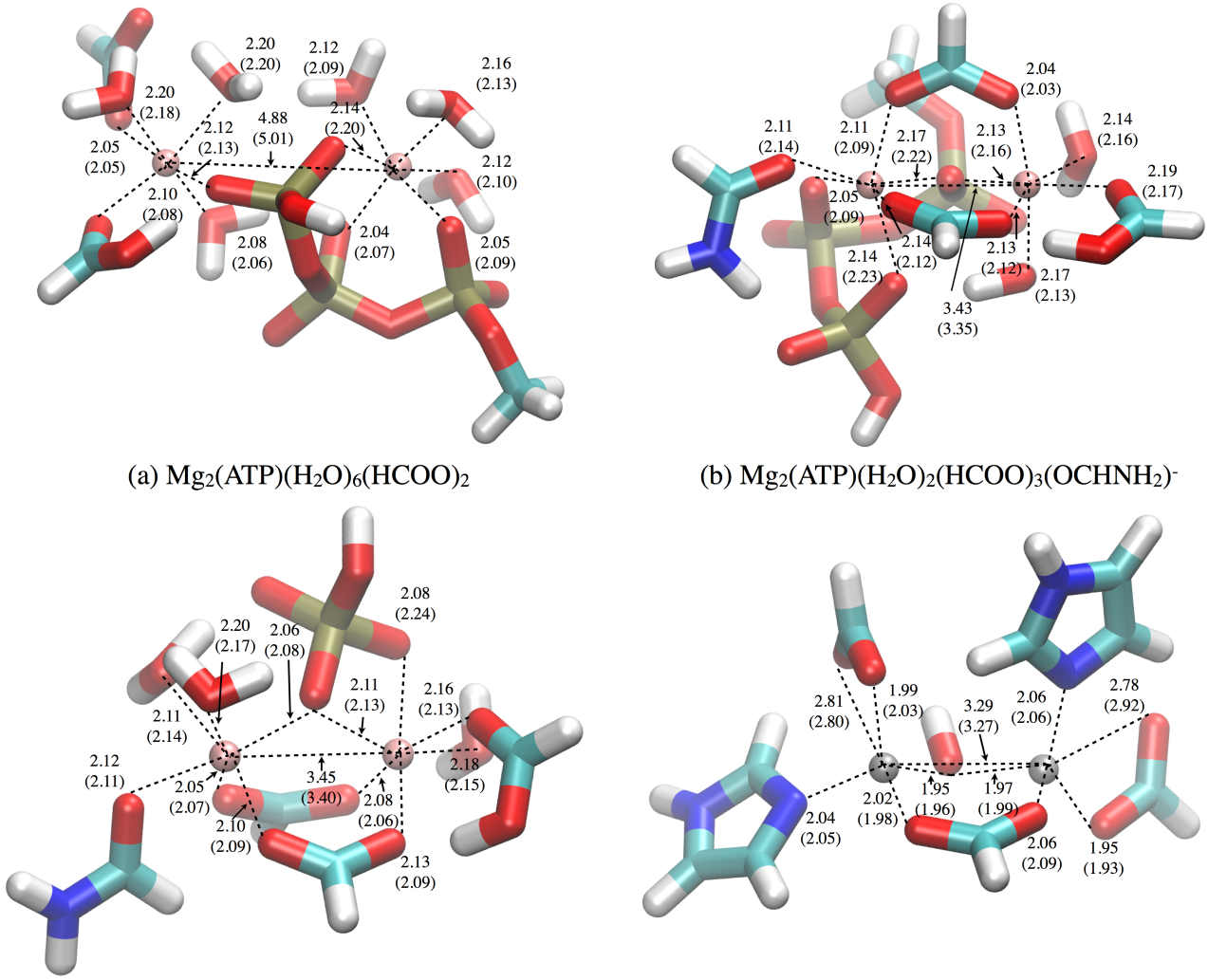
[a] Energies except PM6 are calculated at 0 K excluding zero point energy and thermal corrections. All numbers are given in kcal/mol. Im stands for imidazole, $C_3N_2H_3$ is the deprotonated imidazole ligand, ac stands for CH_3COO^- and ach represents for CH_3COOH. [b] Basis set aug-cc-pVTZ. [c] Basis set 6-31+G(d,p). [d] B3LYP/aug-cc-pVTZ single point calculations on top of DFTB3/3OB geometries.

3.2 (b) is simplified from polymerase η·DNA·dATP complex crystal structure (PDB ID 4ECQ(138)) which is involved in DNA synthesis. The formation of phosphodiester bond is recently captured by time-resolved X-ray crystallography.(138) In the (a)(b) active sites, one Mg^{2+} is tridentated to $\alpha-,\beta-,\gamma-$ oxygens in ATP. Fig. 3.2 (c) is modified from type II topoisomerase (topoII)-DNA cleavage complex crystal structure (PDB ID 3L4K(139)). This enzyme cleaves and ligates DNA using tyrosine as the nucleophilic agent. Following the nucleophilic attack towards DNA backbones, a covalent phosphotyrosyl bond that links topoII and partial DNA is formed. It was long believed catalytic reaction could be aided via a single Mg^{2+}. More recently, it was found that the canonical two-metal-ion mechanism may be possible.(139, 140) Therefore, two Mg^{2+} ions are modeled in the reactant state. Fig. 3.2 (d) is the active site in Aeromonas Protelytica Aminopeptidase (AAP) (PDB ID 1AMP(141)), which is a prototypical member of the bi-zinc enzyme family.

As shown in Fig. 3.2, DFTB3 is tested against B3LYP/6-31+G(d,p) optimized structures. The overall structures from DFTB3 reproduce B3LYP structures very well, particularly the metal ion-ligand bond distances. Mg-Mg/Zn-Zn distances are also excellent with the largest deviation of 0.13 Å. The monodentate or bridging binding modes of carboxylate are correctly described by the 3OB parameter set in comparison to B3LYP. These calculations demonstrate that DFTB3/3OB is useful for modeling bimetallic centers in enzymes with good accuracy.

To further benchmark DFTB3/3OB in describing transition state and reaction energy barrier, we study an active site model of AP in the gas phase. AP catalyzes the hydrolytic reactions of various phosphates via a two-step mechanism: an oxygen neocleophile first attacks the phosphorus, then a water (hydroxide) replaces the leaving group in the subsequent step that is essentially the reverse of the first step.(142) In our previous work,(48) we studied the first step for the hydrolysis of a well-studied phosphate diester $MpNPP^-$ in R166S AP and showed that the transition states are synchronous in nature. To help understand the intrinsic errors in DFTB3/3OB for the reactions of interest, we investigate the active site model in the gas phase with QM calculations. Specifically we study the same reaction for two additional phosphate diesters methyl 3-nitrophenyl phosphate ($MmNPP^-$) and methyl phenyl phosphate (MPP^-) and compare the energetics and nature of transition state to $MpNPP^-$. For simplicity, we only study the α orientation of different substrates.

As shown in Fig. 3.3, there is generally good agreement between DFTB3/3OB and B3LYP/6-31+G(d,p) results in terms of geometries; the P-O distances, Zn^{2+}-Zn^{2+} distances and coordination structures are similar



(a) $Mg_2(ATP)(H_2O)_6(HCOO)_2$

(b) $Mg_2(ATP)(H_2O)_2(HCOO)_3(OCHNH_2)^-$

(c) $Mg_2(H_2O)_3(HCOO)_3$ $(OCHNH_2)(Phos)$

(d) $Zn_2(Im)_2(OH)(HCOO)_3$

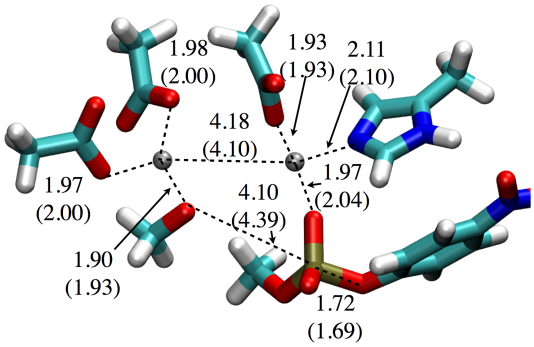
Figure 3.2: Optimized structures of di-nuclear Mg^{2+}/Zn^{2+} complexes of simplified protein active site (a) PDB ID 1A49; (b) PDB ID 4ECQ; (c) PDB ID 3L4K; (d) PDB ID 1AMP. Distance are given in angstroms. The numbers without parentheses are obtained at the B3LYP 6-31+G(d,p) level; those with parentheses are DFTB3 values.

at the two QM computational levels for both reactant and transition states. The Zn-O bond distance error is within 0.03 Å and is a major improvement over MIO parameters which systematically overestimate Zn-O bond length by 0.1 Å (MIO data not shown here). Zn^{2+}-Zn^{2+} distance is underestimated in DFTB3/3OB by $\sim$0.1 Å in the reactant and is shrunk $\sim$0.1 Å more in the transition state in comparison to B3LYP distances. In terms of the nature of the transition state, both DFTB3 and B3LYP predict a synchronous type of structure with the tightness coordinate (i.e. P-O^{ligand}+P-$O^{nucleophile}$) 4.0-4.2 Å even though individual P-O bond is overestimated or underestimated by 0.1 Å in DFTB3, indicating DFTB3 provides a good description for the potential energy surface around the transition state for the three substrates.

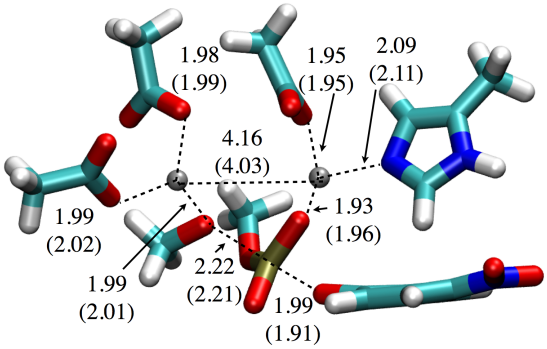
Regarding the reaction energy barrier, Table 3.18 shows that the qualitative trend is consistent among various computational levels although there are substantial differences in the quantitative values. Previous calculations(48) of proton affinities for the leaving group indicate that MP2 gives results close to the experimental values, so MP2/6-311++G(d,p) single point energies is used as the reference here. Numbers span a broad range with different DFT methods without dispersion and the results become much closer after adding the D3 dispersion; an explicit consideration of dispersion is important for a large active site discussed here. Compared to MP2, DFTB3 gives reasonable reaction energy for $MpNPP^-$, but underestimates the barrier of $MmNPP^-$ by $\sim$4 kcal/mol and overestimates the barrier of MPP^- by $\sim$9 kcal/mol. However, MP2 single point energies on top of DFTB3 and B3LYP/6-31+G(d,p) geometries differ by $\sim$1 kcal/mol for all the substrates, indicating DFTB3 produces satisfactory geometries and also highlighting the importance of single-point energy correction with a reliable *ab initio* method at DFTB3 geometries. DFTB3 reaction barriers are close to B3LYP results without D3 correction; this is not surprising because DFTB3 is parameterized based on B3LYP (see Methods) and inherits some errors. M06 behaves the best among three DFT functionals as compared to MP2, especially with the D3 dispersion included. PBE systematically underestimates the energy barrier. Further optimization of the D3 dispersion model developed for DFTB3(98) will further improve DFTB3/3OB energies.

3.3.4 Mg^{2+} and Zn^{2+} Ions Solvation Structure and Relative pK_a

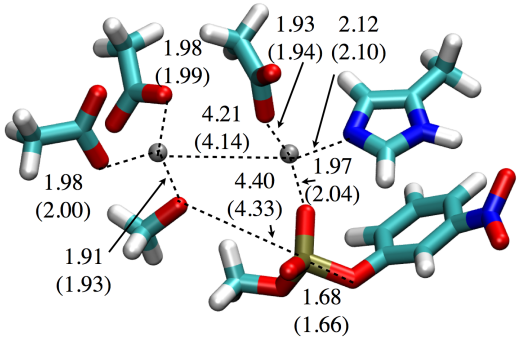
The metal ion-O radial distribution function ($g(r)$) and coordination numbers are shown in Fig. 3.4 for the two cations in solution. A well-defined first solvation shell is observed in both cases. The first peak in DFTB3/MM hydrated Mg^{2+} occurs at 2.07 Å, in good agreement with X-ray diffraction data of 2.09 Å



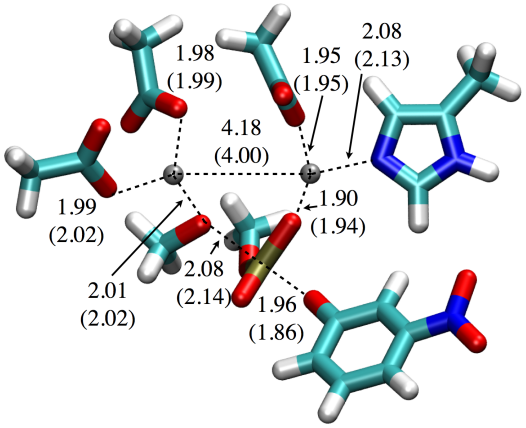
(a) MpNPP- reactant state



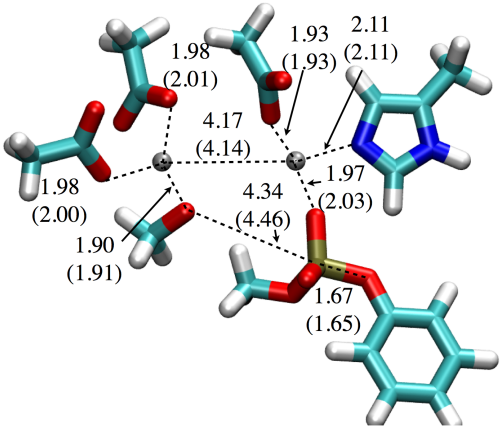
(b) MpNPP- transition state



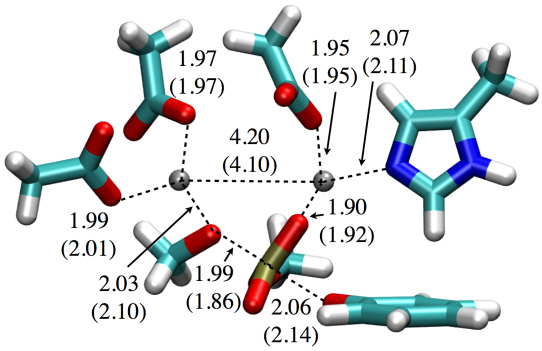
(c) MmNPP- reactant state



(d) MmNPP- transition state



(e) MPP- reactant state



(f) MPP- transition state

Figure 3.3: AP active site gas phase model with MpNPP$^-$, MmNPP$^-$ and MPP$^-$. Distance are given in angstroms. The numbers without parentheses are obtained at the B3LYP 6-31+G(d,p) level; those in the parentheses are DFTB3 values. (a) MpNPP$^-$ reactant state; (b) MpNPP$^-$ transition state; (c) MmNPP$^-$ reactant state; (d) MmNPP$^-$ transition state; (e) MPP$^-$ reactant state; (f) MPP$^-$ transition state;

Table 3.18: Calculated activation barrier (in kcal/mol) for diester hydrolysis in the active site model of AP in the gas phase[a]

substrate	DFTB3	Single points at DFTB3 geometries				Single points at B3LYP geometries			
		B3LYP	M06	PBE	MP2	B3LYP	M06	PBE	MP2
$MnNPP^-$	8.8	9.2 (4.2[b])	5.7 (5.1)	5.7 (3.1)	7.6	12.0 (6.7)	6.8 (6.1)	7.2 (4.3)	8.8
$MmNPP^-$	9.7	9.2 (9.4)	12.1 (12.4)	6.8 (7.5)	13.6	15.3 (12.0)	13.2 (12.8)	11.0 (9.3)	13.2
MPP^-	20.0	16.2 (9.1)	12.4 (10.5)	11.6 (7.1)	9.6	20.1 (11.6)	13.0 (10.8)	14.0 (8.6)	10.4

[a] The geometries are optimized either at DFTB3/3OB or B3LYP/6-31+G(d,p) level. Single point energy calculations are carried out at B3LYP, M06, PBE and MP2 method at two levels of geometries using a large basis set of 6-311++G(d,p) [b] The number with parentheses are obtained with corresponding DFT functional plus D3 dispersion.

$\pm$ 0.04 Å(143) and with previous AIMD simulations in the range of 2.08-2.13 Å.(144–148) The MM force field predicts the Mg^{2+}-O distance to be shorter than DFTB3/MM model with a sharper peak at 2.0 Å. The inner coordination sphere remains similar with gas phase structure. For the cluster $[Mg(H_2O)_6]^{2+}$ in the gas phase, Mg^{2+}-O distance is 2.10 Å in both DFTB3 and B3LYP/aug-cc-pVTZ optimized structures. For Zn^{2+}, the first maximum from DFTB3/MM is at 2.11 Å, consistent with X-ray absorption measurements 2.06 $\pm$ 0.02 Å(149) and earlier AIMD results 2.07-2.13 Å(144, 145, 150). Classical MM model gives the first peak at 2.08 Å, although the peak is higher and narrower than in the DFTB3/MM model. The small difference in the location of the first peak in DFTB3/MM may be traced back to the gas phase model: for the cluster $[Zn(H_2O)_6]^{2+}$ in the gas phase, Zn^{2+}-O is 2.16 Å in DFTB3 versus 2.12 Å in B3LYP/aug-cc-pVTZ.

The integral of the first solvation shell for both cations gives the coordination number of 6. The large QM region results predict very similar first solvation shell as with the small QM region, though they differ slightly for the second solvation shell; the large QM region leads to a broader peak for the second solvation shell than the small QM region and MM model. It is worth noting that in the large QM region simulation of Zn^{2+}, water exchange between first and second solvation shells (coordination number remains the same), whereas no exchange of water is seen in Mg^{2+} simulations. These observations are in agreement with experimental findings; for Mg^{2+}, experimental results suggest that water molecules could reside in the first shell up to a few microseconds,(151, 152) while the water residence time for Zn^{2+} is only hundreds of picoseconds.(153)

To further benchmark the energetics of DFTB3/3OB, we use the DTSC-TI approach to calculate the relative pK_a of Mg^{2+}/Zn^{2+}-bound water with the DFTB3/MM model (Fig. 3.5). Altogether six intermediates states are used to obtain the $\partial F/\partial\lambda$ values. In each window, the data collected during the first 250 ps is

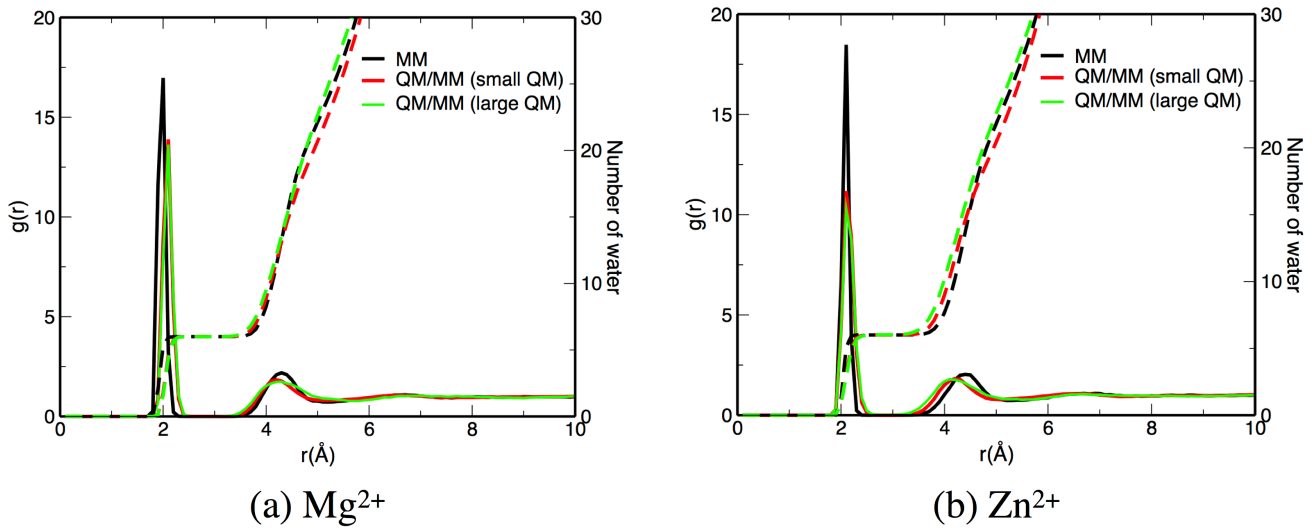


Figure 3.4: Radial distribution function of water oxygen around Mg^{2+}/Zn^{2+} in 20 Å water droplet.

discarded as part of the equilibration process, and the remaining 500 ps data is then subject to statistical analysis to obtain the proper average of free energy derivative. The statistical errors associated with the free energy are on the order of 0.2-0.5 kcal/mol. In all cases, the linear dependence of the free energy derivatives on λ holds well with a correlation coefficient (R^2) higher than 0.99. Integrating $\partial F/\partial\lambda$ over λ gives a value of 132.9 kcal/mol for Mg^{2+} with the small QM region, 124.1 kcal/mol for Zn^{2+} with the small QM region, 130.0 kcal/mol for Mg^{2+} with the large QM region and 125.5 kcal/mol for Zn^{2+} with the large QM region, leading to a ΔpK_a of 6.4 pK_a units with the small QM region and 3.3 pK_a units with the large QM region. The latter is in good agreement with the experimental value of 3.2 pK_a units (12.4 for a Mg^{2+}solution and 9.2 for hydrated Zn^{2+}).(144, 154) The better performance of the large QM region results highlights the importance of using a QM description for the second solvation shell, which allows coordination number fluctuation in the Zn^{2+} solution, especially at the deprotonated state. We also calculated absolute pK_a of Mg^{2+}/Zn^{2+}-bound water as a further benchmark. Compared to the experimental values, the closet computational estimate from current work is 14.0 for $Mg^{2+}$and 10.4 for Zn^{2+}.

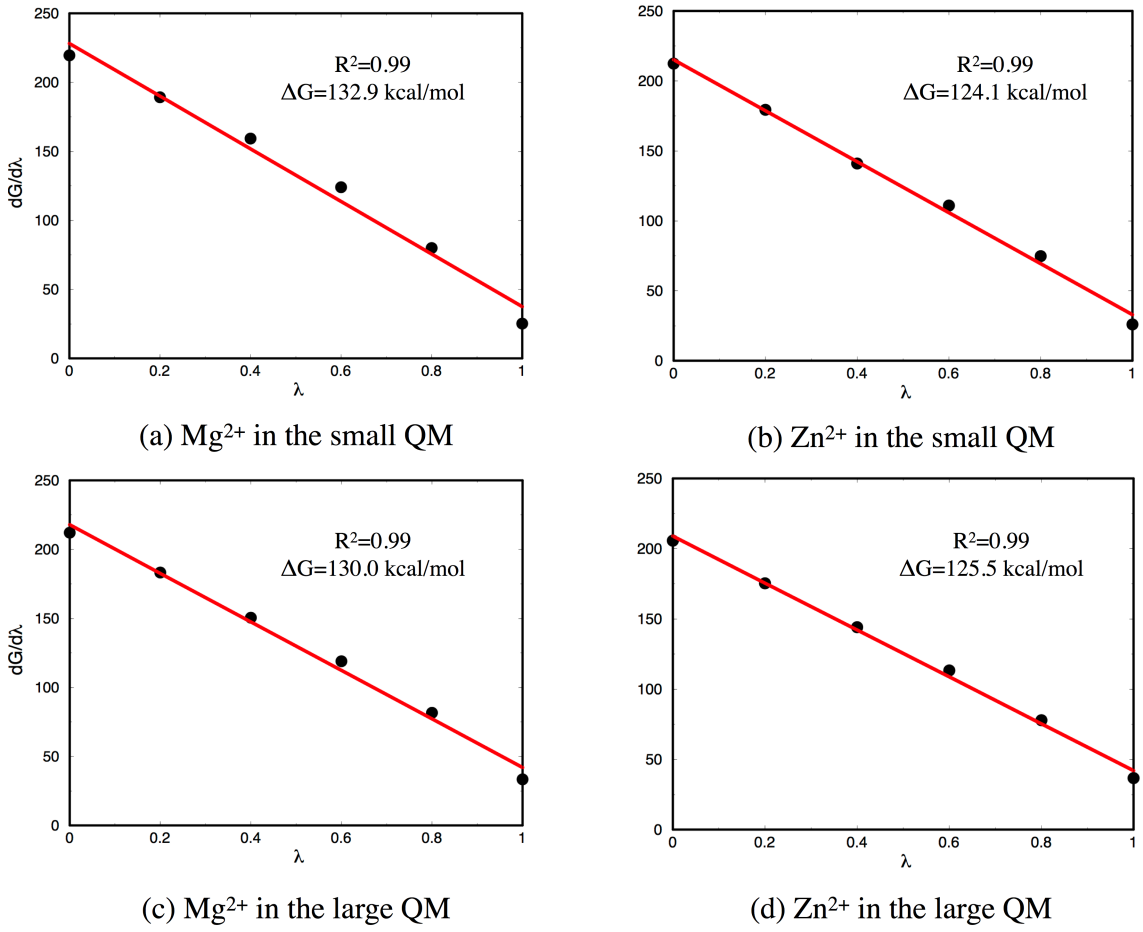


Figure 3.5: The free energy derivatives as the function of λ for (a) Mg^{2+}-water droplet simulations with the small QM region; (b) Zn^{2+}-water droplet simulations with the small QM region; (c) Mg^{2+}-water droplet simulations with the large QM region; (d) Zn^{2+}-water droplet simulations with the large QM region.

3.3.5 Structural Flexibility of the Active Site in Myosin

Myosin-II (thereafter referred to as myosin) is one of the best-characterized molecular motors, which couple large scale conformation transitions with ATP binding and hydrolysis. Despite many previous efforts from both experimental and theoretical studies,(155–159) the complex nature of "mechanochemical coupling" still remains to be well understood.

As can been seen in Figs. 3.6 (a), binding modes obtained from the DFTB3/MM simulation reproduce the key features of the crystal structure. In the discussion of our earlier work,(114) the weakened interaction between Mg^{2+}-nucleotide and Ser237 as reflected by the larger distance fluctuation was proposed to indicate the destabilization of Switch I following ATP hydrolysis. As a test of the DFTB3/3OB magnesium parameters in the enzyme, DFTB3/MM equilibrium simulations are carried out with two nucleotide chemical states (ATP or ADP·Pi) and the distances between Mg^{2+} and four non-water ligands are monitored. As shown in Figs. 3.6 (b) and (c), the distances between Mg^{2+} and ligands are reasonably stable over 1 ns; the most obvious difference observed between the two nucleotide states is the fluctuation of Mg^{2+}-O^{Ser237} distance, for which the average and fluctuation are 2.16 Å and 0.08 Å in the ATP state and 2.24 Å and 0.11 Å in the ADP·Pi state. The magnitude of fluctuation observed in current DFTB3/MM simulations is notably less compared with previous work (MM simulations and SCC-DFTBPR/MM simulation, in which the Mg-O^{Ser237} distance reached $\sim$3.5 Å), which led to coordination number shifting from 6 to 5 after ATP hydrolysis(114). To further evaluate the situation, PMFs are calculated for the distance between Mg^{2+} and O^{Ser237} in the ATP and ADP·Pi states. As seen in Fig. 3.6 (d), dissociation of Ser237 is indeed energetically more favorable in the ADP·Pi state than the ATP state by a few kcal/mol. Nevertheless, there is a significant energetic cost to fully dissociate the Ser237 side chain from the Mg^{2+}, even in the ADP·Pi state. Further experimental analysis based on infrared spectroscopy can potentially generate additional insights into this issue(160).

3.3.6 Force Fields Comparison in AP Active Site

We have benchmarked DFTB3/3OB for the AP active site with a gas phase model in earlier discussion. As an additional test, DFTB3/MM simulations for the R166S AP enzyme are carried out and the results are compared to two popular MM force fields for zinc. The comparison of equilibrated reactant structures by Coulomb(14), SLEF1(25) and DFTB3/MM are shown in Fig. 3.7, with the crystal structure with bound phosphate as a reference. Binding modes obtained from the DFTB3/MM simulation reproduce the key fea-

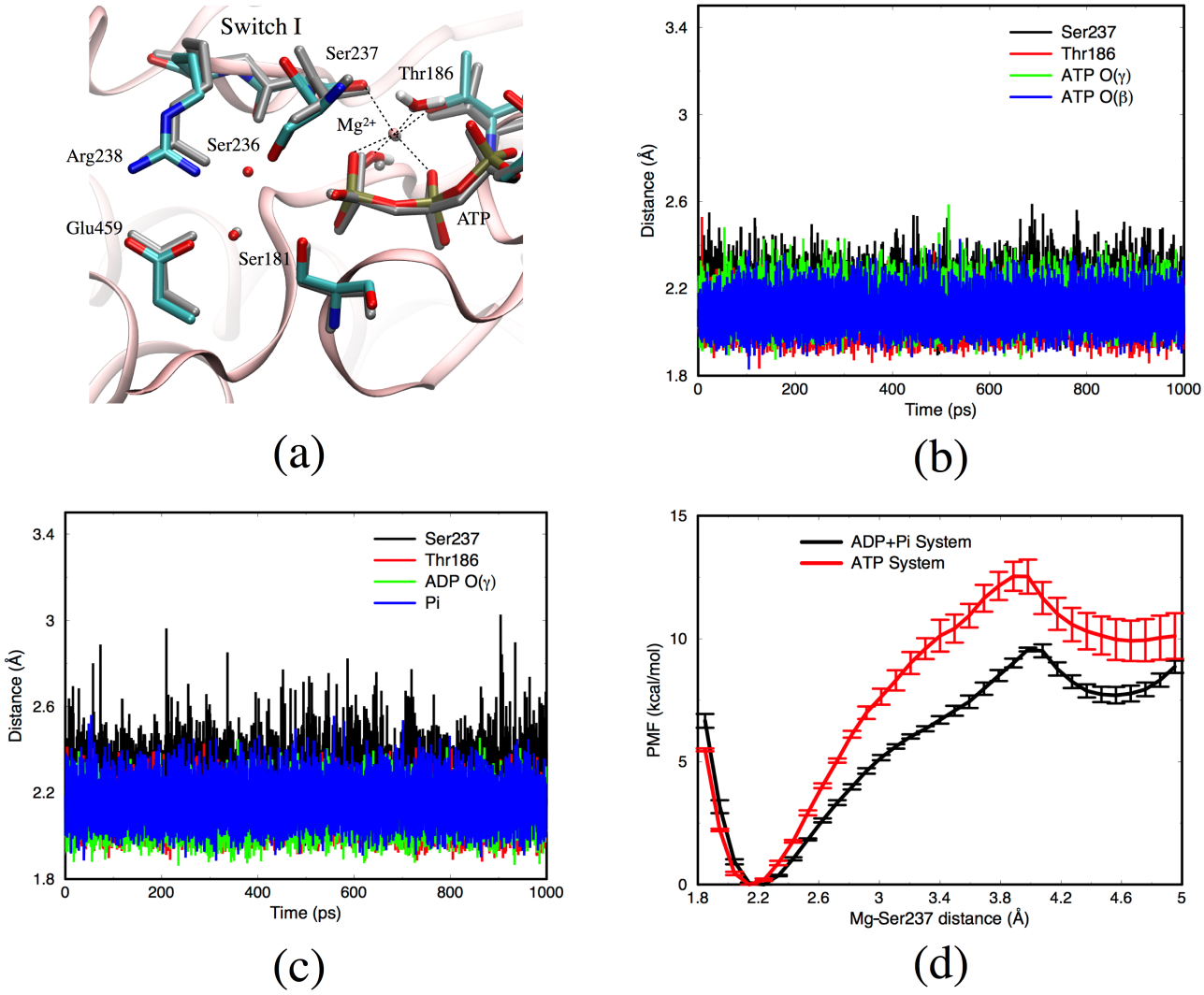


Figure 3.6: Structural properties of the Myosin active site during equilibrium QM/MM MD Simulations with different nucleotide chemical states (ATP or ADP·Pi). (a) Overlay of average DFTB3/MM structure and crystal structure (gray) of Myosin-II motor domain active site in the ATP state. The VO_4 moiety in the crystal structure (PDB code 1VOM) is replaced by a PO_3 phosphate group and a presumptive lytic water molecule. Instantaneous distances between Mg^{2+} and oxygen atoms in its four nonwater ligands at (b) ATP state; (c) ADP·Pi state. Ser237 refers to O^{Ser237} and Thr186 refers to O^{Thr186}. (d) PMF calculations for the ATP/ADP·Pi states. The reaction coordinate is the distance between Mg^{2+} and O^{Ser237}.

tures of the crystal structure and previous SCC-DFTBPR/MM simulations(48) very well (shown in Fig. 3.7 (c) and (d)). The conventional Coulomb scheme (Fig. 3.7 (a)) leads to a much shortened Zn^{2+}-Zn^{2+} distance and significant changes in the active site geometry: Ser102 and Asp327 are stabilized by two zinc ions simultaneously; Asp369 (not shown in Fig. 3.7 for clarity) is bidentate with the $Zn1^{2+}$ in the Coulomb model but monodentate in the crystal and DFTB3/MM structures. By comparison, coordination geometry is notably improved by the SLEF1 force field (Fig. 3.7 (b)), even though the Zn^{2+}-Zn^{2+} (3.7 Å) distance is still substantially shorter than that in the crystal structure (4.3 Å). It is perhaps not surprising that SLEF1 does not provide a correct distance between zinc ions as SLEF1 was not parameterized for binuclear zinc motif and designed to be compatible with the Amber99SB force field rather than with the CHARMM force field. Nevertheless, SLEF1 correctly describes the binding modes of Ser102 and Asp369 to the zinc ions; yet Asp327 still has a strong preference to bridge the two zinc cations due to the short distance between them. Moreover, an extra His372 is pulled closer to bind to $Zn2^{2+}$; this might have been caused by the stronger attractive interaction between zinc ion and nitrogen in the short range part of the SLEF1 force field. In short, DFTB3/MM produces encouraging results for modeling a challenging bimetallic zinc active site, demonstrating the greater applicability of DFTB3/MM to metalloenzymes even when only structural properties are considered.

3.4 Concluding Remarks

Magnesium and zinc play many important roles in metalloenzymes and various chemical systems. An efficient and accurate QM description for Mg/Zn is critical for the use of theoretical methods to study the reaction mechanism of these systems. In this work, we extend the parameterization of the approximate density functional tight binding theory, DFTB3, to magnesium and zinc. The parameterization is performed in a framework consistent with the DFTB3/3OB set for CHNOPS; hence this parameter set can be used to study the chemistry of Mg/Zn moieties in solution and biological systems.

Benchmark calculations in the gas phase and in the condensed phase with a QM/MM framework demonstrate that the current parameterization generally leads to reliable structures and semiquantitative energetics (with typically a MAD of $\sim$3-5 kcal/mol compared to high level *ab initio* calculations). In the gas phase, DFTB3/3OB are tested with a diverse range of molecules that reflect the typical coordination environment of magnesium and zinc in biomolecules, including di-nuclear metal sites. We focus on the geometries, ligand dissociation energies and ligand proton affinities. The results calculated at the DFTB3/3OB level are compared

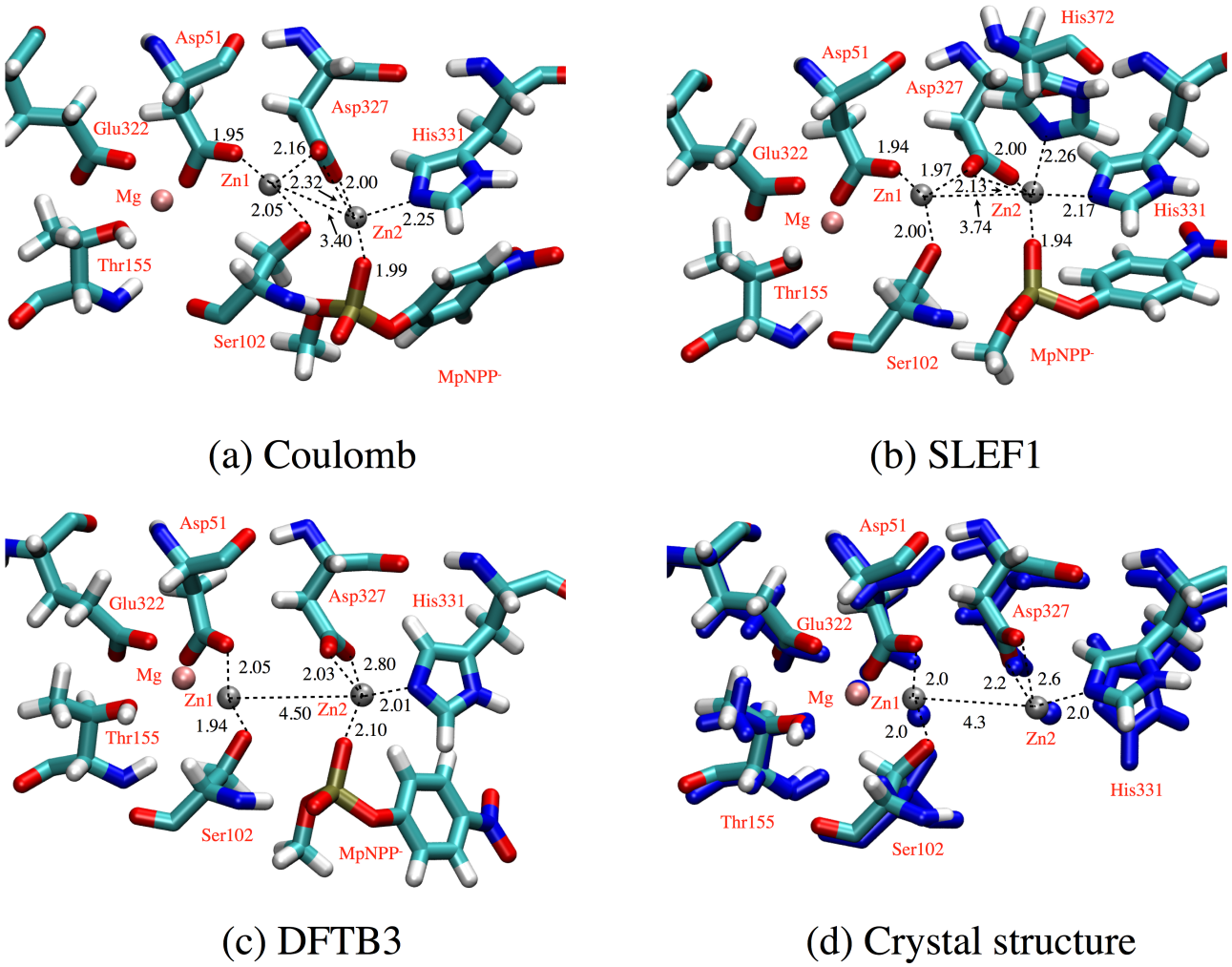


Figure 3.7: Structural properties of the AP active site during equilibrium QM/MM MD Simulations. A snapshot for the reactant state with (a) Coulomb MM force field (b) SLEF1 zinc MM force field (c) DFTB3/3OB (d) overlay of crystal and DFTB3/MM structure (blue), with average key distances in angstroms labeled. Asp369, His370 and His412 are omitted for clarity. Substrate inorganic phosphate in the crystal structure (3CMR) is not shown.

to B3LYP, *ab initio* (G3B3, MP2) and a popular semiempirical method PM6 as well as the older parameterization, MIO. In general, DFTB3/3OB shows substantial improvement over DFTB3/MIO and outperforms PM6 in many aspects within our test sets, particularly in terms of structural properties and ligand proton affinities. Compared to high-level calculations, DFTB3/3OB is very successful at predicting geometries. Large errors can still be found in certain cases, especially with charged ligands. Even for these cases, single point energy calculations with high level QM methods generally give very reliable energetics, suggesting that changes of structures during typical processes of interest (e.g., ligation dissociation or ligand deprotonation) are well captured with DFTB3/3OB.

Due to computationally efficiency and general robustness for geometry predictions, DFTB3/MM calculations are expected to be effective for studies in condensed phase systems. For example, our solution benchmark study for the magnesium, zinc ion have reproduced experimental solvation structures and relative pK_a of metal bound water, especially when both first and second solvation shells are treated as DFTB3/3OB. For the enzyme benchmark, DFTB3/3OB are successful at producing active site structures in myosin and AP relative to available crystal structures; for AP, DFTB3/3OB also leads to semi-quantitative energetics for an active site model, although our calculations using several DFT functionals and MP2 highlight the importance of careful calibration of energetics for large metal sites in enzymes, including consideration of dispersion interactions. These studies have laid the ground work for using DFTB3/MM simulations with state-of-the-art sampling techniques such as metadynamics(161) and finite-temperature string methods(162–164) to study the chemical processes in myosin and AP; high-level single point QM/MM calculations are still required for improving the quantitative nature of the free energy surfaces. Finally, we emphasize that there are still remaining limitations in DFTB3, such as the treatment of interaction between Mg/Zn and highly charged/polarizable ligands (e.g., hydroxide and SH^-), thus it is essential to continue the formal development of DFTB3, such as adding multipole terms for the charge fluctuations and formulating a better description of polarization and short-range Pauli repulsions.

References

[1] S. J. LIPPARD and J. M. BERG, *Principles of Bioinorganic Chemistry*, University Science Books, Mill Valley, CA, 1994.

[2] C. ANDREINI, I. BERTINI, G. CAVALLARO, G. L. HOLLIDAY, and J. M. THORNTON, *J. Biol. Inorg. Chem.* **13**, 1205 (2008).

[3] K. J. WALDRON, J. C. RUTHERFORD, D. FORD, and N. J. ROBINSON, *Nature* **460**, 823 (2009).

[4] C. BERGMAN, Y. KASHIWAYA, and R. L. VEECH, *J. Phys. Chem. B* **114**, 16137 (2012).

[5] T. STEITZ and J. STEITZ, *Proc. Natl. Acad. Sci. USA* **90**, 6498 (1993).

[6] G. PARKIN, *Chem. Rev.* **104**, 699 (2004).

[7] W. MARET and Y. LI, *Chem. Rev.* **109**, 4682 (2009).

[8] W. N. LIPSCOMB and N. STRATER, *Chem. Rev.* **96**, 2375 (1996).

[9] D. S. AULD, *BioMetals* **14**, 271 (2001).

[10] Y.-M. LEE and C. LIM, *J. Mol. Biol.* **379**, 545 (2008).

[11] S. C. HOOPS, K. W. ANDERSON, and K. M. MERZ, JR., *J. Am. Chem. Soc.* **113**, 8262 (1991).

[12] M. B. PETERS, Y. YANG, B. WANG, L. FÜSTI-MOLNÁR, M. N. WEAVER, and K. M. MERZ, *J. Chem. Theory Comput.* **6**, 2935 (2010).

[13] C. S. BABU and C. LIM, *J. Phys. Chem. A* **110**, 691 (2006).

[14] R. H. STOTE and M. KARPLUS, *Proteins* **23**, 12 (1995).

[15] F. DUARTE, P. BAUER, A. BARROZO, B. A. AMREIN, M. PURG, J. ÅQVIST, and S. C. L. KAMERLIN, *J. Phys. Chem. B* **118**, 4351 (2014).

[16] N. GRESH, G. A. CISNEROS, T. A. DARDEN, and J.-P. PIQUEMAL, *J. Chem. Theory Comput.* **3**, 1960 (2007).

[17] J. C. WU, J.-P. PIQUEMAL, R. CHAUDRET, P. REINHARDT, and P. REN, *J. Chem. Theory Comput.* **6**, 2059 (2010).

[18] J. ZHANG, W. YANG, J.-P. PIQUEMAL, and P. REN, *J. Chem. Theory Comput.* **8**, 1314 (2012).

[19] H. B. YU, T. W. WHITFIELD, E. HARDER, G. LAMOUREUX, I. VOROBYOV, V. M. ANISIMOV, A. D. MACKERELL JR., and B. ROUX, *J. Chem. Theory Comput.* **6**, 774 (2010).

[20] D. V. SAKHAROV and C. LIM, *J. Comp. Chem.* **30**, 191 (2008).

[21] B. DE COURCY, J.-P. PIQUEMAL, and N. GRESH, *J. Chem. Theory Comput.* **4**, 1659 (2008).

[22] N. GRESH, J.-P. PIQUEMAL, and M. KRAUSS, *J. Comput. Chem.* **26**, 1113 (2005).

[23] U. RYDE, *Biophys. J.* **77**, 2777 (1999).

[24] R. WU, P. HU, S. WANG, Z. CAO, and Y. ZHANG, *J. Chem. Theory Comput.* **6**, 337 (2010).

[25] R. WU, Z. LU, Z. CAO, and Y. ZHANG, *J. Chem. Theory Comput.* **7**, 433 (2011).

[26] N. V. PLOTNIKOV, S. C. L. KAMERLIN, and A. WARSHEL, *J. Phys. Chem. B* **115**, 7950 (2011).

[27] H. HU and W. YANG, *Annu. Rev. Phys. Chem.* **59**, 573 (2008).

[28] R. B. WU, Z. X. CAO, and Y. K. ZHANG, *Prog. Chem.* **24**, 1175 (2012).

[29] L. H. WANG, X. Y. YU, P. HU, S. BROYDE, and Y. K. ZHANG, *J. Am. Chem. Soc.* **129**, 4731 (2007).

[30] M. R. A. BLOMBERG, T. BOROWSKI, F. HIMO, R. Z. LIAO, and P. E. M. SIEGBAHN, *Chem. Rev.* **114**, 3601 (2014).

[31] S. JONAS and F. HOLLFELDER, *Pure Appl. Chem.* **81**, 731 (2009).

[32] J. K. LASSILA, J. G. ZALATAN, and D. HERSCHLAG, *Annu. Rev. Biochem.* **80**, 669 (2011).

[33] F. DUARTE, B. A. AMREIN, and S. C. L. KAMERLIN, *Phys. Chem. Chem. Phys.* **15**, 11160 (2013).

[34] M. J. S. DEWAR and W. THIEL, *J. Am. Chem. Soc.* **99**, 4899 (1977).

[35] M. J. S. DEWAR and W. THIEL, *J. Am. Chem. Soc.* **99**, 4907 (1977).

[36] W. THIEL and A. A. VOITYUK, *J. Phys. Chem.* **100**, 616 (1996).

[37] M. J. S. DEWAR, E. G. ZOEBISCH, E. F. HEALY, and J. J. P. STEWART, *J. Am. Chem. Soc.* **107**, 3902 (1985).

[38] J. J. P. STEWART, *J. Comput. Chem.* **10**, 209 (1989).

[39] M. C. HUTTER, J. R. REIMERS, and N. S. HUSH, *J. Phys. Chem. B* **102**, 8080 (1998).

[40] J. J. P. STEWART, *J. Mol. Model.* **10**, 155 (2004).

[41] Z. CAI, P. LOPEZ, J. R. REIMERS, Q. CUI, and M. ELSTNER, *J. Phys. Chem. A* **111**, 5743 (2007).

[42] P. IMHOF, F. NOÉ, S. FISCHER, and J. C. SMITH, *J. Chem. Theory Comput.* **2**, 1050 (2006).

[43] A. SORKIN, D. G. TRUHLAR, and E. A. AMIN, *J. Chem. Theory Comput.* **5**, 1254 (2009).

[44] M. J. S. DEWAR and K. M. MERZ, *Organometallics* **7**, 522 (1988).

[45] E. N. BROTHERS, D. SUAREZ, D. W. DEERFIELD, and K. M. MERZ, *J. Comput. Chem.* **25**, 1677 (2004).

[46] M. BRÄUER, M. KUNERT, E. DINJUS, M. KLUSSMANN, M. DÖRING, H. GÖRLS, and E. ANDERS, *J. Mol. Struct. (Theochem)* **505**, 289 (2000).

[47] J. J. P. STEWART, *J. Mol. Model.* **13**, 1173 (2007).

[48] G. HOU and Q. CUI, *J. Am. Chem. Soc.* **134**, 229 (2012).

[49] G. HOU and Q. CUI, *J. Am. Chem. Soc.* **135**, 10457 (2013).

[50] V. LÓPEZ-CANUT, M. ROCA, J. BERTRÁN, V. MOLINER, and I. TUÑÓN, *J. Am. Chem. Soc.* **132**, 6955 (2010).

[51] V. López-Canut, M. Roca, J. Bertrán, V. Moliner, and I. Tuñón, *J. Am. Chem. Soc.* **133**, 12050 (2011).

[52] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai, and G. Seifert, *Phys. Rev. B* **58**, 7260 (1998).

[53] M. Elstner, *J. Phys. Chem. A* **111**, 5614 (2007).

[54] M. Gaus, Q. Cui, and M. Elstner, *WIREs Comput. Mol. Sci.* **4**, 49 (2014).

[55] T. Kruger, M. Elstner, P. Schiffels, and T. Frauenheim, *J. Chem. Phys.* **122**, 114110 1 (2005).

[56] K. W. Sattelmeyer, J. Tirado-Rives, and W. Jorgensen, *J. Phys. Chem. A* **110**, 13551 (2006).

[57] N. Otte, M. Scholten, and W. Thiel, *J. Phys. Chem. A* **111**, 5751 (2007).

[58] T. Kubar, Z. Bodrog, M. Gaus, C. Köhler, B. Aradi, T. Frauenheim, and M. Elstner, *J. Chem. Theory Comput.* **9**, 2939 (2013).

[59] G. Zheng, H. A. Witek, P. Bobadova-Parvanova, S. Irle, D. G. Musaev, R. Prabhakar, K. Morokuma, M. Lundberg, M. Elstner, C. Köhler, and T. Frauenheim, *J. Chem. Theory Comput.* **3**, 1349 (2007).

[60] G. Dolgonos, B. Aradi, N. H. Moreira, and T. Frauenheim, *J. Chem. Theory Comput.* **6**, 266 (2010).

[61] B. Grundkötter-Stock, V. Bezugly, J. Kunstmann, G. Cuniberti, T. Frauenheim, and T. A. Niehaus, *J. Chem. Theory Comput.* **8**, 1153 (2012).

[62] M. Elstner, Q. Cui, P. Munih, E. Kaxiras, T. Frauenheim, and M. Karplus, *J. Comput. Chem.* **24**, 565 (2003).

[63] N. H. Moreira, G. Dolgonos, B. Aradi, A. L. Da Rosa, and T. Frauenheim, *J. Chem. Theory Comput.* **5**, 605 (2009).

[64] Q. Cui, M. Elstner, E. Kaxiras, T. Frauenheim, and M. Karplus, *J. Phys. Chem. B* **105**, 569 (2001).

[65] M. ELSTNER, T. FRAUENHEIM, and S. SUHAI, *J. Mol. Struct. (Theochem)* **632**, 29 (2003).

[66] D. RICCARDI, P. SCHAEFER, Y. YANG, H. YU, N. GHOSH, X. PRAT-RESINA, P. KÖNIG, G. LI, D. XU, H. GUO, M. ELSTNER, and Q. CUI, *J. Phys. Chem. B* **110**, 6458 (2006).

[67] Q. CUI, M. ELSTNER, and M. KARPLUS, *J. Phys. Chem B* **106**, 2721 (2002).

[68] Y. YANG, H. YU, and Q. CUI, *J. Mol. Biol.* **381**, 1407 (2008).

[69] D. RICCARDI, S. YANG, and Q. CUI, *Biochim. Biophys. Acta* **1804**, 342 (2010).

[70] Y. YANG, Y. P. MIAO, B. WANG, G. L. CUI, and K. M. MERZ, JR., *Biochem.* **51**, 2606 (2012).

[71] D. K. CHAKRAVORTY, B. WANG, C. W. LEE, D. P. GIEDROC, and K. M. MERZ, JR., *J. Am. Chem. Soc.* **134**, 3367 (2012).

[72] D. G. XU, Q. CUI, and H. GUO, *Int. Rev. Phys. Chem.* **33**, 1 (2014).

[73] N. H. MOREIRA, A. L. DA ROSA, and T. FRAUENHEIM, *App. Phys. Lett.* **94**, 193109 1 (2009).

[74] S. G. HOLTHAUS, S. KÖPPEN, T. FRAUENHEIM, and L. C. CIACCHI, *J. Chem. Theory Comput.* **8**, 4517 (2012).

[75] M. GAUS, Q. CUI, and M. ELSTNER, *J. Chem. Theory Comput.* **7**, 931 (2011).

[76] Y. YANG, H. YU, D. YORK, Q. CUI, and M. ELSTNER, *J. Phys. Chem. A* **111**, 10861 (2007).

[77] M. GAUS, A. GOEZ, and M. ELSTNER, *J. Chem. Theory Comput.* **9**, 338 (2013).

[78] M. GAUS, X. LU, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **10**, 1518 (2014).

[79] Q. CUI and M. ELSTNER, *Phys. Chem. Chem. Phys.* **16**, 14368 (2014).

[80] M. GAUS, C.-P. CHOU, H. WITEK, and M. ELSTNER, *J. Phys. Chem. A* **113**, 11866 (2009).

[81] J. P. PERDEW, K. BURKE, and M. ERNZERHOF, *Phys. Rev. Lett.* **77**, 3865 (1996).

[82] Z. BODROG, B. ARADI, and T. FRAUENHEIM, *J. Chem. Theory Comput.* **7**, 2654 (2011).

[83] A. D. BECKE, *Phys. Rev. A* **38**, 3098 (1988).

[84] A. D. BECKE, *J. Chem. Phys.* **98**, 5648 (1993).

[85] C. LEE, W. YANG, and R. G. PARR, *Phys. Rev. B* **37**, 785 (1988).

[86] L. A. CURTISS, K. RAGHAVACHARI, P. C. REDFERN, V. RASSOLOV, and J. A. POPLE, *J. Chem. Phys.* **109**, 7764 (1998).

[87] L. A. CURTISS, K. RAGHAVACHARI, P. C. REDFERN, and J. A. POPLE, *J. Chem. Phys.* **112**, 7374 (2000).

[88] L. A. CURTISS, P. C. REDFERN, and K. RAGHAVACHARI, *J. Chem. Phys.* **123**, 124107 1 (2005).

[89] F. CHENG, R. ZHANG, X. LUO, J. SHEN, X. LI, J. GU, W. ZHU, J. SHEN, I. SAGI, R. JI, K. CHEN, and H. JIANG, *J. Phys. Chem. B* **106**, 4552 (2002).

[90] M. N. WEAVER, K. M. MERZ, D. MA, H. J. KIM, and L. GAGLIARDI, *J. Chem. Theory Comput.* **9**, 5277 (2013).

[91] F. A. HAMPRECHT, A. COHEN, D. J. TOZER, and N. C. HANDY, *J. Chem. Phys.* **109**, 6264 (1998).

[92] P. J. WILSON, T. J. BRADLEY, and D. J. TOZER, *J. Chem. Phys.* **115**, 9233 (2001).

[93] A. D. BECKE, *J. Chem. Phys.* **107**, 8554 (1997).

[94] H. L. SCHMIDER and A. D. BECKE, *J. Chem. Phys.* **108**, 9624 (1998).

[95] S. M. TEKARLI, M. L. DRUMMOND, T. G. WILLIAMS, T. R. CUNDARI, and A. K. WILSON, *J. Phys. Chem. A* **113**, 8607 (2009).

[96] M. ELSTNER, P. HOBZA, T. FRAUENHEIM, S. SUHAI, and E. KAXIRAS, *J. Chem. Phys.* **114**, 5149 (2001).

[97] R. SEDLAK, T. JANOWSKI, M. PITONAK, J. REZAC, P. PULAY, and P. HOBZA, *J. Chem. Theory Comput.* **9**, 3364 (2013).

[98] T. RISTHAUS and S. GRIMME, *J. Chem. Theory Comput.* **9**, 1580 (2013).

[99] B. R. BROOKS, C. L. BROOKS, A. D. MACKERELL, L. NILSSON, R. J. PETRELLA, B. ROUX, Y. WON, G. ARCHONTIS, C. BARTELS, S. BORESCH, A. CAFLISCH, L. CAVES, Q. CUI, A. R. DINNER, M. FEIG, S. FISCHER, J. GAO, M. HODOSCEK, W. IM, K. KUCZERA, T. LAZARIDIS, J. MA, V. OVCHINNIKOV, E. PACI, R. W. PASTOR, C. B. POST, J. Z. PU, M. SCHAEFER, B. TIDOR, R. M. VENABLE, H. L. WOODCOCK, X. WU, W. YANG, D. M. YORK, and M. KARPLUS, *J. Comput. Chem.* **30**, 1545 (2009).

[100] W. IM, S. BERNÈCHE, and B. ROUX, *J. Chem. Phys.* **114**, 2924 (2001).

[101] P. SCHAEFER, D. RICCARDI, and Q. CUI, *J. Chem. Phys.* **123**, 014905 1 (2005).

[102] R. H. STOTE, D. J. STATES, and M. KARPLUS, *J. Chim. Phys.* **88**, 2419 (1991).

[103] W. L. JORGENSEN, J. CHANDRASEKHAR, J. D. MADURA, R. W. IMPEY, and M. L. KLEIN, *J. Chem. Phys.* **79**, 926 (1983).

[104] A. D. MACKERELL, D. BASHFORD, BELLOTT, R. L. DUNBRACK, J. D. EVANSECK, M. J. FIELD, S. FISCHER, J. GAO, H. GUO, S. HA, D. JOSEPH-MCCARTHY, L. KUCHNIR, K. KUCZERA, F. T. K. LAU, C. MATTOS, S. MICHNICK, T. NGO, D. T. NGUYEN, B. PRODHOM, W. E. REIHER, B. ROUX, M. SCHLENKRICH, J. C. SMITH, R. STOTE, J. STRAUB, M. WATANABE, J. WIÓRKIEWICZ-KUCZERA, D. YIN, and M. KARPLUS, *J. Phys. Chem. B* **102**, 3586 (1998).

[105] C. N. ROWLEY and B. ROUX, *J. Chem. Theory Comput.* **8**, 3526 (2012).

[106] J.-P. RYCKAERT, G. CICCOTTI, and H. J. BERENDSEN, *J. Comp. Phys.* **23**, 327 (1977).

[107] P. KOLLMAN, *Chem. Rev.* **93**, 2395 (1993).

[108] T. SIMONSON, G. ARCHONTIS, and M. KARPLUS, *Acc. Chem. Res.* **35**, 430 (2002).

[109] G. LI and Q. CUI, *J. Phys. Chem. B* **107**, 14521 (2003).

[110] C. A. SMITH and I. RAYMENT, *Biochemistry* **35**, 5404 (1996).

[111] M. A. GEEVES and K. HOLMES, *Annu. Rev. Biochem.* **68**, 687 (1999).

[112] I. RAYMENT, *J. Biol. Chem.* **271**, 15850 (1996).

[113] G. LI and Q. CUI, *J. Phys. Chem. B* **108**, 3342 (2004).

[114] H. YU, L. MA, Y. YANG, and Q. CUI, *PLoS Comput. Biol.* **3**, 0199 (2007).

[115] D. BASHFORD and M. KARPLUS, *Biochemistry* **29**, 10219 (1990).

[116] C. L. BROOKS III and M. KARPLUS, *J. Mol. Biol.* **208**, 159 (1989).

[117] D. RICCARDI and Q. CUI, *J. Phys. Chem. A* **111**, 5703 (2007).

[118] X. LU and Q. CUI, *J. Phys. Chem. B* **117**, 2005 (2013).

[119] G. TORRIE and J. VALLEAU, *J. Comput. Phys.* **23**, 187 (1977).

[120] S. KUMAR, D. BOUZIDA, R. H. SWENDSEN, P. A. KOLLMAN, and J. M. ROSENBERG, *J. Comput. Chem.* **13**, 1011 (1992).

[121] P. J. O'BRIEN, J. K. LASSILA, T. D. FENN, J. G. ZALATAN, and D. HERSCHLAG, *Biochemistry* **47**, 7663 (2008).

[122] S. FISCHER and M. KARPLUS, *Chem. Phys. Lett.* **194**, 252 (1992).

[123] Y. ZHAO and D. G. TRUHLAR, *Theor. Chem. Acc.* **120**, 215 (2008).

[124] S. E. S. GRIMME, J. ANTONY and H. KRIEG, *J. Chem. Phys.* **132**, 154104 1 (2010).

[125] M. J. FRISCH, G. W. TRUCKS, H. B. SCHLEGEL, G. E. SCUSERIA, M. A. ROBB, J. R. CHEESEMAN, G. SCALMANI, V. BARONE, B. MENNUCCI, G. A. PETERSSON, H. NAKATSUJI, M. CARICATO, X. LI, H. P. HRATCHIAN, A. F. IZMAYLOV, J. BLOINO, G. ZHENG, J. L. SONNENBERG, M. HADA, M. EHARA, K. TOYOTA, R. FUKUDA, J. HASEGAWA, M. ISHIDA, T. NAKAJIMA, Y. HONDA, O. KITAO, H. NAKAI, T. VREVEN, J. A. MONTGOMERY, JR., J. E. PERALTA, F. OGLIARO, M. BEARPARK, J. J. HEYD, E. BROTHERS, K. N. KUDIN, V. N. STAROVEROV, R. KOBAYASHI, J. NORMAND, K. RAGHAVACHARI, A. RENDELL, J. C. BURANT, S. S. IYENGAR, J. TOMASI, M. COSSI, N. REGA, J. M. MILLAM, M. KLENE, J. E. KNOX, J. B. CROSS, V. BAKKEN, C. ADAMO, J. JARAMILLO, R. GOMPERTS, R. E. STRATMANN, O. YAZYEV, A. J. AUSTIN, R. CAMMI, C. POMELLI, J. W. OCHTERSKI, R. L. MARTIN, K. MOROKUMA, V. G.

ZAKRZEWSKI, G. A. VOTH, P. SALVADOR, J. J. DANNENBERG, S. DAPPRICH, A. D. DANIELS, Ö. FARKAS, J. B. FORESMAN, J. V. ORTIZ, J. CIOSLOWSKI, and D. J. FOX, Gaussian 09 Revision A.1, Gaussian Inc. Wallingford CT 2009.

[126] T. DUDEV, J. A. COWAN, and C. LIM, *J. Am. Chem. Soc.* **121**, 7665 (1999).

[127] T. SON, M. ROUX, and M. ELLENBERGER, *Nucleic Acids Res.* **2**, 1101 (1975).

[128] T. R. J. COHN, M; HUGHES, *J. Biol. Chem.* **237**, 176 (1962).

[129] E. BISHOP, S. KIMBER, D. ORCHARD, and B. SMITH, *Biochim. Biophys. Acta* **635**, 63 (1981).

[130] S. L. HUANG and M. D. TSAI, *Biochemistry* **21**, 951 (1982).

[131] F. RAMIREZ and J. MARECEK, *Biochim. Biophys. Acta* **589**, 21 (1980).

[132] H. TAKEUCHI, H. MURATA, and I. HARADA, *J. Am. Chem. Soc.* **110**, 392 (1988).

[133] J. C. LIAO, S. SUN, D. CHANDLER, and G. OSTER, *Eur. Biophys. J.* **33**, 29 (2004).

[134] C. STAMPER, B. BENNETT, T. EDWARDS, R. C. HOLZ, D. RINGE, and G. PETSKO, *Biochemistry* **40**, 7035 (2001).

[135] M. ENGLERT, S. XIA, C. OKADA, A. NAKAMURA, V. TANAVDE, M. YAO, S. H. EOM, W. H. KONIGSBERG, D. SOELL, and J. WANG, *Proc. Natl. Acad. Sci. USA* **109**, 15235 (2012).

[136] M. J. JEDRZEJAS and P. SETLOW, *Chem. Rev.* **101**, 607 (2001).

[137] T. M. LARSEN, M. M. BENNING, I. RAYMENT, and G. H. REED, *Biochemistry* **37**, 6247 (1998).

[138] T. NAKAMURA, Y. ZHAO, Y. YAMAGATA, Y. JIN HUA, and W. YANG, *Nature* **487**, 196 (2012).

[139] B. H. SCHMIDT, A. B. BURGIN, J. E. DEWEESE, N. OSHEROFF, and J. M. BERGER, *Nature* **465**, 641 (2010).

[140] G. PALERMO, M. STENTA, A. CAVALLI, M. DAL PERARO, and M. DE VIVO, *J. Chem. Theory Comput.* **9**, 857 (2013).

[141] B. CHEVRIER, C. SCHALK, H. D'ORCHYMONT, J. RONDEAU, D. MORAS, and C. TARNUS, *Structure* **2**, 283 (1994).

[142] J. E. COLEMAN, *Annu. Rev. Biophys. Biomol. Struct.* **21**, 441 (1992).

[143] Y. MARCUS, *Chem. Rev.* **88**, 1475 (1988).

[144] L. BERNASCONI, E. J. BAERENDS, and M. SPRIK, *J. Phys. Chem. B* **110**, 11444 (2006).

[145] S. RIAHI, B. ROUX, and C. N. ROWLEY, *Can. J. Chem.* **91**, 552 (2013).

[146] H. J. KULIK, N. MARZARI, A. A. CORREA, D. PRENDERGAST, E. SCHWEGLER, and G. GALLI, *J. Phys. Chem. B* **114**, 9594 (2010).

[147] A. BHATTACHARJEE, A. B. PRIBIL, B. R. RANDOLF, B. M. RODE, and T. S. HOFER, *Chem. Phys. Lett.* **536**, 39 (2012).

[148] T. IKEDA, M. BOERO, and K. TERAKURA, *J. Chem. Phys.* **127**, 074503 1 (2007).

[149] A. KUZMIN, S. OBST, and J. PURANS, *J. Phys.: Condens. Matter* **9**, 10065 (1997).

[150] N. REGA, G. BRANCATO, A. PETRONE, P. CARUSO, and V. BARONE, *J. Chem. Phys.* **134**, 074504 1 (2011).

[151] J. NEELY and R. CONNICK, *J. Am. Chem. Soc.* **92**, 3476 (1970).

[152] L. HELM and A. MERBACH, *Coord. Chem. Rev.* **187**, 151 (1999).

[153] P. SALMON, M. BELLISSENFUNEL, and G. HERDMAN, *J. Phys.: Condens. Matter* **2**, 4297 (1990).

[154] B. E. DOUGLAS, D. H. MCDANIEL, and J. J. ALEXANDER, *Concepts and Models of Inorganic Chemistry*, John Wiley and Sons, 1994.

[155] R. VALE and R. MILLIGAN, *Science* **288**, 88 (2000).

[156] C. BUSTAMANTE, D. KELLER, and G. OSTER, *Acc. Chem. Res.* **34**, 412 (2001).

[157] Y. GAO, W. YANG, R. MARCUS, and M. KARPLUS, *Proc. Natl. Acad. Sci. USA* **100**, 11339 (2003).

[158] M. STRAJBL, A. SHURKI, and A. WARSHEL, *Proc. Natl. Acad. Sci. USA* **100**, 14834 (2003).

[159] M. J. MCGRATH, I.-F. W. KUO, S. HAYASHI, and S. TAKADA, *J. Am. Chem. Soc.* **135**, 8908 (2013).

[160] F. XIA, T. RUDACK, C. K. Q. CUI, and K. GERWERT, *J. Am. Chem. Soc.* **116**, 12522 (2012).

[161] A. LAIO and F. L. GERVASIO, *Rep. Prog. Phys.* **71**, 126601 1 (2008).

[162] W. E, W. REN, and E. VANDEN-EIJNDEN, *J. Chem. Phys.* **126**, 164103 1 (2007).

[163] L. MARAGLIANO, A. FISCHER, E. VANDEN-EIJNDEN, and G. CICCOTTI, *J. Chem. Phys.* **125**, 024106 1 (2006).

[164] E. VANDEN-EIJNDEN and M. VENTUROLI, *J. Chem. Phys.* **130**, 194103 1 (2009).

Chapter 4

Charging Free Energy Calculations Using the Generalized Solvent Boundary Potential (GSBP) and Periodic Boundary Condition[1]

4.1 Introduction

Electrostatic interactions are crucial to the structure and function of biomolecules.(1–3) Numerous efficient computational approaches have been proposed and implemented for calculating electrostatics in complex biomolecular systems. The most widely adopted approach is Ewald summation(4) and its various variations, such as the particle mesh Ewald (PME)(5), in the context of periodic boundary condition (PBC). Potential artifacts associated with the enforced periodicity have been thoroughly discussed in the literature and considered well understood(6–9). An alternative protocol often employed when only part of the system is of interest is the finite sphere boundary condition, where one describes the key ("inner") region of the system microscopically while the rest (the "outer" region), which generally contains bulk solvent, ions and the remaining biomolecule, is described with either continuum (macroscopic) electrostatics or a semi-macroscopic(10–12) model. A notable example is the Generalized Solvent Boundary Potential (GSBP) approach developed by Roux and co-workers(13) for classical simulations and later adopted by our group(14, 15) and others(16) for

[1]Reprinted (adapted) with permission from "Charging Free Energy Calculations Using the Generalized Solvent Boundary Potential (GSBP) and Periodic Boundary Condition: A Comparative Analysis Using Ion Solvation and Oxidation Free Energy in Proteins", X. Lu and Q. Cui, *J. Phys. Chem. B* **117**, 2005 (2013). Copyright 2013 American Chemical Society.

hybrid QM/MM simulations. By expanding the Green's function for the inner region electrostatic problem with a basis set, the GSBP provides an efficient approach for propagating molecular dynamics trajectories for the region of interest. This makes it possible to compute meaningful free energy changes associated with local events such as binding(17) and catalysis(18–20) in very large biomolecular systems, which remain computationally very expensive to handle using PBC based calculations.

Although several studies have carefully analyzed the performance of GSBP based classical(17, 21) and QM/MM simulations(16, 22, 23) for realistic biomolecules, and the general findings support the use of GSBP for processes that localize to regions far ($>$ 15-20 Å) from the microscopic/continuum boundary, a direct and systematic comparison between GSBP and PME based free energy simulations has not been carried out. As GSBP gets applied to increasingly complex biomolecules, it is important to understand the impact of various approximations inherent in the model and means to improve the results in quantitative applications. Specifically, the GSBP partitions a solvated macromolecule into the "inner" microscopic region and the "outer" macroscopic (continuum electrostatics) region. The outer region is fixed and described with distinct effective dielectric constants (e.g., $1-4$ for the protein and 80 for bulk water), and the interface between the inner and outer regions is also assumed to be fixed and molecularly smooth; this makes it possible to use a basis set expansion for the relevant Green's function. Clearly, GSBP is designed to tackle problems that do not involve large collective motions in the system. Nevertheless, it is important to understand for an apparently local process, such as the oxidation of a metal ion in the active site, what are the consequences of those inherent approximations.

For example, it is well appreciated that macroscopic concepts such as the dielectric "constant" do not readily apply to highly heterogeneous systems like proteins at the molecular scale(10–12, 24–26). Therefore, in principle it is desirable to adopt a more sophisticated dielectric model than a single effective (or empirical) dielectric constant to describe the fixed outer region, whose dielectric response is not explicitly sampled during the simulation. Prior to such development, however, it is important to establish the impact of the outer region; in realistic biomolecules, it is possible that active site properties are largely dictated by nearby interactions and the outer region contribution is effectively screened out by the inner region protein and solvent molecules. Moreover, the approximation of a rigid and smooth dielectric boundary between the inner and outer regions is expected to be most reliable when the net charge of the inner region is conserved; when the net charge changes,

such as during oxidation/reduction, (de)protonation or an alchemical mutation that involves charged groups, it is possible that the approximations introduce a significant error that needs to be corrected. It is important to understand when such errors become large and how can they be corrected.

In the current work, we compare GSBP and PBC based protocols for charging free energy simulations in both solution and proteins, although the observations and insights should apply to the general comparison of free energy simulations using finite-sphere vs. periodic boundary setups. There are several motivations to focus on charging free energy simulations for this comparative analysis. First, as stated above, the main goal is to understand the impact of various approximations in GSBP; studying processes that involve a net change in the total charge is most revealing in this context. Second, as discussed in great depth in recent studies(27–36), PME based free energy calculations are subject to several subtleties (although, as also discussed below, the magnitude of correction might be large) when the net charge of the system changes (see Fig. 4.1). Therefore, systematically comparing GSBP and PME results for charging free energy simulations of more diverse systems helps better understand these subtle but important correction terms.Third, in many realistic applications, the net charge of the system changes; examples include calculations of solvation free energy of ions(37, 38), redox potential(12, 39–41) and pK_a(11, 22, 42) values in solution and proteins, and relative binding energies of ligands that differ in their net charge(43, 44). Finally, from a practical perspective, charging free energy simulations have been shown to often follow linear response(19, 45–47); this allows adopting the linear response approximation (LRA(48)), which substantially reduces the computational cost when many free energy calculations need to be carried out although care must be taken to establish the validity of LRA for the specific problem in hand(42, 46) (see **Supporting Information**).

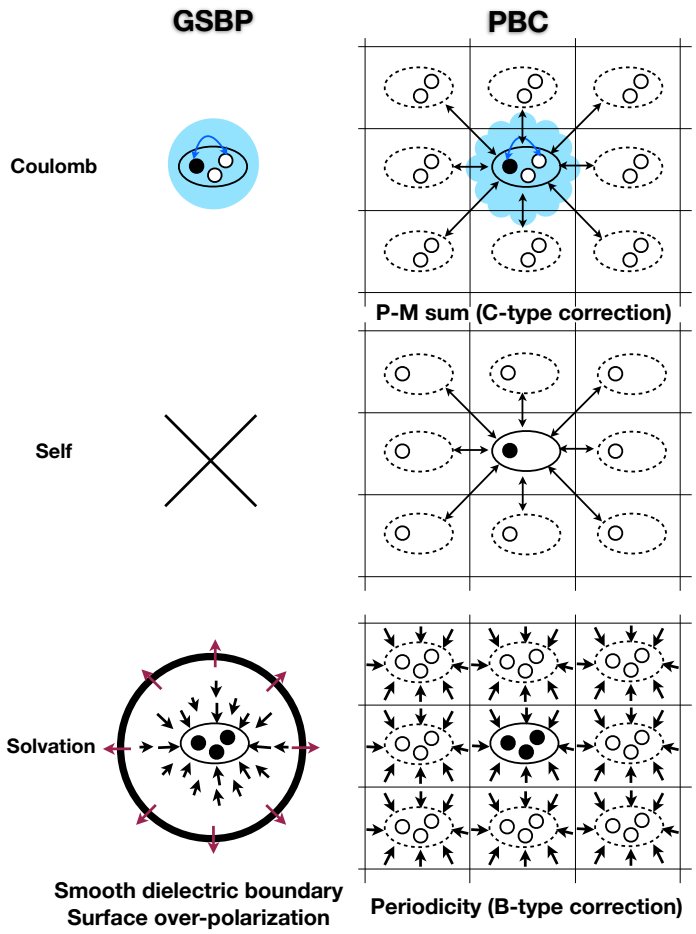


Figure 4.1: A scheme that illustrates the corrections necessary for GSBP and PBC (PME) simulations for electrostatic free energy simulations (modified based on Fig. 1 of Ref.(8)). The filled dot indicates the reference charge and unfilled dots for other charges; the single-headed arrows indicate polarization of the nearby solvent/protein dipoles. For PBC (PME) simulations, the corrections need to be considered include(36, 49) type B, which is due to the self energy and periodicity induced solvation terms, and type C, which is due to the use of an inappropriate particle (P) based summation scheme for Coulomb interactions; for systems studied here, only the type C correction (ΔG_{P-M} in Table 4.2) is found to be large and considered. For GSBP simulations, the corrections include those due to preferentially oriented water molecules (indicated by red arrows) at the inner/outer boundary (i.e., over-polarization, ΔG_{s-pol} in Table 4.2) and the smooth dielectric boundary approximation (corrected with Poisson-Boltzmann calculations, see Tables 4.4 and 4.6).

In the following, we first describe the computational methods and simulation details in Sect.4.2; results and discussions are presented in Sect.4.3. We summarize this work with conclusions in Sect.4.4.

4.2 Computational Methods

4.2.1 Charging free energy of ions in solution

As an important benchmark and reference(27–36), we first study the charging free energy of ions in solution using PBC and GSBP. Three common ions are chosen, Na^+, K^+, and Cl^-. The Lennard-Jones (LJ) parameters for these ions are taken from the CHARMM22 force field.(50) As to the water molecules, they are described with the modified TIP3P model in CHARMM(50, 51) (simply referred to as the TIP3P model below); we also carry out a few simulations with the SPC water model.(52) Moreover, to test the "type C" correction for PME based charging free energy simulations discussed in Ref.(36) (see discussions in Sect.4.3), which depends on the sign of the trace of the quadrupole moment of water, we also study the charging free energy of ions in "charge reversed" water, in which the partial charges on the water oxygen and hydrogen adopt the opposite sign (i.e., oxygen with +0.834 e and hydrogen with -0.417 e).

In the PBC simulations, an ion is solvated by 1,014 water molecules in a 31.5 Å cubic box with the density of 1.0 g/cm^3. The box dimensions and temperature are controlled with MD simulations in the NPT ensemble(4, 53); the temperature is 300K and pressure is 1 bar. Electrostatic interactions are treated with PME(54), and the van der Waals interactions are treated with a cutoff scheme with the switch function turned on between 10 Å and 12 Å. The PME parameters used are $\kappa = 0.33$, B-spline interpolation order of 6, and 32 grid points for the charge mesh.

In the GSBP simulations, the system is set up with an inner region of a specific radius (8, 20 or 25 Å). Because a 2 Å buffer region is required by the smooth boundary condition between the explicit solvents and the bulk dielectric continuum, GSBP simulations with a 8/20/25 Å inner region contains a water droplet of radius 6/18/23 Å. All molecules are located in the inner region, and the outer region is described with a constant dielectric continuum with $\varepsilon_w = 80$. To be consistent with the GSBP protocol, electrostatic interactions among inner region atoms are treated with the extended electrostatic model(55) in which interactions beyond 12 Å are modeled with multipolar expansions, including the dipolar and quadrupolar terms. The reaction field matrix $\mathbf{M}$ is evaluated using 400/900/961 spherical harmonics for GSBP simulations with a 8/20/25 Å inner

region. Temperature is controlled by coupling the system to the Nosé-Hoover thermostat(4, 56, 57) at 300K.

In both boundary setups, the charging ion is fixed at the center of the box/droplet and all bonds involving hydrogen are constrained with the SHAKE algorithm,(58) allowing a 1 fs time step for MD propagation. As mentioned in the Introduction, the charging free energy is computed with a LRA protocol(59) by running MD simulations with the charge of the ion either fully turned on ($\lambda = 0$) or turned off ($\lambda = 1$). The charging free energy is then calculated as $\Delta G = \frac{1}{2}(\langle q_I \Phi(0)\rangle_{\lambda=0} + \langle q_I \Phi(0)\rangle_{\lambda=1})$, where q_I is the charge (± 1) of the ion and $\Phi(0)$ is the electrostatic potential at the position of the ion; the quantity $q_I\Phi(0)$ is also referred to as the "energy gap" in the following discussions. The free energy derivatives are calculated until convergence is reached for each window after an initial equilibration period of 500 ps; statistical errors are estimated by block average(4). For correction terms for the charging free energies associated with PBC and GSBP simulations, see Sect.4.3.

As to the calculation of the electrostatic potential, $\Phi(0)$, it is straightforward in GSBP simulations. As discussed in previous work(13, 14), the electrostatic potential at a position $\mathbf{r}$ in the inner region is given by the sum of inner region charges, the outer region and the reaction field term, so energy gap for a single configuration is calculated by taking the total electrostatic energy difference between two charged states,

$$\begin{aligned} \Delta U_{GSBP} &= \Delta q_I \Phi(0)^{outer} + \Delta q_I \Phi(0)^{inner} + \frac{1}{2}[(q_I\Phi(0)_{rf})_{\lambda=0} - (q_I\Phi(0)_{rf})_{\lambda=1}] \\ &= [\sum_{\alpha \in inner} q_\alpha \Phi^{outer}(\mathbf{r}_\alpha) + U^{inner-inner} + \frac{1}{2}\sum_{\alpha \in inner} q_\alpha \Phi^{inner}_{rf}(\mathbf{r}_\alpha)]_{\lambda=0-\lambda=1.} \end{aligned} \tag{4.1}$$

$\Phi^{outer}(\mathbf{r})$ is the static field from the charged in the outer region acting onto the metal site taking into account the effect of the dielectric shielding of the environment, this term is independent of the instantaneous configuration of the inner region because of the smooth dielectric boundary approximation, so it can be calculated once and stored for computational efficiency. $U^{inner-inner}$ corresponds to the bare Coulomb interactions between inner region charges. Φ^{inner}_{rf} depends on the instantaneous positions of the inner region charges, so in practice, atomic charge distribution in the inner region are expressed in the form of a basis set expansion,

$$\rho^{inner}(\mathbf{r}) = \sum_m c_m b_m(\mathbf{r}). \tag{4.2}$$

The coefficients c_m of the expansion can be written as

$$c_m = \sum_n S^{-1}_{mn} Q_{n.} \tag{4.3}$$

with the elements of the overlap matrix S_{mn} given by

$$S_{mn} = \int d\mathbf{r} b_m(\mathbf{r}) b_n(\mathbf{r}). \tag{4.4}$$

Q_n are the generalized multipole moments, they are expressed as

$$Q_n = \sum_{\alpha \in inner} q_\alpha b_n(\mathbf{r}_\alpha). \tag{4.5}$$

The basis set representation of the inner-inner contribution to the electrostatic solvation energy is then given in a matrix form (we assume an orthonormal basis set is used),

$$\frac{1}{2} \sum_{\alpha \in inner} q_\alpha \Phi_{rf}^{inner}(\mathbf{r}_\alpha) = \frac{1}{2} \sum_{mn} Q_m M_{mn} Q_{n.} \tag{4.6}$$

M is the generalized reaction field matrix and represents the numerical reaction field Green's function between generalized multipoles Q_m,

$$M_{mn} = \int d\mathbf{r} b_m(\mathbf{r}) \Phi_{rf}(\mathbf{r}; b_n(\mathbf{r})). \tag{4.7}$$

where $\Phi_{rf}(\mathbf{r}; b_n(\mathbf{r}))$ is the reaction field at position $\mathbf{r}$ due to the basis function $b_n(\mathbf{r})$. **M** is independent of the instantaneous charge distribution of inner region, thus it needs to be computed once and stored for efficient simulations.

Rewriting the third term of Eq. (1), it becomes

$$\Delta U_{GSBP} = [\sum_{\alpha \in inner} q_\alpha \Phi^{outer}(\mathbf{r}_\alpha) + U^{inner-inner} + \frac{1}{2} \sum_{mn} Q_m M_{mn} Q_n]_{\lambda=0-\lambda=1.} \tag{4.8}$$

For PME simulations, the electrostatic potential is split into real space, reciprocal space and self energy terms, so energy gap for a single configuration between two charge states is also calculated by taking the total electrostatic energy difference,

$$\begin{aligned} \Delta U_{PME} &= \Delta q_I \Phi(0)^{real} + \Delta q_I \Phi(0)^{recip} + \Delta U^{self}_{(\lambda=0-\lambda=1)} \\ &= [\frac{1}{2} \sum_{i,j=1}^{N} \sum_{|\mathbf{n}=0|}^{\infty} q_i q_j \frac{erfc(\kappa|\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|} + \frac{1}{2V} \sum_{i,j=1}^{N} \sum_{k \neq 0} q_i q_j \frac{4\pi}{k^2} \times \\ &\quad exp(-\frac{k^2}{4\kappa^2}) cos(k\mathbf{k} \cdot \mathbf{r}_{ij}) - \frac{\kappa}{\sqrt{\pi}} \sum_{i=1}^{N} q_i^2]_{\lambda=0-\lambda=1} \end{aligned} \tag{4.9}$$

4.2.2 Oxidation free energy for a metal site in a protein

To study oxidation potential in a protein, we have chosen a somewhat unconventional model system: the metal site in a copper chaperone protein, CueR. CueR is a transcription regulator that responds to the binding of Cu(I) ion(60–62); it is *not* known to be a redox active protein. Nevertheless, we chose it as a model system because its metal site has a similar distance from the protein/solvent boundary as blue copper proteins, which are well studied redox systems(40, 41). Compared to the blue copper proteins, CueR is larger in size and likely more flexible due to its structural topology; therefore, CueR is better suited for comparing PBC and GSBP simulations, such as probing the effect of constraining the position of protein atoms far (> 20Å) from the redox site.

The enzyme model is constructed based on the crystal structure(60), and a bonded model is parameterized for the metal ion M, where M-ligand coordination interactions are modeled as covalent bonds and active site structural features are maintained by applying bond, angle and dihedral terms to the metal ion and its immediate ligands (Fig. 4.2). For the purpose of this work, only the partial charge of the metal ion is changed in the "oxidation free energy" calculations; it is changed from +0.1158 e ($\lambda = 0$, the "reduced" state) to +1.1158 e ($\lambda = 1$, the "oxidized" state) and then to +2.1158 e ($\lambda = 2$, the "doubly-oxidized" state). Similar to the charging free energy of ions in solution, the oxidation free energies in CueR are calculated using the LRA protocol by running MD simulations at $\lambda = 0, 1$ and 2; energy gaps are then collected from those trajectories and then averaged to obtain the two oxidation free energies: $\Delta G = \frac{1}{2}(\langle \Delta q_I \Phi(0) \rangle_{\lambda=0} + \langle \Delta q_I \Phi(0) \rangle_{\lambda=1})$, where Δq_I is the charge difference between the oxidized and reduced states of the metal ion M and $\Phi(0)$ is the electrostatic potential at M. Previous studies of reduction potential/pK_a in proteins suggested that LRA may not always apply to protein systems(42). In our calculations, we have carefully monitored the distribution of the energy gaps and found that they closely follow Gaussian statistics in all three λ windows (see **Supporting Information**), suggesting that LRA is applicable; moreover, we have carried out simulations for multiple λ windows for two representative types of PBC simulations (rigid protein and fully flexible protein) and explicitly verified the validity of LRA (see **Supporting Information**).

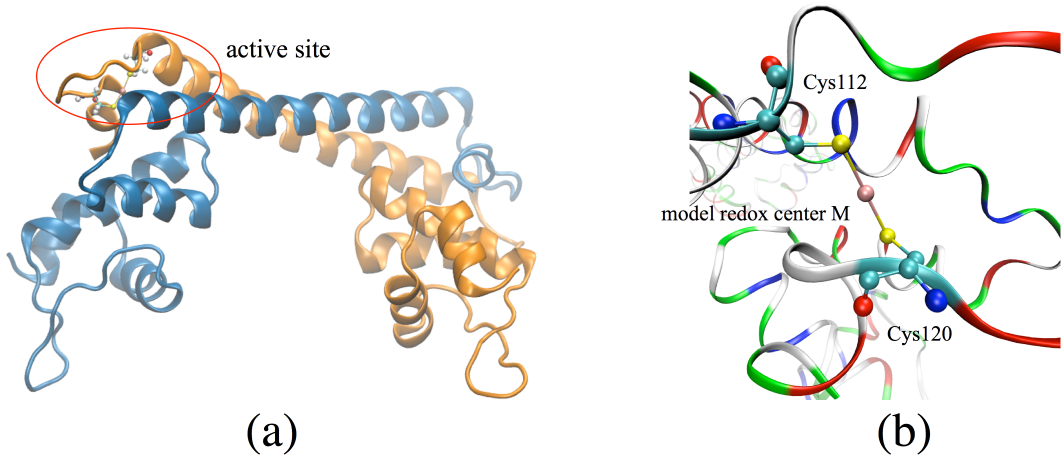


Figure 4.2: The CueR protein used as the model system for oxidation free energy calculations with different boundary conditions. (a) The protein is a homodimer and shown in the ribbon form. One of the metal sites is highlighted by a ellipse and the metal ion M along with its Cys ligands are shown in the ball-and-stick form. (b) A closer view of the metal site (color coding for the ribbon: red, acidic; blue, basic; green, polar; white, nonpolar). A bonded model is parameterized for the metal and its two Cys ligands. Three oxidation states of the metal site are studied, the reduced state ($\lambda = 0$, partial charge of M is +0.1158 e), the oxidized state ($\lambda = 1$, partial charge of M is +1.1158 e) and the doubly-oxidized state ($\lambda = 2$, partial charge of M is +2.1158 e).

In the PBC simulations, water molecules are added to solvate the protein and the final system is a rectangular unit cell with dimensions of 64 Å $\times$ 64 Å $\times$ 120 Å; it contains 4,022 protein atoms and 15,907 TIP3P water molecules. Sodium and chloride ions are added to reach a concentration of $\sim$0.15 M and to neutralize the total charge of the system. To better compare PBC and GSBP simulations, in addition to PBC simulations with a fully flexible protein, PBC simulations are also carried out with either the entire or part of the protein fixed; the water molecules are fully mobile in all PBC simulations. The fully rigid PBC simulations are compared to GSBP simulations with also a rigid protein; such a comparison eliminates errors that originate from sampling different protein conformations in the PBC and GSBP simulations. In PBC simulations with a partially rigid protein, the same set of atoms as in the outer region of the GSBP setup are fixed; with atoms beyond 20 (8) Å of the metal site fixed, 1,021 (146) protein atoms are left fully flexible. Isothermal-isobaric ensemble is used when the entire protein is flexible; NVT ensemble is used when either the entire or part of the protein is fixed.

In the GSBP setup, the protein is partitioned into a spherical inner region (with a radius of 8 or 20 Å) while the remaining portion of the system is in the outer region, and only atoms in the inner region undergo microscopic dynamics. Protein atoms in the outer region are fixed. For the inner region of 20 Å, the initial structure for all λ windows is an equilibrated structure from the PBC simulation for a fully flexible CueR in the $\lambda = 0$ charge state; water molecules are added by superimposing the system with a water droplet of 18 Å radius centered at M and removing water molecules with oxygen atom within 2.8 Å from any heavy atoms in the initial structure. For the inner region of 8 Å, initial protein structure of each λ is taken from PBC simulation with the corresponding charge state and a fully flexible protein; the number of explicit water molecules in the GSBP simulations are determined from the average number of waters located within the 6 Å droplet centered at M in the PBC simulations. In all GSBP simulations, the protein dielectric constant is $\varepsilon_p = 1$, the water dielectric constant is 80; no salt concentration is considered, although test calculations indicate that including 0.15 M of salt in the Poisson-Boltzmann (PB) calculations has very small effects. The extended electrostatic model is used to treat the electrostatic interactions among inner region atoms as described above. The reaction field matrix **M** is evaluated using 289/49 spherical harmonics for GSBP simulations with a 20/8 Å inner region. Nosé-Hoover thermostat is used to control the temperature at 300 K.

With both boundary conditions, the CHARMM22 force field(50) is used for protein residues except for the metal ion M and its Cys ligands, which are parameterized based on cluster DFT calculations (S. Yang and Q. Cui, unpublished). During all MD simulations, all bonds involving hydrogen are constrained with the SHAKE algorithm,(58) and the time step is set as 1 fs. All simulations are equilibrated for 500 ps and then sampled for 3 to 9 ns, depending on the flexibility of protein. In all free energy derivatives calculations, atoms that are bound to the metal ion M are also treated as point charges (using the NBXMOD command in CHARMM); this facilitates the comparison with analysis using PB calculations (see below).

4.2.3 Analysis of differences between PBC and GSBP simulations

As discussed in greater detail in the next section, PBC (PME) and GSBP free energy results differ substantially in numerical values for both solution and enzyme (CueR) cases. Qualitatively, this is understood in terms of several corrections necessary for both types of simulations (see Fig. 4.1). For GSBP, the correction is mainly related to the polarization of water molecules at the boundary of the finite sphere, although the

smooth dielectric boundary approximation is also shown below to require extra care when the inner region is highly charged. For PME, the largest correction observed in this work (using the PME implemented in CHARMM(63)) is due to the use of particle-based summation scheme in computing electrostatic potential at the site of interest(49). To understand these corrections more quantitatively, we analyze electrostatic potentials in the systems of interest using PB calculations with the PBEQ module in CHARMM(64).

4.2.3.1 Analysis of electrostatic potential in a spherical water droplet

To understand the difference between PBC and GSBP for the charging free energy in solution, we analyze the average electrostatic potentials in three types of spherical water droplets at 300 K: a droplet with the orientation disorder limit (ODL) ensemble(36), a droplet with GSBP and a droplet with vacuum beyond the boundary; a quartic restraining potential is applied to water molecules in all setups: $V_{geo}(r) = k_{geo} \times (r - \delta r_{off})^2 \times ((r - \delta r_{off})^2 - P_1) + k_{geo} P_1^2/4$ for $-\sqrt{P_1/2} \leq (r - \delta r_{off}) \leq \sqrt{P_1/2}$, where k_{geo}=0.5 kcal·mol^{-1}·Å^{-1}, $P_1 = 2.25$ Å^2 and the offset distance δr_{off} is 15.5 Å. The ODL ensemble refers to a water simulation in the absence of electrostatic interactions; charges are put back for the analysis of electrostatic potentials. Both TIP3P and SPC water models are tested because TIP3P is widely used together with the CHARMM force field, and SPC was used in the original analysis of Hünenberger and co-workers(8, 36, 49). In all droplet simulations, the radius is 18 Å and the droplet consists of 748 water molecules. The extended electrostatic model is used to approximate exact Coulomb interactions without any cutoff. Each simulation runs for 1 ns and the last 1000 snapshots (500 ps) are used for analysis. To evaluate the electrostatic potential distribution in the sphere, 10,000 sites are chosen randomly in each configuration; electrostatic potential due to all solvent charges at each site is calculated by averaging over all the grid points inside a 1 Å^3 cubic box using the PBEQ module(64) in CHARMM.

4.2.3.2 Poisson-Boltzmann analysis for the smooth dielectric boundary in GSBP

To better understand the errors that arise from the smooth boundary approximation in GSBP (see discussions in Sect.4.3), we re-evaluate the energy gap of selected number of snapshots from GSBP simulations using PB calculations; for each snapshot, this involves two PB calculations that evaluate the direct solvent contributions and reaction field contribution from the charging ion, respectively. In the PB calculations, all explicit water molecules are retained and regarded as part of the solute, which is assigned to a dielectric constant

of 1; the implicit bulk solvent has a dielectric constant of 80. The optimized radii of Nina et al.(65) based on experimental solvation free energies of small molecules are employed to determine the van der Waals surface of the microscopic solute region. Solvent-solute dielectric boundary is determined by expanding beyond the van der Waals surface with a solvent probing radius; three values (0.8, 1.4 and 2.0 Å) have been tested here. To contrast with the rough molecular surface, PB calculations with a smooth dielectric boundary (the "droplet" option in PBEQ) have also been carried out.

4.3 Results and Discussions

4.3.1 Charging free energy of ions in a water droplet: correction terms for PME and GSBP simulations

The calculated charging free energy for Na^+ and Cl^- using different boundary conditions and water models are given in Table 4.1. With the TIP3P water model, the PBC results are comparable with those of Hummer et al,(27) who used slightly different LJ parameters for the ion and SPC water; although only data for 30 Å box are shown, the results show very little box-size dependence when a larger box (60 Å) or smaller boxes are used (data not shown). The GSBP results are consistent with those obtained with the spherical solvent boundary potential (SSBP)(66). However, there is a substantial difference of approximately $\pm$10 kcal/mol between PBC and GSBP results for a specific type of ion: GSBP overestimates the charging free energy of Na^+ but underestimates the value for Cl^-. This is not due to the difference in the solvation structure of the ions, as the ion-water oxygen radial distribution functions calculated from PBC and GSBP are practically identical for both Na^+ and Cl^- (data not shown). Rather, the difference is due to corrections needed for the treatment of electrostatics in GSBP and PBC simulations.

Table 4.1: Uncorrected charging free energy (in kcal·mol^{-1}) calculated using PBC and GSBP boundary conditions[a]

Water model and ion type	PBC (32 Å[b])	GSBP (20 Å[c])	$\Delta G_{PBC,GSBP}$
TIP3 water[d] + Na^+	-97.9 (0.6[f])	-107.7 (0.6)/-105.1[g]	9.8 (0.8)
TIP3 water + Cl^-	-100.4 (0.7)	-91.8 (0.6)	-8.6 (0.9)
SPC water + Na^+	-96.5 (0.7)/-97.3[h]	-107.9 (0.7)	11.4 (1.0)
"Charge-reversed water"[e] + Na^+	-176.6 (0.7)	-165.9 (0.7)	-10.7 (1.0)
"Charge-reversed water" + Cl^-	-59.5 (0.4)	-68.6 (0.8)	9.1 (0.9)

[a] The uncorrected values are obtained by directly integrating the computed free energy derivatives from CHARMM simulations. As discussed in the text, the PBC results need to be revised with a "type C" correction due to the use of particle based summation scheme in PME (ΔG_{P-M}), and GSBP results need to be supplemented with a surface polarization correction (ΔG_{s-pol}). These corrections are summarized in Table 4.2. [b] Cubic box length for PBC simulations. [c] Radius of inner region for GSBP simulations. [d] Modified TIP3P water model in CHARMM. [e] Reverse the charges on the TIP3P water model, so that the oxygen is +0.834 e and the hydrogen is -0.417 e. [f] Numbers before and after slash are, respectively, from this work and Ref.(66), which used SSBP to compute the Na^+ charging free energy in a droplet of 100 water molecules. [g] Numbers before and after slash are, respectively, from this work and Ref.(27) [h] The values in parentheses are statistical errors based on block average.

According to Kastenholz et al.,(49), two types of corrections are relevant to the simulations considered here (see Fig. 4.1). The first is the P-M type correction (referred to as type C_1 correction in Ref.(49)), which accounts for (a nearly constant) the difference in the electrostatic potential computed by summation in the particle (P) and molecular (M) schemes. It was argued that the M scheme is more physical(49); for example, in an orientation disorder limit (ODL), the P scheme leads to an incorrect net external potential in the water droplet while the M scheme does not. Since the P scheme is used in PME, while the M scheme is used for GSBP (group based cutoff with extended electrostatics), the P-M correction needs to be considered when comparing PBC and GSBP charging free energies. The magnitude of the P-M correction can be estimated in two ways: i). an analytical expression for the correction electrostatic potential, $\Delta\Phi^{anal}_{ODL} = (6\varepsilon_0)^{-1}\rho\gamma$, which was derived based on an isotropic quadrupole (IQ) liquid(36); ε_0 is the vacuum dielectric permittivity, ρ the liquid number density and γ the trace of the quadrupole moment of the water model. ii). numerically evaluate the average electrostatic potential or external electrostatic potential (potential at a site that is not too close to any atom in the solvents) of the ODL water ensemble, $\Delta\Phi^{numr}_{ODL}$. The $\Delta\Phi^{anal}_{ODL}$ is 17.8 kcal·mol^{-1}·e^{-1} for the TIP3P water model and 19.1 kcal·mol^{-1}·e^{-1} for SPC, while $\Delta\Phi^{numr}_{ODL}$ is 12.4 kcal·mol^{-1}·e^{-1} for TIP3P and 17.8 kcal·mol^{-1}·e^{-1} for SPC (Table 4.2). Since the IQ liquid deviates from "real" water models, the

analytical value $\Delta\Phi_{ODL}^{anal}$ may not be quantitatively appropriate for realistic simulations, thus we use the second option to calculate the P-M correction required for the charging free energy; this takes the form of $\Delta G_{P-M} = -\Delta q_I \Delta\Phi_{ODL}^{numr}$ and is needed for the PBC result.

Table 4.2: The average potential (in kcal·mol^{-1}·e^{-1}) in spherical water droplets and the derived correction terms for PME/GSBP charging free energy simulations

Finite size boundary condition	Average potential[d]
ODL[a]	-12.4/-12.4/-17.8/-16.8
GSBP[b]	-10.5/-11.0/-12.6/–
Vacuum[c]	-10.3/-11.0/-12.4/-13.7
ΔG_{P-M} [e]	-12.4/-12.4/-17.8/-16.8
ΔG_{s-pol} [f]	-1.9/-1.4/-5.2/–

[a] ODL refers to the orientation disorder limit ensemble(36) (simulations in the absence of electrostatic interactions) with a quartic boundary potential (see text). [b] The outer region of GSBP is the dielectric continuum with $\varepsilon = 80$; water molecules near the interface are subject to the same quartic boundary potential. [c] The water droplet is surround by vacuum; water molecules near the interface are subject to the same quartic boundary potential. [d] For each entry, the numbers are calculated based on the modified TIP3P water model in an 18 Å water droplet, the modified TIP3P water model in a 23 Å water droplet, the SPC water model in an 18 Å water droplet; the last value, when available, is from Ref.(36), which used the SPC water model and a Lennard-Jones boundary potential instead of the quartic boundary potential used here. [e] The P-M correction for the PME charging free energy(36, 49) calculated based on the average electrostatic potentials from the ODL ensemble. [f] The surface polarization correction for GSBP charging free energy calculated based on the difference in the average potential between ODL and GSBP simulations.

The second correction is a surface (fluid-continuum interface) polarization contribution (referred to as the type C_2 correction in Ref.(49)), which accounts for the spurious anisotropy of water molecules close to the boundary of finite sphere simulations. This is illustrated explicitly in Fig. 4.3, which plots the orientational distribution (between the water dipole and surface normal of the droplet) for water molecules within 1.5 Å from the boundary. Clearly, the surface water molecules in GSBP have a preferential orientation compared with bulk water or the ODL ensemble, and this leads to an increase of electrostatic potential at the charging ion; the magnitude of the polarization is similar when a water droplet is placed in vacuum (with the same radial quartic potential applied to the surface water) instead of a dielectric continuum of $\varepsilon = 80$ in GSBP. We evaluate the effect due to this artificial polarization by taking the difference of the average electrostatic potential inside a water droplet sampled with regular GSBP and an ODL ensemble; the corresponding correction to the charging

free energy is then, $\Delta G_{s-pol} = -\Delta q_I \Delta \Phi_{surf}$, and needs to be added to the GSBP result.

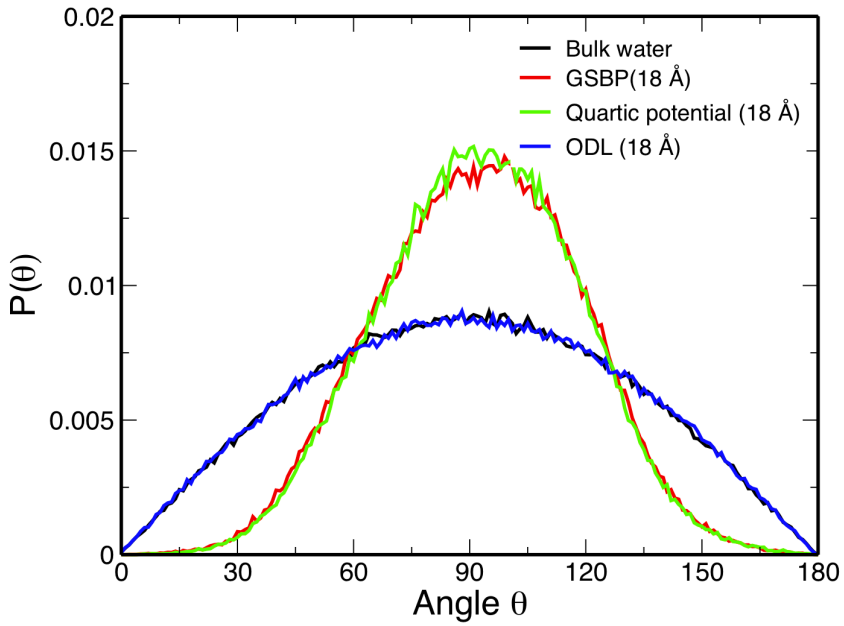


Figure 4.3: The orientation distribution of water molecules in bulk (based on PBC calculations), an ODL ensemble and two other water droplet simulations; for the last three cases, orientations for water molecules within 1.5 Å from the droplet surface are shown. The orientation is defined as the angle between the water dipole and the normal vector of the local surface.

As shown in Table 4.2, the average potential in a 18 Å water droplet sampled with ODL and GSBP using the TIP3P model is -12.4 kcal·mol^{-1}· e^{-1} and -10.5 kcal·mol^{-1}·e^{-1}, respectively. These imply that, for the charging free energy of a cation, the P-M correction (ΔG_{P-M}) is -12.4 kcal/mol and the surface polarization correction (ΔG_{s-pol}) is -1.9 kcal/mol. Including these corrections to the uncorrected PBC and GSBP results summarized in Table 4.1 lead to consistent charging free energies for Na^+ as -110.3 kcal/mol with PBC and -109.6 kcal/mol with GSBP. Alternatively speaking, the difference of -10.5 kcal/mol between the P-M correction and surface polarization correction explains the difference between uncorrected PBC and GSBP charging free energies for Na^+. Moreover, since both ΔG_{P-M} and ΔG_{s-pol} depend linearly on the sign of the

charging ion (q_I), their values as well as the difference between them adopt the opposite sign when considering the charging free energy of an anion; this is consistent with the observation noted above (Table 4.1) that uncorrected GSBP result is lower than the uncorrected PBC result for Cl^- by approximately +10 kcal/mol. Similar electrostatic potentials are also computed for a larger water droplet with a radius of 23 Å (Table 4.2), suggesting that the potentials are insensitive to the droplet size for GSBP or SSBP simulations.

As a further validation, we examine the charging free energies of Na^+ and Cl^- in a "charge-reversed" TIP3P water droplet (private communication with Gabriel Rocklin). In the charge-reversed water model, the partial charges on oxygen and hydrogen are inverted from those in the standard TIP3P model; i.e., oxygen has +0.834 *e* and hydrogen has -0.417 *e*. Due to the pair-wise additive nature of a classical Coulomb model, inverting the partial charges do not change the water-water interactions. However, the charge inversion flips the sign of the quadrupole trace of the water model and therefore the sign of both ΔG_{P-M} and ΔG_{s-pol} for charging free energy calculations. Indeed, as shown in Table 4.1, the difference between uncorrected GSBP and PBC charging free energies in "charge-reversed" water is of the same absolute magnitude as with the TIP3P model for a specific ion, albeit with the opposite sign.

With the SPC model, the average potential in an 18 Å water droplet sampled with an ODL ensemble and a radial quartic restraining potential is -17.8 kcal·mol^{-1}·e^{-1} and -12.4 kcal·mol^{-1}·e^{-1}, respectively; these are in good agreement with values reported in Ref.(49) despite that a L-J restraining potential was used for surface water molecules. Accordingly, the expected difference between uncorrected PBC and GSBP charging free energies for a cation in SPC water is about -12.6 kcal/mol, which is close to the computed values of 11.4 kcal/mol for Na^+ (see Table 4.1).

Finally, since both ΔG_{P-M} and ΔG_{s-pol} do not depend on the characteristics of the ion other than its charge, we expect that they cancel out for the relative charging free energy of ions of the same charge. Indeed, as shown in Table 4.3, GSBP and PBC give very consistent relative charging free energy for Na^+ and K^+, even for GSBP calculations with a small radius (8 Å).

Table 4.3: Relative charging free energy for cations (in kcal·mol^{-1}) calculated using PBC and GSBP boundary conditions

Boundary condition	Na^+	K^+	$\Delta G_{Na^+,K^+}$
PBC (32 Å[a])	-97.9 (0.6[c])	-77.3 (0.5)	-20.6 (0.8)
GSBP (25 Å[b])	-107.7 (0.5)	-87.1 (0.5)	-20.6 (0.7)
GSBP (20 Å[b])	-107.7 (0.6)	-86.8 (0.6)	-20.9 (0.8)
GSBP (8 Å[b])	-108.9 (0.6)	-89.1 (0.7)	-19.8 (0.9)

[a] Cubic box length for PBC simulations. [b] Radius of inner region for GSBP simulations. [c] The values in parentheses are statistical errors based on block average.

4.3.2 Charging free energy of Na^+ in a water droplet with additional charge: smooth boundary in GSBP simulations

As discussed below in Sect.4.3.3, GSBP appears to be problematic when dealing with an inner region that has a high net charge. To better understand the situation, we study the charging free energy of Na^+ in water in the presence of several (1-5) Cl^- ions with both PBC and GSBP simulations. As shown in Table 4.4, the PBC results do not vary much at all with up to 5 Cl^- in a 30 Å box; the uncorrected value remains close to -98.0 kcal/mol. By contrast, the uncorrected GSBP result steadily increases from -107.7 kcal/mol to -116.7 kcal/mol as the number of Cl^- increases from 0 to 5. As a result, the difference between uncorrected GSBP and PBC results significantly increases from 9.8 kcal/mol to 18.2 kcal/mol as the number of Cl^- increases to 5.

Qualitatively, we believe that this additional problem for GSBP is due to the fact that the uncertainty associated with computing electrostatic solvation free energy using the smooth dielectric boundary approximation in GSBP becomes increasingly large as the inner region charge increases. For the charging free energy of a cation in a water droplet, the dominant GSBP contribution is the Born term associated with the net charge of the inner region: $(Q+1)^2/2R_{in} - Q^2/2R_{in} = (2Q+1)/2R_{in}$, where Q is the charge of the inner region without considering the charging ion and R_{in} is the effective size of the inner region. Clearly, this term becomes large when Q is large; similarly, any uncertainty in the estimation of R_{in} would also lead to a significant error in the GSBP contribution when Q is large.

To circumvent the problem, we re-evaluate the energy gaps from selected (1000) snapshots from GSBP simulations using PB calculations; the free energy derivatives are corrected by taking the difference between

PB and GSBP energy gaps for these snapshots. The use of exact PB solution with a detailed dielectric boundary to postprocess the MD conformations was also used by Simonson in previous work.(43) As shown in Table 4.4, this works very well. For example, with the typically recommended value for the solvent probe radius of 1.4 Å, the PB-corrected GSBP charging free energy for Na^+ in the presence of 5 Cl^- becomes -109.0 kcal/mol, which is close to the value of -107.7 kcal/mol without any Cl^-; the standard deviation for the PB correction among the 1000 snapshots is about 0.7 kcal/mol (for more detailes, see **Supporting Information**). The PB correction is fairly insensitive to the solvent probe radius. With a smooth dielectric boundary of radius 20 Å, which is the same as in the GSBP set-up, however, the PB correction is essentially zero, further confirming that the underlying problem is due to the uncertainty associated with computing electrostatic solvation free energy using a smooth dielectric boundary. Therefore, for charging free energy simulations with GSBP, extra care needs to be exercised when the inner region is highly charged; one effective way to circumvent the problem is to perform PB corrections based on selected snapshots(43). This is found effective for both solution and protein (see below) systems.

Table 4.4: Calculated charging free energy (in kcal·mol^{-1}) of Na^+ with different net charges in the systems using PBC and GSBP boundary conditions, and Poisson-Boltzmann correction for the GSBP results

Net charge[a]	PBC (32 Å[b])	GSBP (20 Å[c])	$\Delta G_{PBC,GSBP}$
0	-97.9 (0.6[d])	-107.7 (0.6)	9.8 (0.8)
-1	-98.0 (0.6)	-109.3 (0.6)	11.3 (0.8)
-2	-97.9 (0.5)	-110.9 (0.7)	13.0 (0.9)
-3	-98.4 (0.8)	-113.1 (0.5)	14.7 (0.9)
-4	-98.2 (0.7)	-115.3 (0.4)	17.1 (0.8)
-5	-98.5 (0.6)	-116.7 (0.6)	18.2 (0.8)
Solvent probe radii	PBC (32 Å)	PB[e]	Difference
0.8 Å/1.4 Å/2.0 Å	-98.5 (0.6)	-110.6/-109.0/-109.0 (0.8)	12.1/10.5/10.5 (1.0)
20 Å droplet[f]	-98.5 (0.6)	-117.0 (0.8)	18.5 (1.0)

[a] Different numbers of Cl^- are added to the system to adjust the total net charge; a neutral sodium atom is taken as the reference to specify the net charge of the system. [b] Cubic box length for PBC simulations. [c] Radius of inner region for GSBP simulations. [d] The values in parentheses are statistical errors based on block average. [e] Snapshots used in PB calculations are taken from GSBP simulations with -5 net charge; the dielectric constant is set to 1 for all explicit solvent/ions and 80 for the bulk (continuum) solvent. [f] A smooth dielectric boundary condition with a radius of 20 Å ($\varepsilon = 1$) is used to mimic the GSBP setup.

4.3.3 Oxidation free energy for the metal site in a fixed CueR protein

To better explore the impact of the issues discussed above on GSBP and PBC charging free energy calculations for proteins, we first discuss oxidation free energy calculations for a metal site in a fixed CueR system. In this way, we avoid contributions due to differences in the sampled configurations in GSBP and PBC simulations; these contributions will be discussed in the next subsection.

As shown in Table 4.5, similar to the situation of charging free energy in a water droplet, GSBP and PBC calculations give very consistent results for the relative oxidation free energy for two metal sites that differ by the size of the metal ion. The van der Waals radius of the metal ion is taken to be 1.68 and 0.68 Å, respectively, and the relative oxidation free energy is 11.5 kcal/mol and 10.9 kcal/mol with PBC and GSBP, respectively. The basic trends can be understood by looking at the water accessibility of the metal site. As shown in Fig. 4.4, the metal site is not immediately accessible to water in the "reduced state" ($q_I = +0.1158e, \lambda = 0$) regardless of the van der Waals radius of the metal ion or the boundary condition; this is expected because the partial charges and bonded parameters of the "reduced" metal ion are developed to reflect the buried nature of the metal site in CueR(60). In the "oxidized state" ($q_I = +1.1158e, \lambda = 1$), the metal site becomes more water accessible as indicated by the significant water distribution peak at fairly short metal-oxygen distances (< 3 Å), especially for the metal ion with the smaller van der Waals radius (Fig. 4.4d). For all the states sampled, GSBP and PBC give essentially identical water distributions around the metal site.

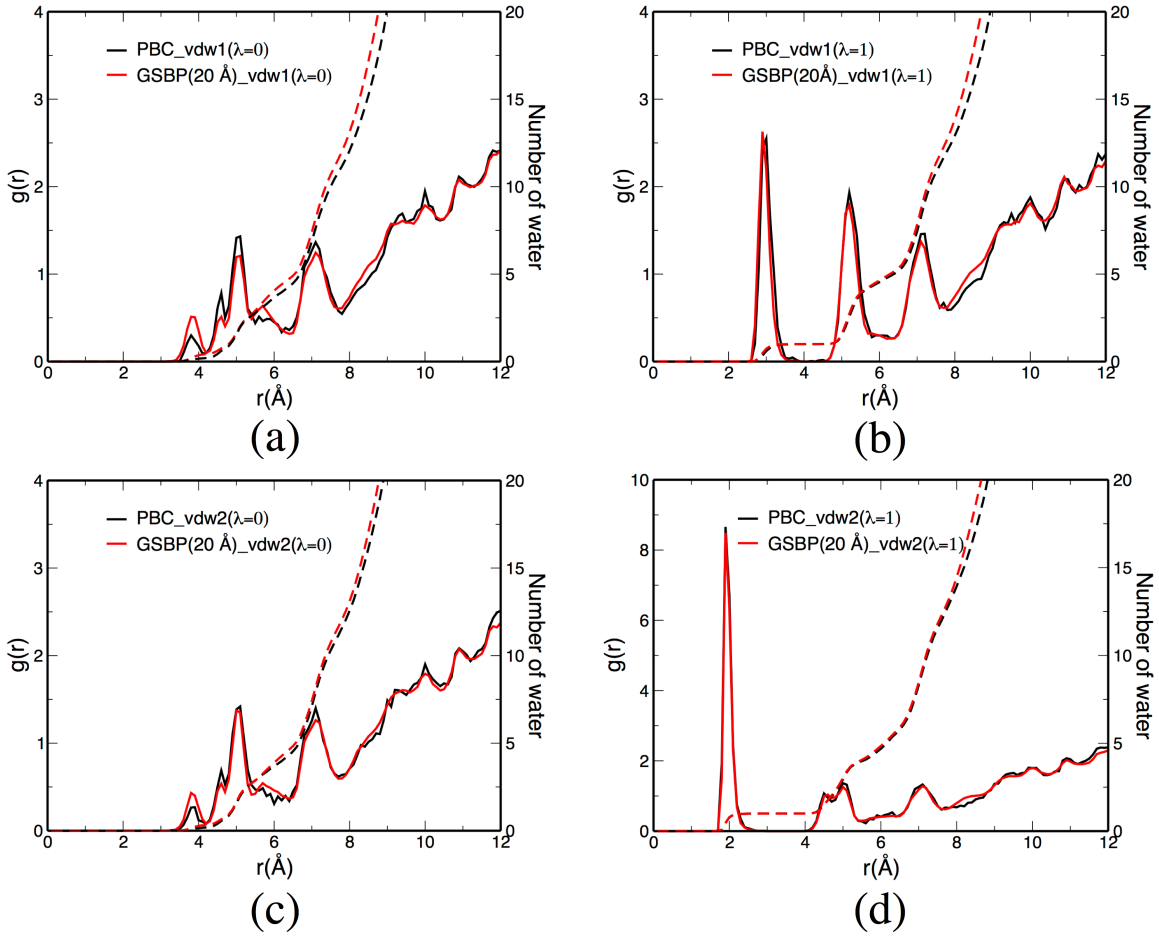


Figure 4.4: Radial distribution function of water oxygen around M in PBC and GSBP (20 Å) simulations of a rigid CueR protein. Different boundary conditions give reasonably similar water distributions. (a) M in the reduced state ($\lambda = 0$) with a van der Waals radius of 1.68 Å; (b) M in the oxidized state ($\lambda = 1$) with a van der Waals radius of 1.68 Å; (c) M in the reduced state ($\lambda = 0$) with a van der Waals radius of 0.68 Å; (d) M in the oxidized state ($\lambda = 1$) with a van der Waals radius of 0.68 Å.

Table 4.5: Oxidation free energy (in kcal·mol^{-1}) for a metal site M in a rigid CueR protein calculated using PBC and GSBP boundary conditions

Boundary condition	M[a]	M'[b]	$\Delta G_{M,M'}$
PBC	-74.5 (1.2[c])	-86.0 (1.3)	11.5 (1.8)
GSBP (20 Å)	-99.6 (0.9)	-110.5 (1.2)	10.9 (1.5)

[a] M is the model redox center in CueR with a van der Waals radius of 1.68 Å. [b] M is the model redox center in CueR with a smaller van der Waals radius of 0.68 Å. [c] The values in parentheses are statistical errors based on block average.

For a specific metal ion, however, the PBC and GSBP oxidation free energies (without the corrections discussed in the last two subsections) differ by as much as 25 kcal/mol, which is significantly different from the value of ∼10 kcal/mol found for charging free energies in solution. The discussion in Sect.4.3.2 suggests that this deviation is due to the high charge of the inner region in the GSBP simulations of CueR. Indeed, with a sphere of 20 Å that centers on the metal site, the inner region has a charge of -8.12 *e*, leaving the outer region with a charge of -3.88 *e*; these charges are not integers because the smooth dielectric boundary is determined by geometry instead of using a group-based criterion. Perturbative analyses are then performed to better understand the role of these net charges on the calculated oxidation free energy; the net charges are modulated by applying the neutralized (not zero) partial charge models for charged residues(67) to selected residues or, as an extreme model, setting all partial charges of protein residues to zero.

As seen in Table 4.6, the calculated oxidation free energy using PBC (∼75 kcal/mol) is in fact fairly insensitive to these charge perturbations to either the inner or outer regions, unless the immediate ligands of the metal ion (Cys112 and Cys120) are also set to zero (leading to a net charge of -0.12 *e* for the inner region) or all protein residues are set to have zero partial charges. This observation suggests that electrostatic interactions are fairly screened beyond the immediate neighborhood of the metal center; this is also supported by the observation that the GSBP result is not sensitive to charge perturbations in the outer region. Once the net charge of the inner region is reduced, the GSBP calculation becomes better behaved and the difference from the PBC result (without including the P-M and surface polarization corrections) approaches the value observed for charging free energies in solution (i.e., ∼10 kcal/mol). There appears to be a small deviations (∼1 kcal/mol), however, between the PBC/GSBP difference even in the limit of a low inner region charge. For example, when all protein partial charges are set to 0, the uncorrected GSBP/PBC difference in the oxidation free energy is ∼11.8 kcal/mol, which is slightly larger than the difference in PBC/GSBP charging free energies

in a water droplet. This is likely because the presence of the protein reduces the fraction of the box occupied by the solvent, thereby modifying the precise values of both ΔG_{P-M} and ΔG_{s-pol}.

Table 4.6: Oxidation free energy (in kcal·mol^{-1}) of a metal site M in a rigid CueR protein calculated with different net charges using PBC and GSBP boundary conditions, and Poisson-Boltzmann correction for the GSBP results

Net charge distribution[a]	PBC[b]	GSBP (20 Å)	$\Delta G_{PBC,GSBP}$
inner: -8.12, outer: -3.88[c]	-74.5 (1.2[f])	-99.6 (0.9)	25.1 (1.5)
inner: -2.12, outer: -3.88[d]	-74.6 (0.7)	-88.3 (0.6)	13.7 (0.9)
inner: -0.12, outer: -3.88[d]	-69.0 (1.0)	-81.1 (0.5)	12.1 (1.1)
inner: -8.12, outer: 0.12[d]	-75.7 (0.8)	-100.0 (0.7)	24.3 (1.1)
inner: 0.0, outer: 0.0[e]	-28.7 (0.5)	-40.5 (0.7)	11.8 (0.9)
Solvent probe radii	PBC	PB[g]	Difference
0.8 Å/1.4 Å/2.4 Å	-74.5 (1.2)	-85.5/-85.8/-86.0 (1.5)	11.0/11.3/11.5 (1.9)
20 Å[h] droplet	-74.5 (1.2)	-99.5 (1.5)	25.0 (1.9)

[a] Net charge adjustment is realized by neutralizing charged sidechains. Numbers following "inner" and "outer" are the net charges in regions that correspond to the inner and outer partitions in the GSBP simulations. [b] PBC simulations employ the same charge sets as the corresponding GSBP simulations. [c] This net charge distribution is obtained when using the standard CHARMM22 force field; the values are not integers because the inner/outer partitioning in GSBP is done based on geometrical criteria. [d] Selected number of charged residues in the inner/outer regions are neutralized by applying the EEF1 partial charges(67). In the third row (with an inner charge of -0.12), the ligands of the metal site are also neutralized. [e] All partial charges on the protein atoms are set to zero. [f] The values in parentheses are statistical errors. [g] Snapshots used in PB calculations are from GSBP simulations with an inner net charge of -8.12 e and an outer net charge of -3.88 e; the dielectric constant is set to 1 for all explicit solvent/protein atoms and 80 for the bulk (continuum) solvent. [h] A smooth dielectric boundary condition with a radius of 20 Å ($\varepsilon = 1$) is used to mimic the GSBP setup.

As a numerical remedy for the problem associated with the high-charge of the GSBP inner region, PB calculations are carried out to re-evaluate the energy gap for selected snapshots (2,000) from GSBP simulations following the protocol established in the last subsection. Similar to the situation of a water droplet, this PB correction scheme is found to be effective with a fairly small standard deviation of 0.7 kcal/mol. (For more detailes, see **Supporting Information**). With a solvent probe radius of 1.4 Å in the PB calculations, the difference between PB corrected GSBP oxidation free energy and PBC value decreases from $\sim$25 kcal/mol to the expected $\sim$11 kcal/mol. With a smooth dielectric boundary in the PB calculations, however, the correction is essentially zero, emphasizing again the importance of carefully considering the dielectric boundary when

evaluating electrostatic solvation free energy in GSBP calculations with a highly charged inner region.

4.3.4 Oxidation free energy for the metal site in a flexible protein

Finally, as the most realistic example, we study the oxidation free energy of the metal site in a flexible CueR and explore how sensitive is the result to the details of the boundary condition.

To better compare PBC and GSBP, in addition to PBC simulations with a fully flexible protein, we also carry out PBC simulations with the region that corresponds to the outer region of a GSBP set-up held fixed to the crystal structure configuration; atoms beyond 8 Å and 20 Å from the metal site are held fixed (see Sect.4.2 for details), and the simulations are referred to +8Å-fixed and +20Å-fixed PBC simulations, respectively. We first compare these three PBC simulations.

As shown in Table 4.7, the first oxidation free energies ($q_I = +0.1158e, \lambda = 0$ to $q_I = +1.1158e, \lambda = 1$) from the fully flexible and +20Å-fixed PBC simulation are fairly similar to each other; the values are -112.7 and -113.3 kcal/mol, respectively. This suggests that the reorganization of protein atoms and water molecules within 20 Å of the metal site make the dominant contribution while the response from atoms further away is minimal due to effective screening. By contrast, the +8Å-fixed PBC simulation (-90.1 kcal/mol) underestimates the oxidation free energy by as much as 22.5 kcal/mol, suggesting that the proper conformational response to the metal oxidation is not captured in the more restrictive +8Å-fixed PBC setup. By comparing the protein conformation and water distribution in different oxidation states from different PBC simulations, it is found that the response of the metal binding loop to the metal oxidation is too restricted in the +8Å-fixed set-up (see Fig. 4.5c), which in turn also limits the solvent accessibility of the metal site in the oxidized ($\lambda = 1$) state; as shown in Fig. 4.5a, on average one less water is found within the first solvation shell of the oxidized metal site in the +8Å-fix simulation as compared to the fully flexible case.

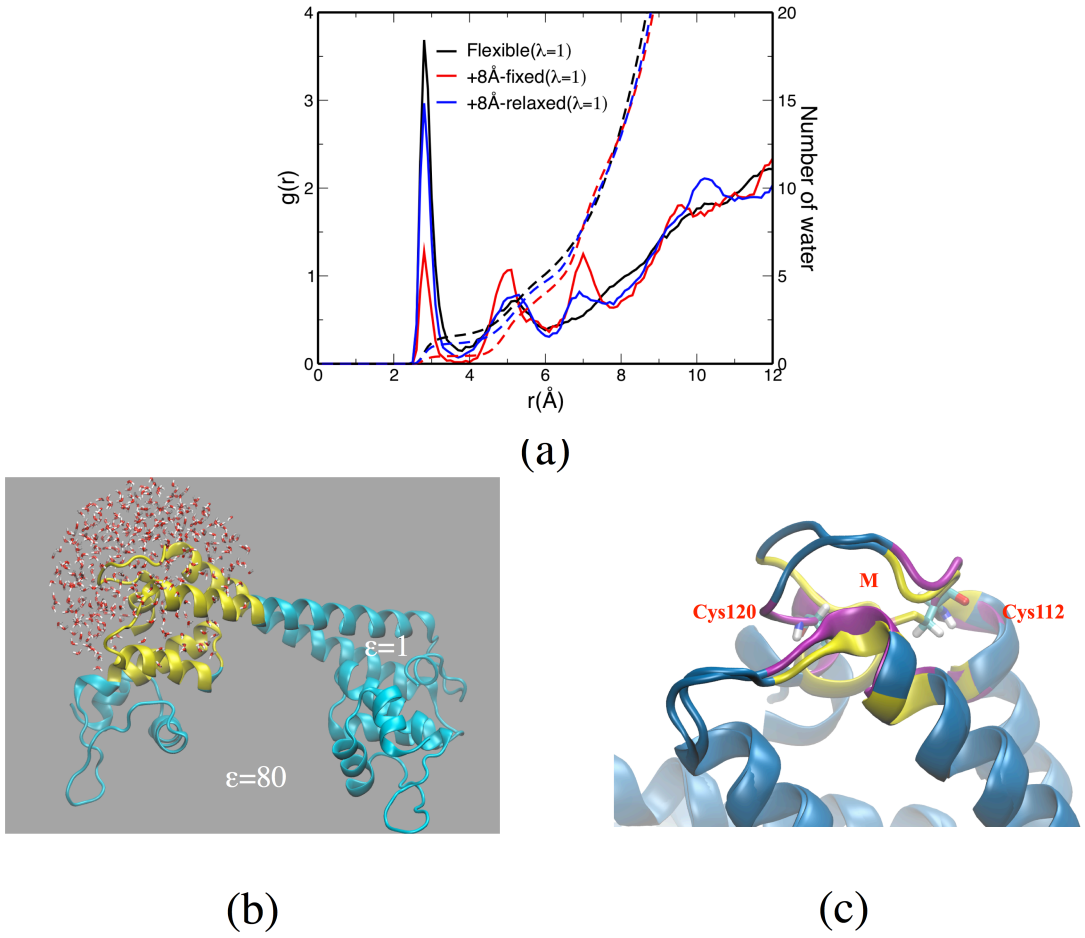


Figure 4.5: Properties of PBC and GSBP simulations with a partially or fully flexible CueR. (a) Radial distribution function of water oxygen around M (oxidized state, $\lambda = 1$) in PBC simulations with different parts of the CueR protein held fixed. (b) A snapshot that illustrates the GSBP (20 Å) setup; protein atoms beyond 20 Å of M are fixed (cyan), those within 20 Å of M are flexible (yellow) and the bulk continuum is indicated with a gray background. (c) Overlay of +8Å-fixed and +8Å-relaxed structures for the oxidize state after 1 ns production run. The flexible protein region from the +8Å-fixed structure is shown in yellow, and that in the +8Å-relaxed structure is shown in purple; the fixed protein region is colored in blue. Clearly, the metal binding loop responds to the metal oxidation and therefore should be allowed to relax for reliable oxidation free energy calculations (see Table 4.7 and text for discussion).

Table 4.7: Oxidation free energy (in kcal·mol^{-1}) of a metal site M in CueR using PBC and GSBP with different parts of protein fixed

System	ΔG ($\lambda = 1 - \lambda = 0$)[a]	ΔG ($\lambda = 2 - \lambda = 1$)[a]	$\Delta\Delta G_{12}$
PBC (fully flexible)	-112.7 (0.8[h])	-255.8 (0.7)	-143.1 (1.1)
PBC (+20Å-fixed[b])	-113.3 (1.3)	-254.7 (1.1)	-141.4 (1.7)
PBC (+8Å-fixed[c])	-90.1 (0.7)	-	-
PBC (+8Å-relaxed[d])	-115.4 (0.6)	-	-
GSBP (20 Å)[e,f]	-132.8/-119.0 (0.7)	-271.4/-259.2 (0.7)	-138.6/-140.2 (1.0)
GSBP (8 Å)[e,g]	-128.9 (0.9)	-273.1 (0.8)	-144.2 (1.2)

[a] $\lambda = 2$ is the "doubly-oxidized" state of M with a partial charge of +2.1158 e, $\lambda = 1$ is the oxidized state with a partial charge of +1.1158 e, and $\lambda = 0$ is the reduced state with a partial charge of +0.1158 e. [b] Fix protein beyond 20 Å of M and all water molecules are mobile. [c] Fix protein beyond 8 Å of M and all water molecules are mobile. [d] An equilibrated structure from the oxidized state of +20Å-fixed simulation is used as the initial structure. [e] The dielectric constant of protein in the outer region is set to 1 for all GSBP simulations. [f] For each entry, the numbers after the slash contain PB correction for the smooth dielectric boundary. [g] The initial structures are equilibrated structures from PBC simulations with the corresponding charge state of M and a fully flexible protein. [h] The values in parentheses are statistical errors based on block average.

To further quantify the importance of these conformational differences, we carry out another set of PBC simulation in which protein atoms beyond 8 Å from the metal site are also held fixed; however, the initial structures in the reduced ($\lambda = 0$) and oxidized ($\lambda = 1$) windows are taken from the equilibrated structures in the respective +20Å-fixed simulations. We refer this new set of simulations as the +8Å-relaxed set, which in principle should have captured most of the conformational response to metal oxidation. As shown in Table 4.7, the +8Å-relaxed PBC result (-115.4 kcal/mol) is indeed much closer to the +20Å-fixed and fully flexible PBC simulations ($\sim$-113 kcal/mol), confirming that it is important to leave flexibility in the model so that rearrangement of nearby structural motifs (e.g., the metal binding loop) occurs without artificial hinderance; it seems, however, less important to sample the thermal fluctuation of atoms beyond 8 Å.

We now move on to GSBP simulations. The comparison of the various PBC results suggests that fixing atoms beyond 20 Å from the metal site has a small impact on the oxidation free energy. Therefore, we expect that GSBP simulations with a 20 Å inner region yield reliable results; the only difference from the +20Å-fixed PBC simulation is the treatment of the bulk solvent and the dielectric boundary of the inner region. Without the PB correction for the smooth dielectric boundary approximation, the GSBP result with a 20 Å inner region differs from the PBC results by about 20 kcal/mol; with the PB correction included, the difference is reduced

to about 6 kcal/mol, which is somewhat different from the GSBP/PBC difference observed for fixed protein simulations (∼12 kcal/mol, see Table 4.6). In other words, it appears that ΔG_{P-M} and ΔG_{s-pol} are not highly transferrable among water and protein systems.

As an additional test along this line, we compute another oxidation free energy in which the metal site is further oxidized ($q_I = +2.1158e, \lambda = 2$) and we compare the difference from the first oxidation free energy using several simulation protocols; the goal is to see to what degree ΔG_{P-M} and ΔG_{s-pol} cancel for the relative oxidation free energy, which is often the quantity of interest, in different computational protocols. As shown in Table 4.7, the $\Delta\Delta G_{12}$ from GSBP (20 Å) simulations after the PB dielectric boundary correction is -140.2 kcal/mol, which is reasonably close to the values from fully flexible (-143.1 kcal/mol) and +20Å-fixed (-141.4 kcal/mol) PBC simulations; even without the PB correction, the $\Delta\Delta G_{12}$ from GSBP (20 Å) is within a few kcal/mol from the PBC results. This is consistent with the finding that solvent distributions near the metal site are similar between the GSBP and those two PBC simulations for both $\lambda = 1$ and $\lambda = 2$ windows (see Fig. 4.6).

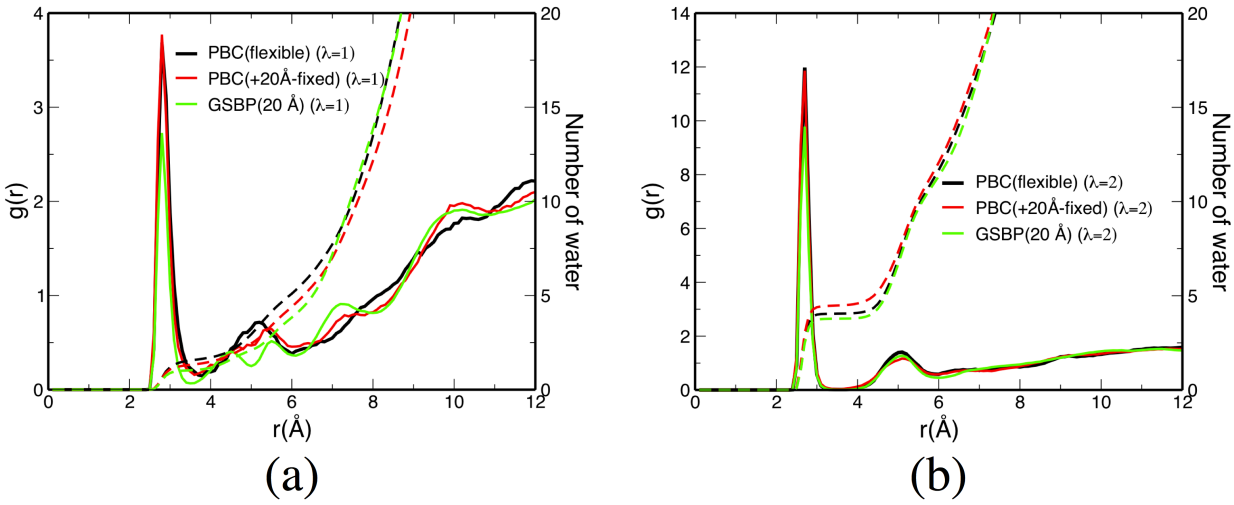


Figure 4.6: Radial distribution function of water oxygen around M in PBC (flexible or 20Å-fixed) and GSBP (20 Å) simulations. (a) M in the oxidized state ($\lambda = 1$). (b) M in the doubly-oxidized state ($\lambda = 2$). Different boundary conditions lead to reasonably similar water distributions.

When a smaller inner region (8 Å) is used for GSBP, analysis of the various PBC results suggests that different structures should be used for different λ windows for reliable results. Therefore, the initial structures

in GSBP (8 Å) are taken from the fully flexible PBC simulations with the corresponding charge state for the metal site. The oxidation free energies obtained from these GSBP (8 Å) simulations are -128.9 kcal/mol for $\Delta G(\lambda = 1 - \lambda = 0)$, -273.1 kcal/mol for $\Delta G(\lambda = 2 - \lambda = 1)$ and therefore a difference of -144.2 kcal/mol for $\Delta\Delta G_{12}$; this is reasonably close to the PBC and GSBP (20 Å) results. However, we note that the absolute values for the oxidation free energies from the GSBP (8 Å) simulations remain fairly different from the PBC results; as shown in Table 4.7, they are in fact somewhat close to the GSBP (20 Å) results without the PB correction for the smooth boundary approximation. Since the inner region of the GSBP (8 Å) simulations do not feature a significant net charge, the effect of the smooth boundary approximation is not significant; test calculations indicate a contribution in the range of 1-2 kcal/mol. Therefore, further studies are needed to elucidate the origin for the larger difference between GSBP (8 Å) and PBC results.

Nevertheless, especially for relative oxidation free energies (i.e., $\Delta\Delta G_{12}$), the GSBP simulations with $\varepsilon_p = 1$ for the outer region leads to good agreement with PBC simulations. This might be due to special features of CueR, which has the metal binding site fairly accessible to the solvent. As a result, the interactions with other charges in the protein are very effectively screened out. For other systems with a deeply buried redox site, a more sophisticated description for the outer region dielectrics might be required. In our study of microscopic pK_a of a deeply buried Glu residue in the transmembrane protein cytochrome c oxidase, for example, the protonation state of titratable residues within 15-20 Å range was found to make notable contributions(19, 68).

4.4 Concluding Remarks

Despite the rapid progress in computational hardwares, it often remains attractive to construct computational models for biomolecules with a finite sphere boundary condition. This allows one to focus on the sampling of a fairly localized region without the need of sampling far away regions. This is clearly most advantageous when the system of interest is very large (e.g., the ribosome) and/or the potential function is expensive (e.g., *ab initio* QM/MM). On the other hand, it should be kept in mind that many approximations are inherent to the finite sphere boundary framework, and whether these approximations make a significant impact on the problem of interest should be evaluated with care. The most obvious issue is whether the process under study implicates collective structural changes or a significant change in the level of solvation in the active site. If so, one should either resort to PBC simulations for better conformational sampling or other specialized simulation techniques to better sample water distributions(69, 70).

In the current work, we focus on the treatment of electrostatic properties by finite sphere boundary condition and compare the results with PBC simulations; electrostatics are of special interest because they often dominate the energetics of processes that occur in biomolecules. Specifically, we compare charging free energies in both solution and protein systems calculated with GSBP and PME simulations. With a change in the net charge, the results of these free energy simulations are sensitive to various approximations and therefore particularly informative to the comparison of simulation protocols.

For simple ions in solution, we find good agreements between GSBP and PBC charging free energies, once the relevant correction terms are taken into consideration. For PBC simulations with PME for long-range electrostatics, the contribution (ΔG_{P-M}) due to the use of a particle rather than molecule based summation scheme in real space is found to be significant, which was first pointed out by Hünenberger and co-workers(36, 49). For GSBP, when the inner region is close to be charge neutral, the key correction is the over polarization of water molecules at the inner/outer dielectric boundary; the magnitude of the correction (ΔG_{s-pol}), however, is relatively minor ($\sim$2 kcal/mol with TIP3P).

For charging (oxidation) free energy in proteins, the situation is more complex although good agreement between GSBP and PBC can still be obtained when care is exercised. For example, we observe that the smooth dielectric boundary approximation inherent to GSBP tends to make significant errors when the inner region is featured with a high net charge. However, the error can be corrected with Poisson-Boltzmann calculations using snapshots from GSBP simulations in a straightforward and robust manner. Due to the more complex charge and solvent distributions, the magnitudes of ΔG_{P-M} and ΔG_{s-pol} in protein simulations appear to be different from those derived for solution simulations. Therefore, it remains somewhat challenging to directly compare absolute charging free energies in PBC and GSBP simulations for protein systems; the relative charging/oxidation free energies, however, are more robust and PBC/GSBP results agree much better even for GSBP simulations with a fairly small inner region.

For the specific protein system (CueR) studied, the use of a low-dielectric "constant" for the outer region is found to introduce only a small error to the oxidation free energy, especially when the inner region is as large as 20 Å; even with a 8 Å inner region, using a dielectric constant of 1 for the outer region leads to a relative oxidation free energy that agrees fairly well with PBC and GSBP (20 Å) simulations. These results are likely due to the fact that the site of interest is fairly solvent accessible and therefore well screened from the

outer region. For a deeply buried site, a more sophisticated treatment for the dielectrics for the outer region(2, 71) might be needed. An important benchmark for the near future, therefore, concerns oxidation/reduction or pK_a calculations in transmembrane proteins. Although several GSBP studies have been carried out for bacteriorhodopsin(19, 20) and cytochrome c oxidase(19, 68), a systematic comparison to PBC simulations and multi-faceted comparisons to available experimental data will be essential to establishing the quantitative accuracy (and limitation) of GSBP for membrane proteins.

Finally, the effect of freezing the protein structure in the outer region is found to be small, unless a very small (8 Å) inner region is used; even in the latter case, the result becomes substantially improved when the nearby metal binding loop is allowed to respond to metal oxidation. In other words, while it is important to leave flexibility in the model so that rearrangement of nearby structural motifs in response to metal oxidation occurs without artificial hinderance, it seems less important to sample the thermal fluctuation of atoms beyond 8 Å. These observations are informative to *ab initio* QM/MM calculations. On one hand, they highlight the importance of carrying out adequate sampling rather than limited structural minimizations. On the other hand, they support a practical approach in which one equilibrates the system using MM or semi-empirical QM/MM with either PBC or large-inner-region GSBP simulations, then carries out *ab initio* QM/MM calculations using GSBP/SMBP(72, 73) but with a smaller inner region.

4.5 Supporting Information

4.5.1 Energy gap distribution

As stated in the main text, LRA is expected to apply if the distribution of the energy gaps closely follow Gaussian statistics.(46) Figure 4.7 shows the energy gap distribution and fitted gaussian plot (with the correlation factor above 99%) for three windows; the behavior is very similar for other windows and therefore not included. These plots indicate that LRA is a valid approximation for our solution and protein systems. Moreover, we have explicitly sampled multiple λ windows for two representative enzymes systems (Figure 4.8); the free energy derivatives fit well with linear relations, with a correlation factor (R^2) of 0.996 for the rigid protein system and 0.986 for the flexible protein case.

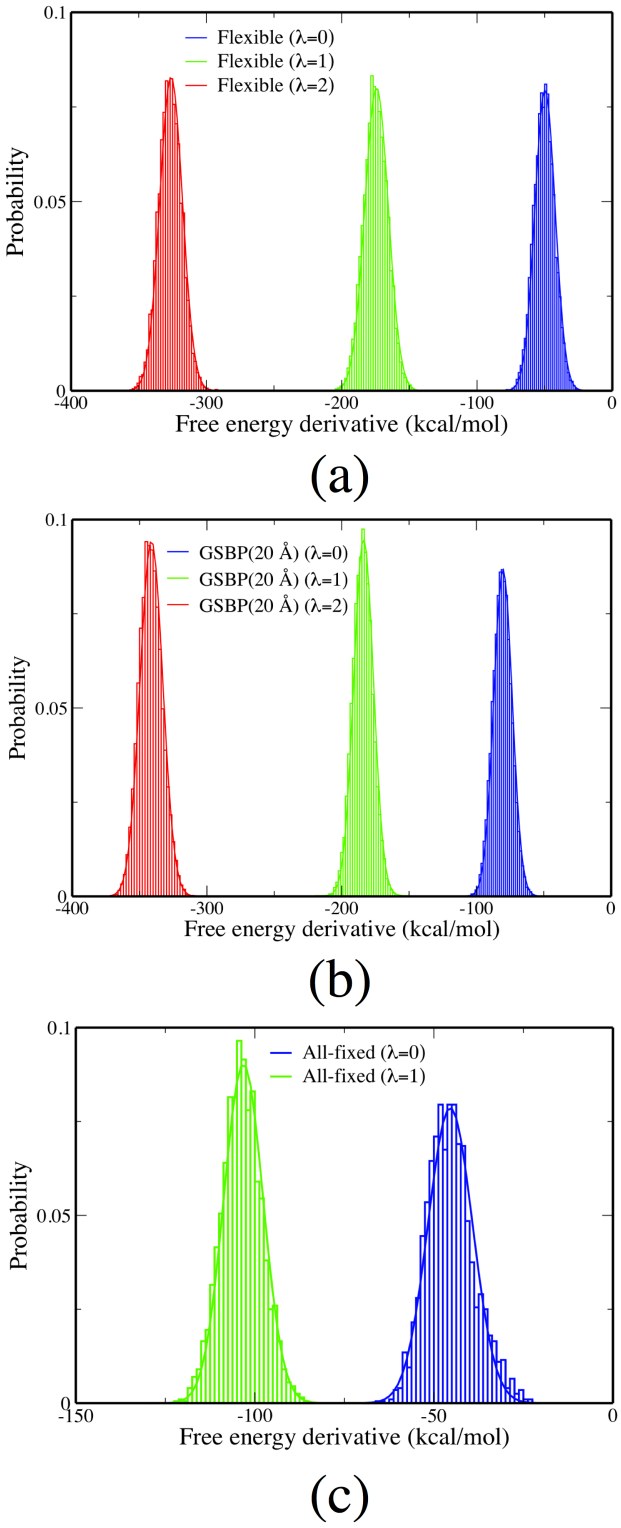


Figure 4.7: Energy gap distributions for $\lambda = 0$, $\lambda = 1$ and $\lambda = 2$ windows in (a) fully flexible PBC, (b) GSBP (20 Å), and (c) rigid protein PBC simulations. For the $\lambda = 1$ windows, the distributions shown correspond to the energy gap between oxidized and reduced states; the distributions for the energy gap between doubly-oxidized and oxidized states also closely follow Gaussian statistics and are not shown here.

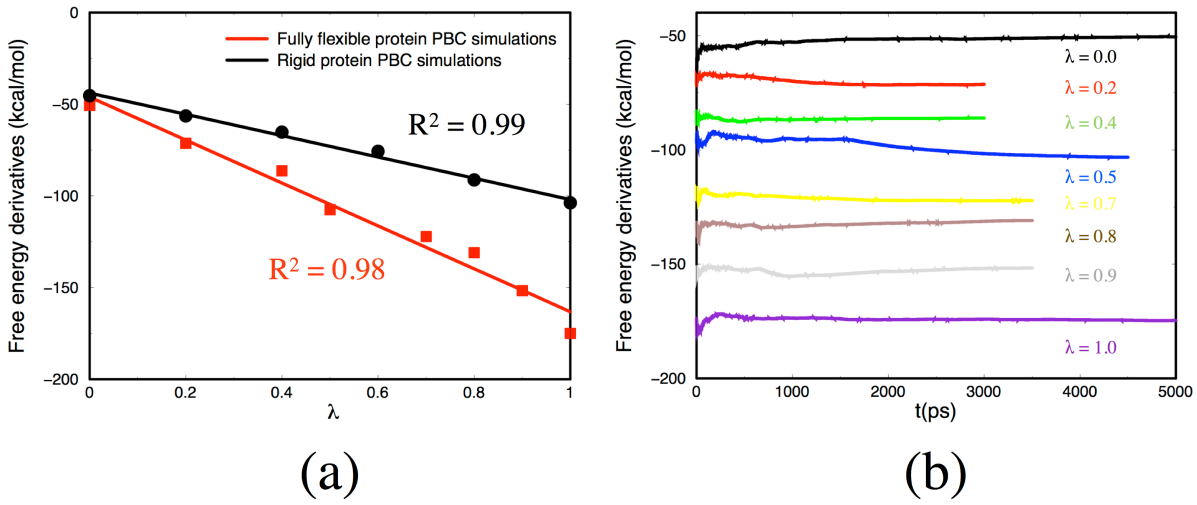


Figure 4.8: (a) Linear fits of the free energy derivatives as functions of λ for the rigid protein and fully flexible protein PBC simulations. (b) Time-dependence of free energy derivatives for different λ windows for the fully flexible protein PBC simulation.

4.5.2 Statistical for the Poisson-Boltzmann (PB) correction for the smooth dielectric boundary approximation in GSBP

The statistics for the PB correction for the smooth dielectric boundary approximation in GSBP (20 Å inner region) are summarized in Table 4.8. 1000 snapshots from solution (Na^+ + 5 Cl^-) simulations are selected to re-evaluate energy gap by PB calculations; 2000 and 6000 snapshots are used for the rigid protein simulations and partially flexible protein simulations, respectively. As can be seen from Table 4.8 and Figure 4.9, the statistical deviation is about ±1 kcal/mol for all the cases. For the solution and rigid protein simulations, randomly picking 200 snapshots for each window also gives a reasonable energy gap correction, while for the partially flexible protein simulation, as few as 300 configurations are found to give reliable PB correction (standard deviation ∼ ±1 kcal/mol).

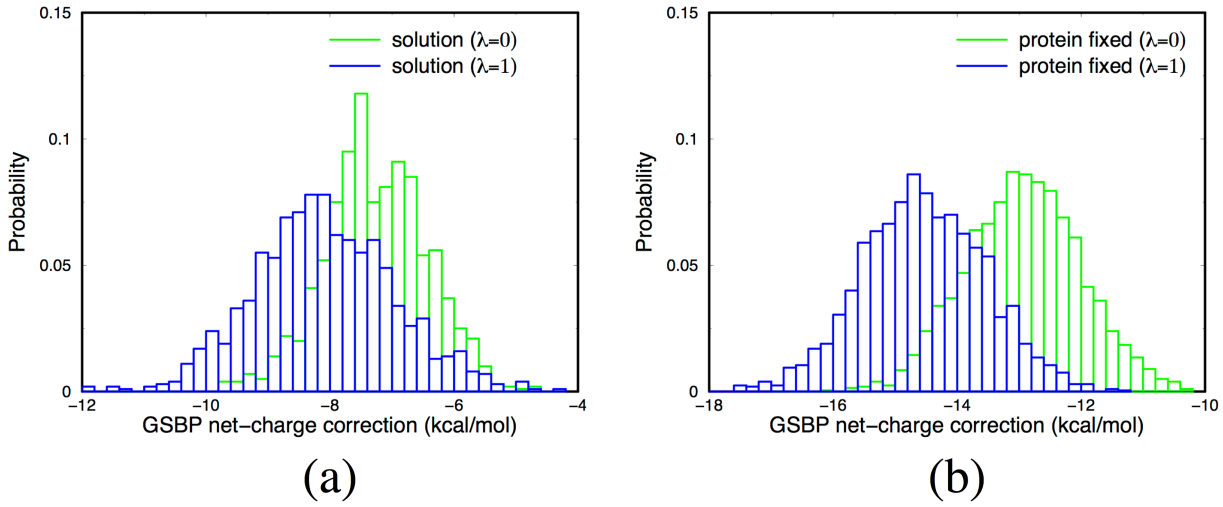


Figure 4.9: The distribution of PB corrections for the energy gaps from two representative windows of GSBP(20 Å) simulations. (a) charging free energy of Na^+ in the presence of five Cl^-, (b) rigid protein simulations. For the partially flexible protein simulations, the distributions for the PB corrections are similar and therefore not shown.

Table 4.8: Standard deviation of Poisson-Boltzmann (PB) correction for the energy gaps from different λ windows of GSBP (20 Å) simulations and for the charging/oxidation free energy

System	$\lambda = 0$	$\lambda = 1$[c]	$\Delta G\ (\lambda = 1 - \lambda = 0\ )$
Na^+ in solution[a]	0.8	1.1	0.7
Rigid protein	0.9	1.0	0.7
Partially flexible protein[b]	1.0	1.2	0.8
System	$\lambda' = 1$[d]	$\lambda = 2$	$\Delta G\ (\lambda = 2 - \lambda = 1\ )$
Partially flexible protein	1.2	1.5	1.0

[a] Na^+ in a water droplet in the presence of five Cl^-; 1000 snapshots are used. [b] Protein atoms within 20 Å inner region sphere are flexible; 6000 snapshots are used. [c] For protein simulations, the energy gap for $\lambda = 1$ is between oxidized ($\lambda = 1$)and reduced ($\lambda = 0$) states [d] The energy gap is that between the doubly-oxidized ($\lambda = 2$) and oxidized ($\lambda = 1$) states

References

[1] Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144–1149.

[2] Warshel, A.; Sharma, P. K.; Kato, M.; Parson, W. W. *Biochim. Biophys. Acta* **2006**, *1764*, 1647–1676.

[3] Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Acad. Natl. Sci. USA* **2001**, *98*, 10037–10041.

[4] Frenkel, D.; Smit, B. *Understanding Molecular Simulations: from Algorithms to Applications*; Academic Press, San Diego, 1996.

[5] Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[6] Kuwajima, S.; Warshel, A. *J. Chem. Phys.* **1988**, *89*, 3751–3759.

[7] Bogusz, S.; Cheatham, T. E.; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070–7084.

[8] Hünenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856–1872.

[9] Weber, W.; Hünenberger, P. H.; McCammon, J. A. *J. Phys. Chem. B* **2000**, *104*, 3668–3675.

[10] King, G.; Lee, F. S.; Warshel, A. *J. Chem. Phys.* **1991**, *95*, 4366–4377.

[11] Schutz, C. N.; Warshel, A. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 400–417.

[12] Warshel, A.; Dryga, A. *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 3469–3484.

[13] Im, W.; Bernèche, S.; Roux, B. *J. Chem. Phys.* **2001**, *114*, 2924–2937.

[14] Schaefer, P.; Riccardi, D.; Cui, Q. *J. Chem. Phys.* **2005**, *123*, 014905 1–14.

[15] Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458–6469.

[16] Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2008**, *4*, 1600–1609.

[17] Ge, X.; Roux, B. *J. Phys. Chem. B* **2010**, *114*, 9525–9539.

[18] Yang, Y.; Yu, H.; Cui, Q. *J. Mol. Biol.* **2008**, *381*, 1407–1420.

[19] Ghosh, N.; Prat-Resina, X.; Gunner, M. R.; Cui, Q. *Biochemistry* **2009**, *48*, 2468–2485.

[20] Goyal, P.; Ghosh, N.; Phatak, P.; Clemens, M.; Gaus, M.; M. Elstner,; Cui, Q. *J. Am. Chem. Soc.* **2011**, *133*, 14981–14997.

[21] Banavali, N. K.; Im, W.; Roux, B. *J. Chem. Phys.* **2002**, *117*, 7381–7388.

[22] Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.

[23] Riccardi, D.; Cui, Q. *J. Phys. Chem. A* **2007**, *111*, 5703–5711.

[24] Simonson, T.; Perahia, D. *J. Am. Chem. Soc.* **1995**, *117*, 7987–8000.

[25] Simonson, T.; Brooks III, C. L., *J. Am. Chem. Soc.* **1996**, *118*, 8452–8458.

[26] Goh, G. B.; B. Garcia-Moreno E.,; Brooks III, C. L., *J. Am. Chem. Soc.* **2011**, *133*, 20072–20075.

[27] Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.

[28] Figueirido, F.; Buono, G. S. D.; Levy, R. M. *J. Chem. Phys.* **1995**, *103*, 6133–6142.

[29] Hummer, G.; Pratt, L. R.; García, A. E. *J. Phys. Chem. B* **1997**, *101*, 3017–3020.

[30] Hummer, G.; Pratt, L. R.; García, A. E.; Garde, S.; Berne, B. J.; Rick, S. W. *J. Phys. Chem. B* **1998**, *102*, 3841–3843.

[31] Åqvist, J.; Hansson, T. *J. Phys. Chem. B* **1998**, *102*, 3837–3840.

[32] Vorobjev, Y. N.; Hermans, J. *J. Phys. Chem. B* **1999**, *103*, 10234–10242.

[33] Babu, C. S.; Yang, P.-K.; Lim, C. *J. Biol. Phys.* **2002**, *2*, 95–113.

[34] Ashbaugh, H. S.; Wood, R. H. *J. Chem. Phys.* **1997**, *106*, 8135–8139.

[35] Darden, T.; Pearlman, D.; Pedersen, L. G. *J. Chem. Phys.* **1998**, *109*, 10921–10935.

[36] Kastenholz, M. A.; Hünenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 124106 1–27.

[37] Reif, M. M.; Hünenberger, P. H.; Oostenbrink, C. *J. Chem. Theo. Comp.* **2012**, *8*, 3705–3723.

[38] Yu, H.; Mazzanti, C. L.; Whitfield, T. W.; Koeppe, R. E.; Andersen, O. S.; Roux, B. *J. Am. Chem. Soc.* **2010**, *132*, 10847–10856.

[39] Li, G.; Zhang, X.; Cui, Q. *J. Phys. Chem. B* **2003**, *107*, 8643–8653.

[40] Olsson, M. H. M.; Hong, G.; Warshel, A. *J. Am. Chem. Soc.* **2003**, *125*, 5025–5039.

[41] Cascella, M.; Magistrato, A.; Tavernelli, I.; Carloni, P.; Rothlisberger, U. *Proc. Nat. Acad. Sci. USA* **2006**, *103*, 19641–19646.

[42] Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

[43] Simonson, T. *J. Phys. Chem. B* **2000**, *104*, 6509–6513.

[44] Woo, H. J.; Roux, B. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6825–6830.

[45] Levy, R. M.; Belhadj, M.; Kitchen, D. B. *J. Chem. Phys.* **1991**, *95*, 3627–3633.

[46] Simonson, T. *Proc. Acad. Natl. Sci. U.S.A.* **2002**, *99*, 6544–6549.

[47] Ghosh, N.; Cui, Q. *J. Phys. Chem. B* **2008**, *112*, 8387–8397.

[48] Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solution*; Wiley, New York, 1991.

[49] Kastenholz, M. A.; Hünenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 224501 1–20.

[50] Mackerell, A. D. et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

[51] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

[52] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel, Dordrecht, 1981.

[53] Andersen, H. C. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

[54] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[55] Stote, R. H.; States, D. J.; Karplus, M. *J. Chim. Phys.* **1991**, *88*, 2419–2433.

[56] Nosé, S. *J. Chem. Phys.* **1984**, *81*, 511–519.

[57] Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.

[58] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. *J. Comp. Phys.* **1977**, *23*, 327–341.

[59] Štrajbl, M.; Hong, G.; Warshel, A. *J. Phys. Chem. B* **2002**, *106*, 13333–13343.

[60] Changela, A.; Chen, K.; Xue, Y.; Holschen, J.; Outten, C. E.; O'Halloran, T. V.; Mondragón, A. *Science* **2003**, *301*, 1383–1387.

[61] Hobman, J. L. *Mol. Microbiol.* **2007**, *63*, 1275–1278.

[62] Rao, L.; Cui, Q.; Xu, X. *J. Am. Chem. Soc.* **2010**, *132*, 18092–18102.

[63] Brooks, B. R. et al. *J. Comp. Chem.* **2009**, *30*, 1545–1614.

[64] Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59–75.

[65] Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.

[66] Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050–9063.

[67] Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Genet.* **1999**, *35*, 133–152.

[68] Goyal, P.; Cui, Q. *ms in preparation* **2012**,

[69] Woo, H.-J.; Dinner, A. R.; Roux, B. *J. Chem. Phys.* **2004**, *121*, 6392–6400.

[70] Chakrabarty, S.; Warshel, A. *Proteins: Struct., Funct., Bioinf.* **2012**, *In press*.

[71] Song, X. *J. Chem. Phys.* **2002**, *116*, 9359–9363.

[72] Benighaus, T.; Thiel, W. *J. Chem. Theory Comput.* **2009**, *5*, 3114–3128.

[73] Zienau, J.; Cui, Q. *J. Phys. Chem. B* **2012**, *116*, 12522–12534.

Chapter 5

Recent Progress and Challenges in QM/MM Free Energy Simulations[1]

5.1 Introduction

Central to most QM/MM applications is the computation of free energy differences. In the context of enzyme catalysis, for example, the relevant quantities(27) include the binding affinity of the substrate (often related but not necessarily equivalent to the Michealis constant K_M) and the free energy profile for the chemical step(s), which contain information about k_{cat}; to understand the proficiency of enzyme catalysis, it is also important to determine the free energy profile for the uncatalyzed reaction(s) in solution. Accordingly, two types of free energy simulations are involved: alchemical simulations for binding and potential of mean force (PMF) simulations for chemical reactions. The division of the techniques is certainly not absolute; for example, PMF type of simulations can also be used to compute the absolute binding free energy of ligands to protein receptors(28).

In this contribution, we briefly review the two types of QM/MM free energy simulations with most examples motivated by our own research. The goal is to illustrate state-of-the-art techniques, recent developments and technical issues relevant for realistic applications using model systems. We focus on problems in which a direct sampling of the QM/MM potential function is important. For direct sampling, it is crucial to de-

[1]Reprinted (adapted) with permission from "QM/MM free energy simulations: recent progress and challenges", X. Lu, D. Fang, S. Ito, Y. Okamoto and Q. Cui, *Mol. Simul.* **42**, 1056 (2016). Copyright 2016 Informa UK Limited, trading as Taylor & Francis Group.

velop a low-level QM/MM potential that provides a semi-quantitative description of the underlying potential energy surface. One approach is to develop system-dependent semi-empirical QM/MM potentials, based on either minimum energy paths(45–47), force-matching(48) or para-dynamics(49). The alternative approach, which is what we follow, is to develop a general-purpose approximate QM method so that refitting is (ideally) not required when studying a new system. In particular, we have been pursuing an approximate density functional theory, density functional tight binding (DFTB(50, 51)), for biologically relevant systems. As reviewed elsewhere recently(52–54), the latest version of the DFTB methodology, referred to as DFTB3(55), provides encouraging results for a fairly broad class of systems of biological interest; the accuracy is often comparable to density functional theory with the generalized gradient approximation (DFT/GGA) and a double-zeta-plus-polarization quality basis set, while the computational cost is similar to conventional semi-empirical methods(56) such as AM1 and PM3, making it routine to conduct nano-second simulations based on DFTB3/MM potentials. For metalloenzyme applications, recent developments have led to promising parameterizations for several metal ions that include the alkali metals(57), magnesium, zinc(58) and copper(59). The DFTB3 method in the current form is most reliable for structural properties, including for fairly complex bi-metallic motifs in several enzymes(40, 41, 58, 60, 61); for energetic properties, the results are less robust as compared to DFT/GGA(62, 63) but can often be improved to satisfying accuracy with single point energy calculations at high QM(/MM) level, making DFTB3 a promising low-level approach in dual-level QM/MM free energy simulations, a topic that we will also discuss here.

In the following, we first summarize the basic theoretical foundation of the two types of QM/MM free energy simulations and relevant techniques (e.g., metadynamics and finite-temperature string) for various applications. Next, we discuss a number of condensed phase model systems to illustrate these free energy methods and relevant technical issues. Finally, we draw a few conclusions and highlight a few pressing challenges for future investigations.

5.2 Theory and Methods

In this section, we first discuss methods for computing reaction free energy profiles and pathways, and then move on to alchemical free energy computations, including multi-level free energy calculations that integrate low- and high-level QM(/MM) potential functions.

5.2.1 Reaction free energy profile and pathways

5.2.1.1 Sampling along pre-determined order parameters

For relatively simple chemical reactions, such as localized proton/atom transfers, umbrella sampling(64) along one or several pre-defined order parameter(s) is likely an effective approach. These order parameters are usually geometrical parameters, such as an antisymmetric stretch involving the atom being transferred and the respective donor and acceptor atoms(65); in some applications, energetic parameters such as the energy gap between the relevant diabatic states(66, 67) are also useful. The potential function in umbrella sampling (U_{US}) takes the following form, which includes both the unbiased potential energy (U_0) and the umbrella potential (U_b),

$$U_{US}^{(i)}(\mathbf{X}) = U_0(\mathbf{X}) + U_b^{(i)}(\xi(\mathbf{X})). \tag{5.1}$$

The umbrella potential of the *i*th window typically takes the form of a harmonic restraint on the order parameter ξ (i.e., $k^{(i)}(\xi - \xi^{(ref,i)})^2$), although more complex functional forms or numerical values can be used in, for example, adaptive umbrella sampling(68); the bias can also take the form of a constraint(69, 70) rather than a restraint, thus fixing the value of the order parameter to be specific reference values, $\xi^{(ref,i)}$. When a restraining potential is used, statistics of the order parameter(s) from all windows are collected and analyzed using, typically, Weighted Histogram analysis (WHAM)(71) to generate the underlying PMF, $W(\xi)$ (Eq. 6.3). In constrained simulations(69), the mean gradient of the order parameter(s), $\langle \nabla_\xi (U_0 - \beta^{-1} \log |J|) \rangle_\xi$, is integrated to generate $W(\xi)$ (where $\langle \cdot \rangle_\xi$ is the conditional average $\int d\mathbf{X}(\cdot) \exp[-\beta U_0] \delta(\xi(\mathbf{X}) - \xi) / \int d\mathbf{X} \exp[-\beta U_0] \delta(\xi(\mathbf{X}) - \xi)$, β is the inverse temperature, and we have formally included the Jacobian of the coordinate transformation, but left the transformation unspecified). Because it is straightforward to decompose atomic contributions to (mean) force when pair-wise MM potential or QM-MM interactions are used, constrained simulations are

valuable for decomposing atomic contributions(72) to $W(\xi)$, provided that care is exercised to include the Jacobian contribution(73, 74).

As discussed in recent studies, the WHAM approach relies on the assumption that data from different windows are drawn from globally equilibrated simulations, thus care needs to be exercised to design the restraining potentials and to monitor convergence(75); a promising alternative is to analyze the data using a discrete transition based reweighting analysis (dTRAM(76), a similar idea was published in Ref.(75)), which provides maximum-likelihood estimates of free energy quantities without the assumption of global equilibrium. Another technical issue often encountered in umbrella sampling for complex system is that the degrees of freedom orthogonal to the order parameters are sampled differently in different windows. One promising approach to improve the convergence is replica-exchange umbrella sampling (REUS(77, 78)), which is a variant of the Hamiltonian replica-exchange method(77, 79). In REUS, different umbrella windows are run in parallel; at a given frequency, an attempt is made to exchange configurations from two windows ($m \leftrightarrow n$) with the acceptance probability following a Metropolis criterion:

$$P[(\mathbf{X}_m, \mathbf{X'}_n) \rightarrow (\mathbf{X'}_m, \mathbf{X}_n)] = \begin{cases} 1 & \Delta \leq 0 \\ e^{-\Delta} & \Delta > 0 \end{cases}, \tag{5.2}$$

where $\Delta = \beta[U_b^m(\xi(\mathbf{X'})) - U_b^m(\xi(\mathbf{X})) - U_b^n(\xi(\mathbf{X'})) + U_b^n(\xi(\mathbf{X}))]$. If there are significant barriers along the orthogonal directions for all important ξ values, REUS alone does not help overcome these barriers; when the orthogonal barrier is only high in certain ξ windows, however, REUS is expected to be effective(80). For the acceptance probability to fall in the reasonable range, exchange is usually attempted for neighboring windows.

Another popular variant of the adaptive umbrella sampling technique is metadynamics(81–83), in which Gaussian potentials are deposited with a time interval of τ_G along a molecular dynamics simulation. The Gaussian potentials, with width δ_G and height ω_G, are functions of the order parameters, which are referred to as the collective variables (CVs) in the metadynamics literature. Addition of the Gaussian functions elevates the potential wells and eventually facilitates transition over barriers in the CV space. In the long time limit, when the dynamics in the CV space becomes diffusive, the accumulated Gaussian potential,

$$V_G(\xi, t) = \omega_G \sum_{t'=\tau_G, 2\tau_G, \cdots, t} \exp\left\{-\frac{[\xi - \xi(\mathbf{X}(t'))]^2}{2\delta_G^2}\right\} \tag{5.3}$$

approaches the negative PMF in the CV space, i.e., $\lim_{t\to\infty} V_G(\xi, t) \sim -W(\xi)$; it is straightforward to generalize Eq.5.3 to multi-dimensional cases. In addition to the choice of the CVs, which is a common issue to

any form of umbrella sampling technique, the convergence behavior of metadynamics simulations has been thoroughly discussed in the literature(81, 84), including the choice of the key parameters (δ_G, ω_G and τ_G). Approximate expressions have been derived for metadynamics, although in practice, the error in the computed PMF profile is more conveniently estimated by comparing independent runs or using block averaging. In well-tempered metadynamics(85), the height of the Gaussian potential is automatically scaled as the simulation proceeds, which leads to a smoother convergence behavior(84),

$$\omega_G(t) = \omega_0 \tau_G \exp\left[-\frac{V_G(\xi,t)}{k_B \Delta T}\right]. \tag{5.4}$$

Here ΔT is an input parameter, which determines the long-time limit of the accumulated Gaussian potential,

$$lim_{t\to\infty} V_G(\xi,t) \sim -\frac{\Delta T}{T+\Delta T} W(\xi). \tag{5.5}$$

For $\Delta T = 0$, regular MD is recovered, while the regular metadynamics corresponds to the limit of $\Delta T \to \infty$.

The challenge for studying reactions in a complex environment is that many degrees of freedom are likely involved. If a large number of order parameters/CVs are chosen, umbrella sampling and metadynamics simulations will be prohibitively expensive for thoroughly exploring the multi-dimensional PMF. One possible solution is to maintain a small number of CVs, but couple metadynamics with replica-exchange/simulated tempering simulations to enhance the sampling of the orthogonal or transverse degrees of freedom(81). Another choice is bias exchange metadynamics(86), in which replica-exchange is conducted for multiple metadynamics simulations that bias different sets of CVs. The other alternative is to search for a minimum free energy path rather than exploring the entire multi-dimensional PMF surface. In this case, only a one-dimensional object is sought for, regardless of the number of CVs used to parameterize the path, thus the computation formally scales much more favorably compared to multi-dimensional umbrella sampling/metadynamics. We turn to this subject in the next section.

5.2.1.2 Free energy path sampling

The search of minimum energy path (MEP), also known as the intrinsic reaction coordinate(87), has a long history in quantum chemistry. For reactions in the liquid phase or biomolecules, which feature very rugged potential energy surfaces, a small number of MEPs are often insufficient for a quantitative description of the reaction(9, 45, 88), since the potential energy profiles can vary greatly among them and it's not trivial to evaluate the weights of these MEPs in terms of contribution to the reaction flux at a finite temperature(89).

The alternative is to determine the path(s) on a free energy surface. Again, there has been many developments motivated by somewhat different applications, and here we focus on those most relevant to QM/MM type of simulations; we limit the discussion to pathways in the configurational space only, as sampling in the trajectory space (transition path sampling(90)) is most important for understanding dynamical (non-equilibrium) effects associated with a reaction. In this context, we again first define the CV space spanned by a set of order parameters most relevant to the reaction process: $\vec{\xi}(\mathbf{X}) = (\xi_1(\mathbf{X}), \xi_2(\mathbf{X}), \cdots, \xi_n(\mathbf{X}))$; the CVs can range from a combination of internal coordinates to a set of Cartesian coordinates for the most essential atoms. Then the free energy surface (or PMF) in this n-dimensional space is defined as,

$$W(\vec{\xi}') = -\beta^{-1} ln\Big[Z^{-1} \int e^{-\beta U(\mathbf{X})} \prod_{i=1}^{n} \delta(\xi_i' - \xi_i(\mathbf{X})) d\mathbf{X}\Big], \tag{5.6}$$

where Z is the configuration integral. The minimum free energy path (MFEP) on this surface is a curve (or a "string") defined by the condition that the projection of the mean force perpendicular to the path vanishes,

$$\Big[\mathbf{M}(\vec{\xi}) \cdot \nabla W(\vec{\xi})\Big]^{\perp} = 0, \tag{5.7}$$

where $M_{ij} = \langle \nabla_{X^*} \xi_i \cdot \nabla_{X^*} \xi_j \rangle_{\vec{\xi}}$ is the metric tensor associated with the transformation to the space of collective variables, and $\mathbf{X}^*$ are mass-weighted coordinates. As discussed in detail in Ref.(91), the MFEP is the path of maximum likelihood for the reaction if two conditions are met: (i). the set of collective variables is adequate to approximate the committor function, which is taken to be the *true* reaction coordinate that describes the mechanism of the reaction(90, 92–96); (ii). most reactive trajectories are contained in narrow 'transition' tubes, which are separated by significant energy barriers.

In practical implementations, the MFEP is optimized with the chain-of-states methods(97), similar to the nudged elastic band(98) and its variations for MEP computations. The string is discretized into a set of N equidistant points (or images(91)) $\vec{\xi}_i$, with $i \in [1 \ldots N]$ and to each point a separate simulation system is assigned, which is used to compute the metric tensor $M(\vec{\xi})$ and the free energy gradient $\nabla W(\vec{\xi})$ that appear in Eq. 6.1. Starting from an initial condition for the string $\vec{\xi}_i^0$, the points $\vec{\xi}_i^m$ are updated iteratively until a discretized version of condition Eq. 6.1 is satisfied. A straightforward iterative method is to first advance the string using steepest descent:(99)

$$\vec{\xi}_i^* = \vec{\xi}_i^m - \frac{\Delta t}{\gamma} \mathbf{M}(\vec{\xi}_i^m) \nabla W(\vec{\xi}_i^m), \tag{5.8}$$

followed by reparametrization(91), i.e., an interpolation step that enforces equal distances between adjacent

images:

$$\hat{\xi}_i^{m+1} = \hat{\xi}_i^* + \lambda_i \vec{\tau}_i. \tag{5.9}$$

In Eq. (5.9), $\vec{\tau}_i$ is the vector tangent to the string at $\vec{\hat{\xi}}_i^*$, and λ_i are chosen such that $|\vec{\xi}_{i+1}^{m+1} - \vec{\xi}_i^{m+1}|$ are equal for $i \in \{1 \ldots N-1\}$.(99, 100).

$\mathbf{M}$ and ∇W in Eq. (6.2) are computed from simulations with harmonic restraints $k \times [\vec{\xi}(\mathbf{X}) - \vec{\hat{\xi}}_i]^2$, and the parameters Δt and γ are adjusted empirically to accelerate convergence to the MFEP.(91, 100) For the special case of Δt=dt, where dt is the time step in the molecular simulation used to compute $\mathbf{M}$ and ∇W, the string was shown to converge to the MFEP rapidly,(101, 102) provided that the friction γ was taken sufficiently large. However, in this case the optimized path satisfies the condition

$$\left[\mathbf{M}(\vec{\xi}) \cdot \nabla W(\vec{\xi}) - \beta^{-1} \nabla \cdot \mathbf{M}(\vec{\xi})\right]^{\perp} = 0, \tag{5.10}$$

which differs from Eq. (6.1) by the temperature-dependent term $\beta^{-1}\nabla \cdot \mathbf{M}$. The differences in the paths and free energies based on Eqs. (6.1) and (5.10) were found to be small for the solvated alanine dipeptide.(101, 102). It is noteworthy that the reparametrization step of Eq. (5.9) is a relatively expensive collective computation involving all points $\vec{\hat{\xi}}_i$ (i.e., possibly all processors), and thus a larger Δt is typically chosen to avoid performing this step too frequently.

An indirect approach to approximate the solution to Eq. (5.10) is the swarm-of-trajectories string method (STS).(103) In STS, a collection of independent unbiased trajectories of (usually short) duration Δt are launched from configurations of the molecular systems conditioned on $\vec{\xi}(\mathbf{X})$=$\vec{\hat{\xi}}_i^m$, and the evolution in Eq. (6.2) is replaced by

$$\hat{\xi}_i^* = \hat{\xi}_i^m + \alpha \langle \vec{\xi}(\mathbf{X}(\Delta t)) - \hat{\xi}_i^m \rangle_{\vec{\xi}(\mathbf{X}(0))=\vec{\hat{\xi}}_i^m}, \tag{5.11}$$

where α is a scaling parameter that can be used to optimize convergence and accuracy (α=1 is a typical starting choice). $\vec{\hat{\xi}}_i^{m+1}$ are then obtained from Eq. (5.9). To iterate further, one must recondition the molecular systems to satisfy $\vec{\xi}(\mathbf{X}) \simeq \vec{\hat{\xi}}^{m+1}$, which can be achieved using e.g., restrained simulations. Although the original STS approach assumed diffusive dynamics of the CV variables under which condition the optimized path is the most probable transition path (i.e., the system has the highest probability of being on the path), it was subsequently shown that the approach is also valid when the CV dynamics is in the inertia regime(102, 104), and the path converges to MFEP.

In practical applications of the string methods, the most important decisions include the choice of the CVs, the initial string ($\vec{\xi}^0$) and the values of the restraining force constants. The choices are often not straightforward and can influence results significantly. We will discuss some of the issues using examples in Sect.5.3.1.2.

Finally, we note that an entirely different approach is to use enhanced sampling techniques to let the QM region overcome chemical barriers without biasing specific geometrical coordinates. Then the free energy surface can be projected onto desirable CVs for analysis. Several attempts have been made in the past under the name of "chemical flooding"(107). The most recent example featured the use of selective integrated tempering sampling (SITS(108)) for the QM atoms in a QM/MM framework. In ITS(109), an effective Hamiltonian is constructed by integrating the distributions at multiple temperatures, and sampling using this effective Hamiltonian allows overcoming enthalpic barriers; in SITS(108), only the effective temperature for selected degrees of freedom and their coupling to the environment is scaled, similar to solute tempering(110), allowing the enhancement of sampling for specific degrees of freedom. The method has been successfully used to study Claisen rearrangement in different solutions(111). In the future, it is interesting to explore whether this promising method remains effective when the QM region is large and therefore there are many ways to partition the energy boost introduced through the higher temperature distributions.

5.2.2 Alchemical and multi-level free energy computations

5.2.2.1 Alchemical free energy simulations

When only the relative free energy of two species/states, rather than the path that connects them, is of interest, alchemical free energy simulations(112–114) are often used. Specifically for QM/MM applications, typical examples include the solvation free energy of a QM solute, and the binding affinity of a QM ligand to a protein active site; in the first case, the two states correspond to the QM solute in the gas phase and water, respectively, while in the latter case, the two states are the QM ligand in water and in the protein active site, respectively. Alternatively, we might be interested in the relative solvation free energy of two QM solutes in water, or the relative binding affinity of two QM ligands to the same protein active site. To minimize the cost of sampling, QM/MM alchemical free energy simulations are often carried out using a thermodynamic cycle pioneered by Gao(5) and Warshel(66) in which pure MM simulations are used as a reference; an example is described in Fig. 5.1a, which describes the solvation free energy calculation of a QM solute.

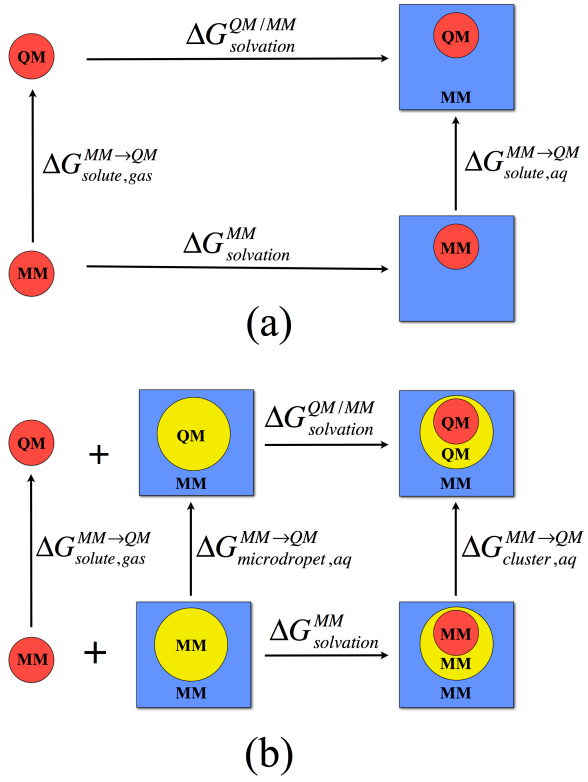


Figure 5.1: Thermodynamic cycles used to compute the solvation free energy of a QM solute in (a) a pure MM environment and (b) when nearby solvent molecules are also described at the QM level. In both cases, a pure MM solvation free energy calculation (bottom horizontal process) is used as a reference; to obtain the desired QM solvation free energies, various vertical processes need to be studied in which the solute (colored red) or a micro droplet (colored yellow) is converted between a MM and a QM model either in the gas phase or in an MM solution environment (in blue).

Following the thermodynamic cycle, the step that requires most sampling is done with an inexpensive MM potential ($\Delta G^{MM}_{solvation}$), while "alchemical mutations" are used to convert the solute between MM and QM representations; such conversion is expected to converge quickly due to the generally small differences between QM and MM potentials for stable species. Using the state nature of free energy, it is easy to see,

$$\Delta G^{QM/MM}_{solvation} = \Delta G^{MM}_{solvation} + (\Delta G^{MM\rightarrow QM}_{solute,aq} - \Delta G^{MM\rightarrow QM}_{solute,gas}). \tag{5.12}$$

This type of thermodynamic cycle can be easily generalized to connect QM descriptions at different levels as well, which we discuss in the next subsection.

For highly polar and charged solutes, it might be preferable to treat also the first solvation shell as QM. In this case, the thermodynamic cycle only requires a minimal degree of revision, as shown in Fig. 5.1b. The additional step involves converting the microdroplet, which corresponds to the first solvation shell solvent molecules in the absence of the solute, between MM and QM representations,

$$\Delta G^{QM/MM}_{solvation} = \Delta G^{MM}_{solvation} + (\Delta G^{MM\rightarrow QM}_{cluster,aq} - \Delta G^{MM\rightarrow QM}_{solute,gas} - \Delta G^{MM\rightarrow QM}_{microdroplet,aq}). \tag{5.13}$$

In terms of computational cost, the alchemical steps that involve MM/QM conversion are most likely to dominate. When QM is an *ab initio* or DFT method, it is desirable to minimize the degree of QM sampling, a topic we discuss in the next subsection. When the QM is a semi-empirical method, direct sampling is possible although it may still be costly. Thus it is fruitful to develop methods to speed up alchemical free energy computations that mix different potential functions. We recently developed the integrated Hamiltonian sampling (IHS)(115), which can be regarded as an extension of the ITS approach(109); it is also intimately related to the enveloping distribution sampling of van Gunsteren and co-workers(116). In IHS, we introduce an effective potential ($U_{IHS}(\mathbf{R})$) whose canonical distribution is the integrated distributions of multiple potential functions,

$$U_{IHS}(\mathbf{R}) = -\frac{1}{\beta}\ln\int_0^1 d\lambda\Omega(\lambda)e^{-\beta U_\lambda(\mathbf{R})}, \tag{5.14}$$

where $\Omega(\lambda)$ is a weight function to be determined (see Ref.(115) for discussion of algorithms) and U_λ takes the usual form,

$$U_\lambda(\mathbf{R}) = (1-\lambda)U_0(\mathbf{R}) + \lambda U_1(\mathbf{R}). \tag{5.15}$$

In the specific case of MM$\rightarrow$QM conversion, U_0 and U_1 correspond to pure MM and QM/MM potential functions, respectively. In general applications, U_0 can be the potential function of a realistic system while U_1

the potential function of a fictitious system introduced to enhance sampling; e.g., specific torsional barriers or non-bonded interactions are scaled down.

5.2.2.2 Multi-level QM/MM free energy methods

As mentioned above, the thermodynamic cycle in Fig.5.1a can be easily generalized to connect low-level and high-level QM potentials in QM/MM simulations. This can be applied to improve the relative free energy of the same molecule in different environments (e.g., solvation free energy) or of different chemical states in the same environment (e.g., reactant vs. transition state in an enzyme active site). In practice, the challenge is that the computational cost for any calculation involving the high-level QM is so high that canonical sampling has to be at a minimal level or avoided altogether. For example, several studies computed the free energy profile (e.g., PMF along a pre-defined order parameter) with a low-level QM/MM potential function, and correction is estimated based on the difference between the low- and high-level QM/MM results along one or several MEPs(117, 118); for systems with a dynamical active site, the energetics of MEPs may fluctuate significantly as functions of the protein conformation(9, 45, 88), thus the correction may also have significant uncertainties. In other studies, sampling is done exclusively with the low-level QM/MM potential, and a one-step free energy perturbation is used to estimate the free energy change to high-level QM/MM(41, 119–122); as expected, the quantitative accuracy and convergence of the results depends critically on the overlap of distributions with different QM/MM potentials(121, 123), thus some authors advocated to carry out at least short sampling using high-level QM/MM potential rather than relying entirely on trajectories sampled with low-level QM/MM(13, 124, 125).

One interesting recent development in multi-level free energy computation is the Non-Boltzmann Bennet (NBB) approach(122), which treats the low-level simulation ensemble as that from a high-level simulation under a biasing potential being nothing but the energy difference between the low- and high-level potentials. This allows, for example, directly connecting the λ_0 window at high-level and λ_1 window at the low-level method (Fig.5.2a); by coupling regular BAR simulations at the low-level for the intermediate λ windows and NBB for the end states (λ_0, λ_n windows), one recovers the free energy difference between the end states with the high-level potential. The working expression for NBB combines the BAR equation and the familiar

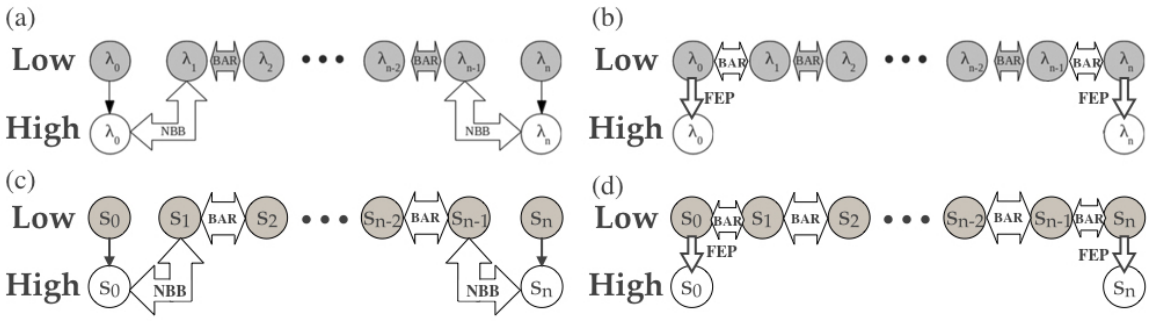


Figure 5.2: Two different routes to compute the free energy change as the potential function is switched between a low level and a high level. (a-b) illustrate the routes for alchemical free energy calaculations and (c-d) illustrate the corresponding calculations along a path (e.g., an optimized string). In the NBB route, a low-level window (e.g., λ_1^L) is connected with a neighboring high-level (e.g., λ_0^H) by re-weighting the λ_0^L data. In the FEP(+BAR) route, free energy change due to a switch in the potential function is calculated only for the same (e.g., λ_0) window; different λ/s windows (e.g., λ_0/λ_1 or s_0/s_1) are connected with the low-level potential function using BAR. See text for the corresponding equations for the free energy changes (compare Eqs. 5.16 and 5.18).

reweighting formula(64); for the high-level λ_0/low-level λ_1 free energy difference, for example, we have,

$$\Delta A^{NBB}(\lambda_0^H,\lambda_1^L) = \beta^{-1} ln\Big(\frac{\Big\langle f(U_{\lambda_0}^H - U_{\lambda_1}^L + C)\Big\rangle_{\lambda_1,L}}{\Big\langle f(U_{\lambda_1}^L - U_{\lambda_0}^H + C) exp(\beta V_0^b)\Big\rangle_{\lambda_0,L}} \Big\langle exp(\beta V_0^b)\Big\rangle_{\lambda_0,L}\Big) + C, \tag{5.16}$$

in which C is the constant to be determined in the BAR iteration, f is the Fermi function: $f(x) = \frac{1}{1+\exp(\beta x)}$, and $V_0^b = U_{\lambda_0}^L - U_{\lambda_0}^H$ in the energy difference between the low- and high-level potential functions for the λ_0 window. The notation $\langle\cdots\rangle_{\lambda_0,L}$ highlights that the ensemble average is over the simulation at the low-level.

The NBB approach has been tested using a series of solvation free energy simulations with both explicit and implicit solvent models(122); the low-level has been taken as either a MM model, or a low-level QM method, while the high-level is taken to be a DFT or correlated *ab initio* method. Encouraging results have been obtained. Nevertheless, the question is whether NBB offers a significant advantage over the more traditional approach in which the low- and high-level methods are connected at only specific windows (e.g., λ_0, λ_n) via one-step free energy perturbation. At the λ_0 window, for instance, one has,

$$\Delta A^{FEP}(\lambda_0^H,\lambda_0^L) = \beta^{-1} ln\Big\langle exp[-\beta(U_{\lambda_0}^H - U_{\lambda_0}^L)]\Big\rangle_{\lambda_0,L} = \beta^{-1} ln\Big\langle exp(\beta V_0^b)\Big\rangle_{\lambda_0,L}. \tag{5.17}$$

Therefore, the λ_0^H and λ_1^L windows can alternatively be connected by combining $\Delta A^{FEP}(\lambda_0^H,\lambda_0^L)$ and a BAR computation for λ_0,λ_1 at the low level (Fig. 5.2b):

$$\begin{aligned}\Delta A^{FEP+BAR}(\lambda_0^H,\lambda_1^L) &= \Delta A^{BAR}(\lambda_0^L,\lambda_1^L)+\Delta A^{FEP}(\lambda_0^H,\lambda_0^L)\\ &= \beta^{-1}ln\Big(\frac{\Big\langle f(U^L_{\lambda_0}-U^L_{\lambda_1}+C)\Big\rangle_{\lambda_1,L}}{\Big\langle f(U^L_{\lambda_1}-U^L_{\lambda_0}+C)\Big\rangle_{\lambda_0,L}}\Big)+C+\beta^{-1}ln\Big\langle exp(\beta V_0^b)\Big\rangle_{\lambda_0,L}\end{aligned} \quad (5.18)$$

Comparing Eqs.5.16 and 5.18, one sees that the two pathways of connecting λ_0^H and λ_1^L windows indeed have similar expressions. As well appreciated in the literature(126), both FEP and BAR methods require a good overlap of distributions between the two states being considered; numerically, BAR is more robust and therefore less sensitive to the degree of overlap, which likely motivated the development of NBB. We note, however, the numerical behavior of Eq.5.18 depends on the overlap between the low- and high-level distributions for the *same* (λ_0) window and the overlap between two λ windows at the *same* (low- level) theory, while Eq.5.16 depends on the overlap between the high-level distribution for λ_1 and the low-level distribution for λ_0; i.e., both the level of theory and λ values are changed *simultaneously* in NBB. Therefore, intuitively, NBB is expected to work less well than the traditional FEP+BAR route, despite the intrinsic numerical advantage of BAR. At a quantitative level, one possible way to compare the two approaches is to evaluate the variances of the two free energy difference estimators. This was done recently by Jia et al.(127), who showed for ΔA(λ_0^H,λ_1^L), for example,

$$\begin{aligned}\sigma^{\mathrm{NBB}}_{\Delta\mathrm{A}} = \frac{1}{N_{\lambda_0}\beta^2}\left(\frac{\Big\langle f^2(U^L_{\lambda_1}-U^H_{\lambda_0}-C)\Big\rangle_{\lambda_0,L}}{\Big\langle f(U^L_{\lambda_1}-U^H_{\lambda_0}-C)\Big\rangle^2_{\lambda_0,L}}-1\right)+\frac{1}{N_{\lambda_1}\beta^2}\left(\frac{\Big\langle f^2(U^H_{\lambda_0}-U^L_{\lambda_1}+C)\Big\rangle_{\lambda_1,L}}{\Big\langle f(U^H_{\lambda_0}-U^L_{\lambda_1}+C)\Big\rangle^2_{\lambda_1,L}}-1\right)\\ +\frac{1}{N_{\lambda_0}\beta^2}\left(\frac{\Big\langle \exp(2\beta(U^L_{\lambda_0}-U^H_{\lambda_0}))\Big\rangle_{\lambda_0,L}}{\Big\langle \exp(\beta(U^L_{\lambda_0}-U^H_{\lambda_0}))\Big\rangle_{\lambda_0,L}}-1\right).\end{aligned} \quad (5.19)$$

The variance is actually the sum of contributions from BAR between $U^H_{\lambda_0}$ and $U^L_{\lambda_1}$ (though with configurations sampled at low-level for both λ values) and FEP between $U^L_{\lambda_0}$ and $U^H_{\lambda_0}$; a similar expression can be obtained for $\sigma^{\mathrm{FEP+BAR}}_{\Delta\mathrm{A}}$ by replacing $U^H_{\lambda_0}$ by $U^L_{\lambda_0}$ in the first two terms on the right hand side of Eq. 5.19. The variance tends to be dominated by the FEP variance term, thus the result confirms our expectation that, with the same degree of sampling, the variance of NBB is unlikely better than the FEP+BAR route. It is nevertheless interesting to analyze the two approaches under broader context since variance alone may not be sufficient to characterize

the reliability of the multi-level free energy result.

The discussions are straightforwardly generalized to free energy correction along a reaction path, such as for images along a string. With the NBB scheme (Fig. 5.2c), for example, the argument of the Fermi function would include also the restraining potentials associated with different images in a string calculation, e.g., $U^H_{s_0} - U^L_{s_1} + \frac{1}{2}k_{s_0}(\vec{\xi} - \vec{\xi_{s_0}})^2 - \frac{1}{2}k_{s_1}(\vec{\xi} - \vec{\xi_{s_1}})^2$. It is of interest to compare the results to the traditional "one-step" FEP route (Fig. 5.2c vs. 5.2d) for specific images (e.g., those correspond to the reactant and transition state), as we do in Sect.5.3.2.2.

5.3 Model and realistic examples

5.3.1 Reaction free energy profile and pathways

5.3.1.1 Umbrella sampling and metadynamics

We'll use a simple model to illustrate the comparison between umbrella sampling, replica-exchange umbrella sampling (REUS) and metadynamics. As illustrated in Fig. 5.4a, the model describes proton transfer between two small water droplets through a "channel"; the confinement effect of the channel is implemented using a set of restraints available in CHARMM(128) and a set of point dipoles are included to stabilize the excess proton in the low-dielectric region of the model. This type of model was used in our previous studies to explore the definition of order parameters for long-range proton transfers(129), to understand the impact of boundary potential, including an external membrane potential(130), on proton transfers, and to compare different parameterizations of DFTB3 for studying the free energy of proton transfers(54). Here we use the model to illustrate the value of REUS for computing free energy profiles over regular umbrella sampling, and to illustrate the flexibility of choosing CVs in metadynamics simulations for gaining new mechanistic insights.

The model includes a water chain consisting of 19 QM water molecules and one excess proton; the distance from the first to the last oxygen is 16 Å. The length of the chain is motivated by the typical size of proton conducting channels in biomolecules. The water chain is held in the model channel by a cylindrical potential to each of the oxygen atoms:

$$E_{cyl} = \frac{1}{2} k_{cyl} (r - r_0)^2 \Theta(r - r_0), \tag{5.20}$$

with k_{cyl} = 125 kcal·mol^{-1}·Å^{-2}; r is the distance of the water oxygen atom from the central axis of the channel, r_0 is 3.5 Å, and Θ is the Heaviside step function. To ensure appropriate solvation of the excess proton at the ends of the water chain, 34 MM water molecules are added to each end of the chain. These "bulk" water molecules are subject to a set of quartic constraints to maintain a cubic shape of 10 × 10 × 5 Å^3. To stabilize the hydronium in the center of the channel, eight dipoles with modest dipole moments composed of two opposite charges (|q| = 0.5 e and d = 1.5 Å) are circularly arranged around the channel. The system is then embedded into a dielectric environment to mimic a lipid membrane (see Fig. 5.4a for dimensionality);

the dielectric constants for the membrane and bulk water are set to be 2 and 80, respectively.

To study the proton transfer, we carry out 1D REUS simulations as well as 2D metadynamics simulations and compare those to our recent study that used the standard 1D umbrella sampling(54). The order parameter used to describe the proton transfer itself is the ζ coordinate defined in our previous studies: it is based on the modified center of excess charge (mCEC)(129) and takes the form,

$$\zeta = \frac{d_{\xi,D} - d_{\xi,A}}{d_{\xi,D} + d_{\xi,A}}, \tag{5.21}$$

where ξ is the mCEC, D denotes the proton donor heavy atom, A is the proton acceptor heavy atom and d is the distance; one MM water molecule in each "bulk" region of the model is fixed and taken to be the donor/acceptor group. Therefore, a ζ value of ± 0.6 represents the excess proton being solvated by water molecules near the entrance/exit of the model channel and corresponds to the PMF minimum (see Figs. 5.3-5.4).

In the metadynamics simulations, a second coordinate is used to describe the level of solvation at the center of the channel. As discussed below, this was motivated by the observation that in the reactant state (i.e., when the excess proton is in the "bulk" region), a vacuum region (or a "bubble") is observed between the excess proton and the other side of the channel; this likely reflects the "amphiphilic" nature of the hydronium(131), which prefers to remain at the air/water interface. It is therefore of interest to explicitly study the coupling between proton transfer and wetting of the channel. The solvation number (81, 132) of the channel center is described using s, which is defined by,

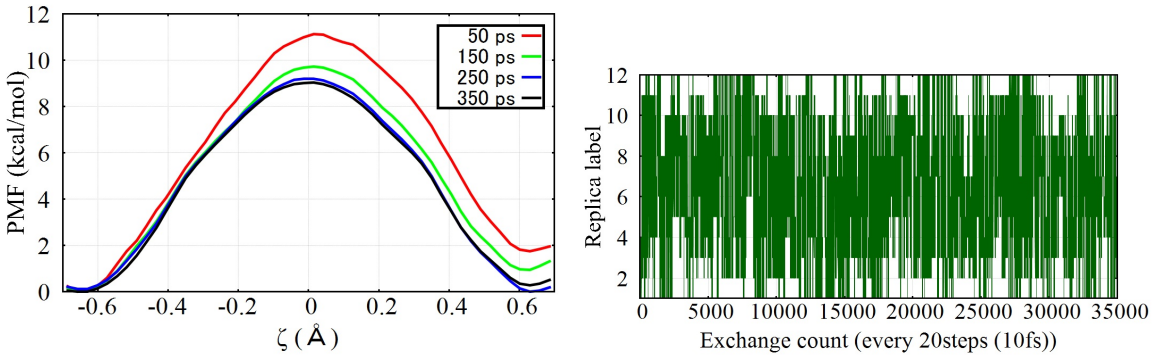
$$s = \sum_{i \in channel\ center} \sum_{j \in O_{solvent}} s_{ij}. \tag{5.22}$$

Here, the channel center is represented by a fixed dummy atom, and

$$\begin{aligned} s_{ij} &= 1 && r_{ij} \leq d_0 \\ &= \frac{1 - \{(r_{ij} - d_0)/r_0\}^6}{1 - \{(r_{ij} - d_0)/r_0\}^{12}} && r_{ij} > d_0 \end{aligned} \tag{5.23}$$

where r_{ij} is the distance between atom i and j and d_0, r_0 are set to be 2.5 and 0.5 Å, respectively.

In REUS, the same 13 set of umbrella windows with equal spacing in ζ as Ref.(54) is used. Exchange of configurations between a randomly selected pair of neighboring windows is attempted every 10 fs. Simulations are run till 350 ps for each replica, thus the total simulation time is comparable to that of Ref.(54).



(a) (b)

Figure 5.3: Results for the proton transfer in a model channel (see Fig. 5.4) from 1D REUS simulation along the ζ coordinate. (a) Convergence of the PMF as a function of simulation time per replica; (b) Illustration of the exchange behavior of replicas during the REUS simulation, using replica 2 as an example.

The DFTB3/3OB model(133) is used to describe the QM region. In the 2D metadynamics simulation, well-tempered metadynamics(85) is employed with a bias factor of 15. The initial values for Gaussian height are set to be 0.1 kcal/mol for ζ and 0.2 kcal/mol for s. A new Gaussian potential is added every 2.5 ps during the simulation so as to leave sufficient time for the system to respond to changes in the CVs. Two sets of DFTB3 O-H repulsive potential are tested, the standard 3OB(133) and a recent 3OBw(54) set developed based on improved description of bulk water. Each set of simulation is ran for 1.5 ns.

With the standard 1D umbrella sampling, the computed PMF in Ref.(54) still suffers from minor convergence issues after 500 ps of sampling per window. Indeed, the PMF is not symmetrical and has a small exothermicity of about 1 kcal/mol with the DFTB3/3OB model; the barrier measured from the more stable minimum is about 10.0 kcal/mol. With the 1D REUS (Fig. 5.3a), we see indeed the computed PMF features a significant level of asymmetry at the early stage of the simulation. As simulation proceeds, however, the degree of asymmetry decreases and a practically symmetrical and converged PMF is obtained after about 250 ps per replica. The diffusive behavior of the exchange moves (Fig. 5.3b) also confirms the efficiency of REUS. Therefore, the simple example illustrates the value of REUS over the standard umbrella sampling in terms of convergence.

As noted above, in the reactant, a vacuum region (or a “bubble”, see Fig. 5.4a) is formed between the excess proton and the other side of the channel. Therefore, it is expected that hydration level change (i.e.,

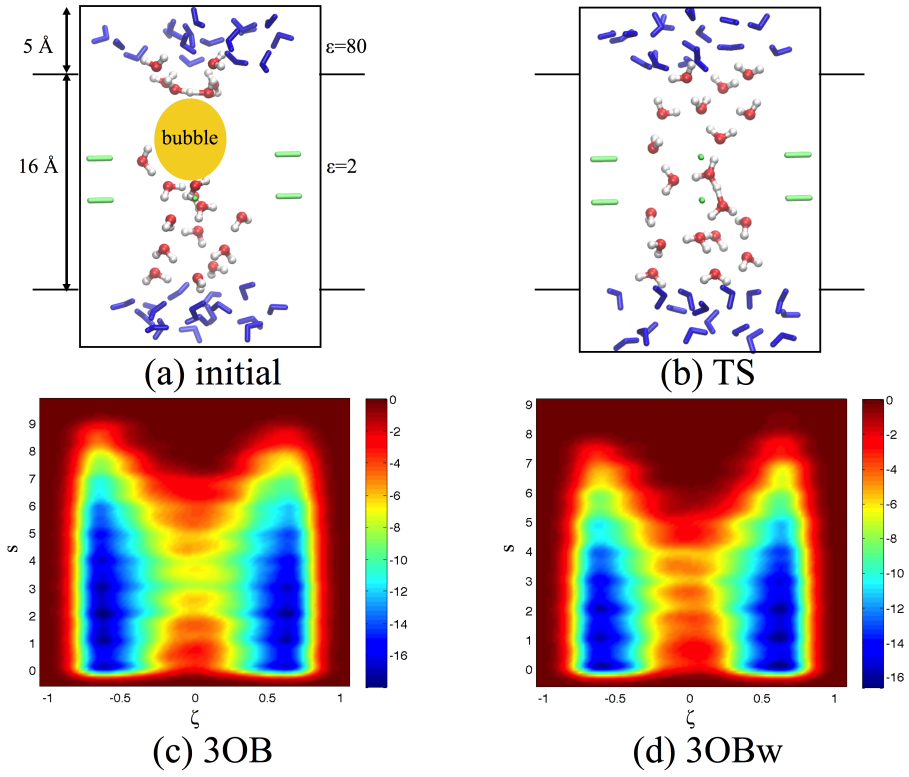


Figure 5.4: 2D metadynamics simulations for proton transfer in a model channel to illustrate the coupling between hydration level change and proton transfer. (a)-(b) Snapshots for the reactant and transition state regions, respectively. The "bulk" MM water molecules are shown in blue, and the QM water molecules are colored by atom type. The dipoles are shown in green. (c)-(d) Two dimensional PMF from well-tempered metadyanmics simulations using DFTB3/3OB(133) (barrier 9.8 kcal/mol) and DFTB3/3OBw(54) (barrier 11.5 kcal/mol), respectively. ζ describes the proton transfer and s describes the level of channel hydration (see text).

wetting) of the channel is explicitly coupled with the proton transfer. The key question regards whether such wetting is an energy demanding process. As shown by the 2D PMF from metadynamics, proton transfer indeed requires a substantial change in the level of hydration. The barrier along ζ is sufficiently low only when s is in the range of 3-4; as s further increases, the barrier along ζ remains fairly constant. However, the free energy profile along s is rather flat for $s \leq 6$ when the excess proton remains in either side of the "bulk" region (i.e., when $\zeta \sim \pm 0.6$). Therefore, wetting of the channel, at least with the current model, appears to be a rather facile process. This explains why the barriers from the 2D metadynamics simulation and the 1D umbrella sampling/REUS calculations are, in fact, very similar, regardless of the DFTB3 model used.

The general behaviors we found here are similar to the findings from a recent study of proton transfer across a carbon nanotube by Voth and co-workers using the MS-EVB model(134). It was found that the presence of the excess proton at one end of the nanotube induces spontaneous wetting of the tube and facilitates the subsequent proton transfer; interestingly, such favorable wetting behavior was not observed with other monovalent cations (e.g., Na^+) or a classical model of the hydronium ion (H_3O^+), suggesting that the delocalized nature of the excess proton in water is essential. This delocalized nature is captured in our study with the use of a QM model (DFTB3), and the relatively similar barrier heights between 3OB and 3OBw suggests that the qualitative behavior is not highly sensitive to the quantitative description of proton solvation.

The fact that local wetting around an excess proton does not appear to be energetically demanding even in a rather hydrophobic environment (low-dielectric in our model and carbon nanotube in Ref.(134)) has important implications to proton transfer in biomolecules. Proton transfer pathways are often identified by examining hydrogen bonding networks(135, 136) visible in protein structures from crystallography or equilibrium MD simulations in the absence of an excess proton. The current discussion highlights that the hydrogen bonding network may in fact change rather significantly in the presence of the excess proton to facilitate its transport. Depending on the specific molecular environment, the wetting process may or may not involve a substantial energy cost. Therefore, one should not rule out proton transfer pathways based purely on hydrogen-bonding network seen in structures obtained in the absence of the excess proton(137), and it is essential to explicitly include the local hydration level(35, 36) as one of the order parameters to characterize the energetics and kinetics of proton transfers.

5.3.1.2 String method in collective variables

To illustrate some of the practical issues in a string simulation, we study a seemingly simple proton transfer reaction inspired by the D-channel in cytochrome c oxidase(34). The model contains two aspartate sidechains separately by an asparagine sidechain, and the entire system is immersed in a water droplet of 20 Å radius. One aspartate is initially protonated, and the proton is transferred from the carboxylate to the carbonyl of the asparagine, which in turn rotates upward to deliver the proton to the upper aspartate (see Fig. 5.5). The positions of the three sidechain models are adjusted to make the proton transfer pathway feasible with only modest conformational isomerization of the asparagine. The model is constructed to explore whether an asparagine sidechain can relay proton transfers in biomolecules. The three sidechains are treated with DFTB3/3OB, and the water environment is treated with TIP3P under the stochastic boundary condition.

For the string calculation, 56 images are included, and the CV space is spanned by all the pairwise distances between heavy atoms in the three amino acid analogues and the distances involved the extra proton, leading to 62 distances. The friction coefficient γ in Eq. 6.2 is 100 ps^{-1}. MD sampling is carried out for 1.5 ns per image at 300 K with a time step of 0.5 fs; string evolution and reparameterizations are performed every 0.5 ps.

A few critical structures along the converged string are shown in Fig. 5.5a to illustrate the process of interest. The calculations are repeated with different values of the restraining force constants, which range from 5 to 20 kcal/mol/Å^2. The structural features of the converged strings do not seem to vary notably with respect to the restraining force constants, while the estimated free energy along the string does vary significantly (not shown). Although the free energy derivatives computed from string simulations were shown to converge as $O(k^{-1})$(91) (where k is the restraint force constant), the optimal value for k in an arbitrarily-dimensioned CV space is not obvious. For example, choosing a very high value precludes the use of replica-exchange between adjacent string points that can substantially accelerate convergence.(100)

To emphasize the importance of choosing a good CV set, we recall that two types of free energies can be computed from string calculations. The first is simply the integral of the n-dimensional PMF, $W(\vec{\xi})$ along the converged path

$$\Delta W(\alpha) = W(\vec{\xi}(\alpha)) - W(\vec{\xi}(0)) = \int_0^\alpha \sum_{i=1}^{n} \frac{d\xi_i(\alpha')}{d\alpha'} \frac{\partial W(\vec{\xi}(\alpha'))}{\partial \xi_i} d\alpha', \tag{5.24}$$

in which the free energy gradient $\nabla W(\vec{\xi}(\alpha))$ is evaluated using restrained MD simulations, and α is used to parametrize the string curve. Since the gradients are evaluated at the discrete images $\vec{\xi}_i \equiv \vec{\xi}(\alpha_i)$, numerical

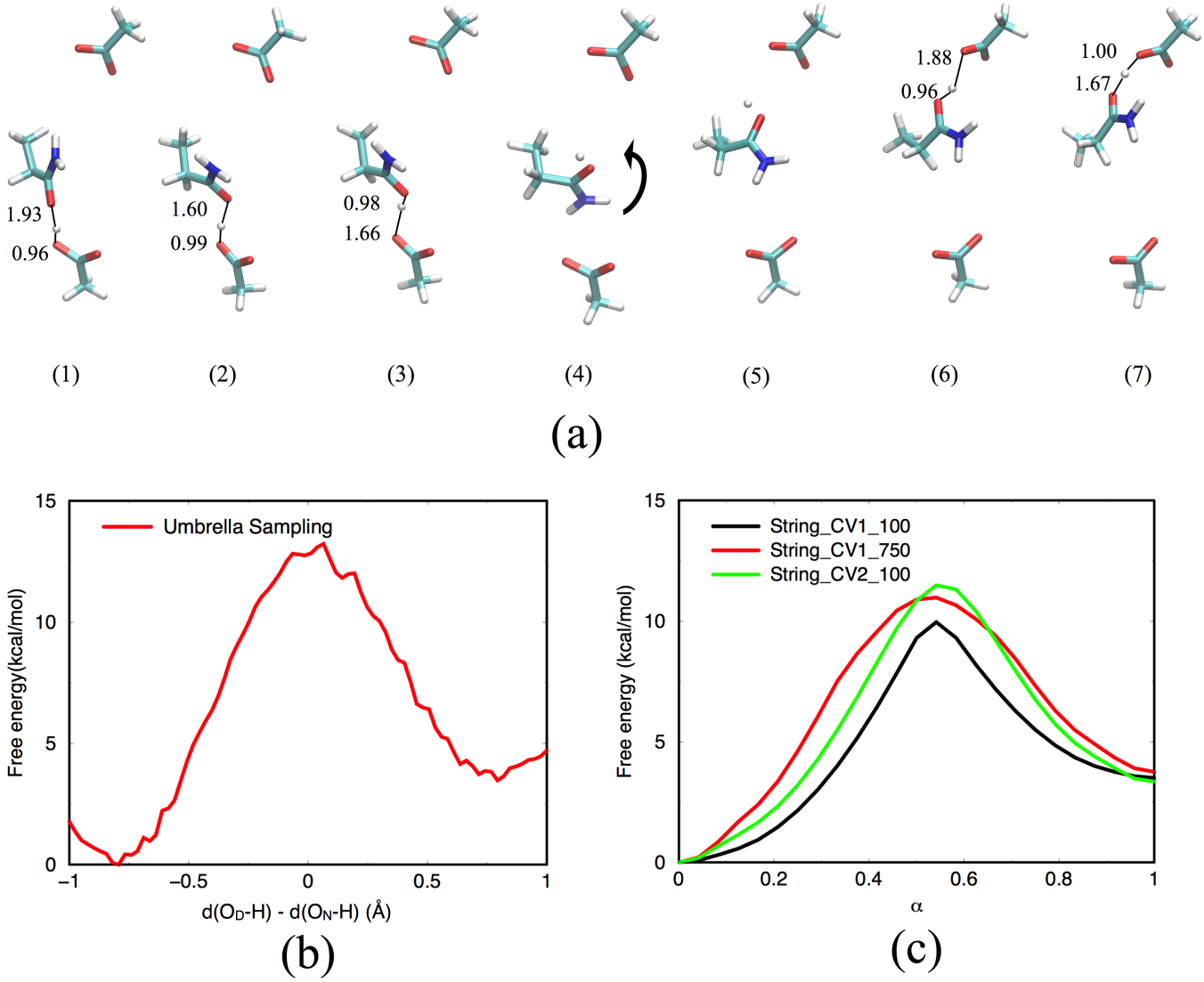


Figure 5.5: Study of a proton transfer process inspired by the D-channel in cytochrome c oxidase. (a) Representative structures along the optimized string with key distances (in Å) labeled (water molecules are excluded for clarity); (b) PMF for the first step of proton transfer (approximately corresponding to 1→3 in panel a) studied by umbrella sampling; (c) PMFs for the first step of proton transfer studied by the string method with different collective variables and restraining force constants (indicated in the legend in kcal/mol/Å^2).

quadrature must be used to evaluate Eq. (6.3), such as the trapezoid rule.(100) This free energy profile, however, does not generally capture the energetics that govern the reaction because $\Delta W(\alpha)$ does not contain the contributions due to the thermal fluctuations of the CVs orthogonal to the string. The proper 1D free energy quantity is the free energy as a function of the committor(100, 138), denoted by $F(\alpha)$, which is obtained by integrating over hypersurfaces orthogonal to the string. $\Delta W(\alpha)$ and $F(\alpha)$ are equivalent only if the transition tube is infinitely narrow or has identical cross sections along the reaction path. Alternatively speaking, entropic effects associated with the CVs orthogonal to the string are not properly included in $\Delta W(\alpha)$. This consideration suggests that one should be particularly careful when choosing the CVs. Too few CVs clearly won't describe the reaction process well (i.e., the committor is poorly parameterized); on the other hand, including too many CVs might effectively freeze out many potentially relevant degrees of freedom and greatly perturb the computed free energy profile. Moreover, including too many CVs also means that finer discretization is needed to avoid missing important changes in a few degrees of freedom.

To illustrate the sensitivity of the free energy profile from string calculations, we focus on the first proton transfer step, which approximately corresponds to the transition from structure 1 to 3 in Fig. 5.5a; to further simplify the situation, the Asn sidechain is fixed to coordinates in structure 1. First, we study the PMF for this step using the regular umbrella sampling, using the antisymmetric stretch, $d(\mathrm{O_{Asp}}\text{-H})$ - $d(\mathrm{O_{Asn}}\text{-H})$, as the order parameter; 21 windows (0.1 Å spacing) are used with 500 ps simulations for each window and the uniform force constant of 300 kcal/mol·Å^2. Next, we carry out string calculations for this first proton transfer step. 25 images are used to discretize the string, and the force constants used in the restrained simulations are varied from 100 to 750 kcal/mol·Å^2 and the friction coefficient γ is set to 100 ps^{-1}. Different CV sets are also used: CV1 set consists of the two distances in the antisymmetric stretch, $\mathrm{d(O_{Asp}}\text{-H)}$ and $\mathrm{d(O_{Asn}}\text{-H)}$; CV2 set includes 13 distances between the transferring proton and all heavy atoms in the model side chains. MD sampling is carried out for 1 ns per image at 300 K with a time step of 0.5 fs; string evolution and reparameterizations are performed every 0.5 ps. For the last 500 ps, image centers are fixed and free energy gradients are calculated and used for analysis.

As shown in Fig. 5.5b, the proton transfer from an aspartic acid to an asparagine in this model is only moderately endothermic by $\sim$4 kcal/mol with a barrier of $\sim$ 13 kcal/mol. Considering that the asparagine model is fixed thus the proton donor/acceptor distance has only a modest degree of flexibility (the O-O distance

changes from ∼2.9 Å in the reactant to ∼2.6 Å in the barrier region), the barrier height computed here should be an upper bound. Therefore, asparagine should not be ruled out as a proton relay group in biological proton transfers. With the string calculations, the calculated free energy profile $\Delta W(\alpha)$ depends on both the choice of CV and the restraining force constant, especially near the barrier. With CV1, the free energy barrier changes by about 1 kcal/mol as the restraining force constant is increased from 100 to 750 kcal/mol/Å^2. With the same restraining force constant, using two different CV sets also leads to a barrier difference of 1-2 kcal/mol; this level of difference is not unexpected considering that the two CV sets differ significantly in terms of dimensionality (2 vs. 13). In all three cases shown in Fig. 5.5c, the barrier appears too low compared to the umbrella sampling result by 2-3 kcal/mol; considering the local nature of the proton transfer, the 1D umbrella sampling result is expected to be more reliable.

The results highlight that free energy calculations using string methods should be treated with great care, even for a fairly simple local proton transfer. For more complex processes that involve both chemistry and conformational transitions, such as those in Fig. 5.5a, the energetics results appear even more sensitive to the choice of the CV space and restraining force constants. A more systematic analysis, including the comparison of $F(\alpha)$ and $\Delta W(\alpha)$, is being carried out and will be presented elsewhere.

5.3.2 Alchemical and multi-level free energy computations

5.3.2.1 Solvation free energy

In our recent study(54), using the thermodynamic cycle in Fig.5.1a, we have computed the solvation free energy of several small molecules at the DFTB3/3OB level relative to the CHARMM22 force field(144), which has been well calibrated for solvation free energies(145, 146) (although the value for charged species remains uncertain(147–149)). As also summarized in Table 5.1 for acetate and acetic acid, the DFTB3/3OB model gives slightly (∼3-5 kcal/mol) more favorable solvation free energy than the CHARMM 22 force field. Although part of such mild degree of oversolvation is probably due to the fact that van der Waals parameters(150) have not been reoptimized for the QM atoms at the DFTB3 level, another likely reason is that the electrostatic interactions between DFTB3 and MM (TIP3P) atoms employed a simple Coulombic model(151). As discussed in Ref.(152), a more appropriate model would consider the finite size of the QM charge distribution and therefore damp the interaction at short range, such as in a Klopman-Ohno (KO) model. Further optimization

Table 5.1: Relative Solvation Free Energy $\Delta\Delta G_{gas\rightarrow aq}$(MM→QM) (in kcal/mol) with Different Size of QM Region[a]

Solute	MM ref. (Exp.)[b]	DFTB3/3OB[c] small QM region	DFTB3/3OB[d] large QM region
CH_3COOH	-3.7 (-6.7)	-3.41 ± 0.07	+1.49 ± 0.11
CH_3COO^-	-85.4 (-77.6)	-5.36 ± 0.04	+3.21 ± 0.09

[a] Thermodynamic cycle shown in Figure 5.1 is used. MM force field is the CHARMM22 force field and water is described with a modified TIP3P model. The QM atoms use the standard CHARMM22 van der Waals parameters. GSBP(154, 155) (inner region 20 Å) is used for all the calculations here. [b] The "MM ref." values are absolute solvation free energies computed following the standard protocol. Values in parentheses are experimental values(156). [c] The small QM region includes the solute only. [d] The large QM region includes the solute and 27 water molecules in the first coordination shell of the solute. To ensure the QM water molecules are always nearest the metal ion, Flexible Inner Region Ensemble Separator (FIRES(157)) restraining potential is imposed to any outer sphere MM water molecules that are closer to the ion than the most distant inner sphere QM water molecule.

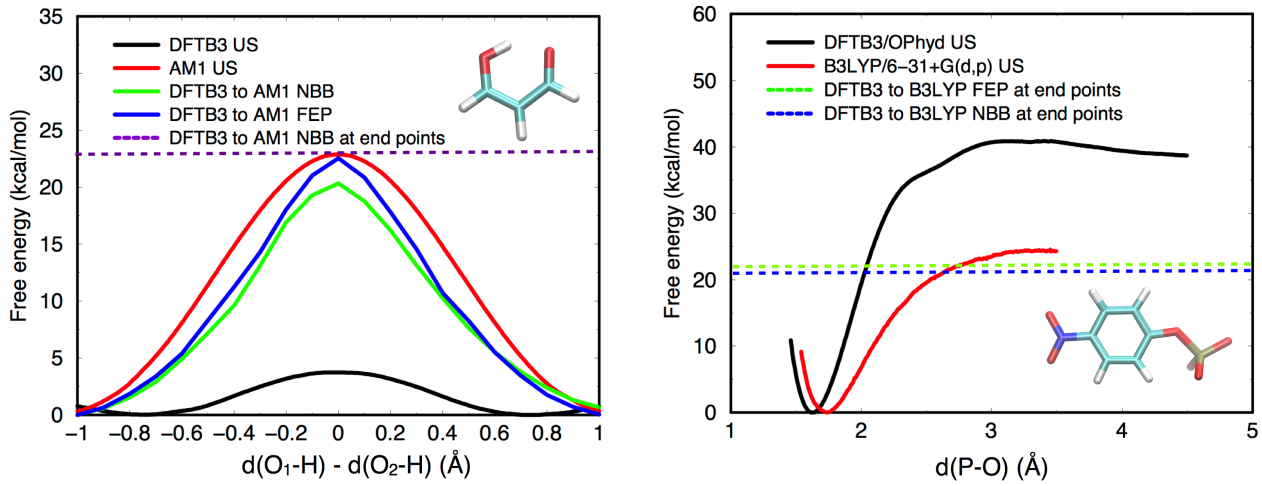
of the KO parameters and (charge-dependent(153)) van der Waals parameters for DFTB3 is an important task for ongoing work that aims to improve the robustness of DFTB3/MM for condensed phase applications(54).

Alternatively, one could describe the first solvation shell of the solute also with the DFTB3 model. This likely reduces the sensitivity of DFTB3/MM simulations to QM/MM coupling scheme and parameters, although for the study of heterogeneous systems like an enzyme, it is likely that any deficiency of QM/MM boundary will eventually propagate back to the active site of interest (Roston and Cui, private communication). To test this hypothesis, we compute the change of solvation free energies of acetate and acetic acid when the solute and its first solvation shell water molecules are converted from CHARMM 22 to DFTB3/3OB. As discussed in Sect.5.2.2.1, this requires a slightly revised thermodynamic cycle as illustrated in Fig. 5.1b. The results are also summarized in Table 5.1. Compared to CHARMM 22, the DFTB3 treatment now leads to slightly smaller solvation free energies, by ∼1-3 kcal/mol. From a technical point of view, the alchemical free energy simulations employ 11 λ windows even spaced between 0 and 1, and each window is equilibrated for 200 ps, which is followed by a production run of 500 ps. As shown in Table 5.1, this level of sampling leads to very small statistical errors as evaluated via block averaging, highlighting the advantage of converting between generally similar models for the solutes (CHARMM22 vs. DFTB3). As shown in Refs.(54, 115), convergence can be further enhanced with the IHS approach.

5.3.2.2 Dual-level QM/MM PMF

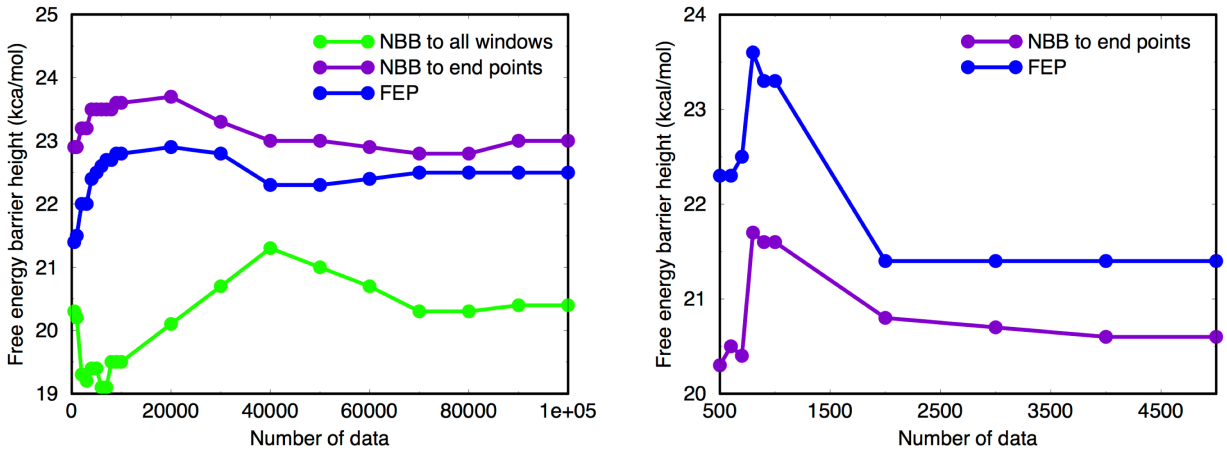
In this last subsection, we discuss several issues related to multi-level QM/MM free energy simulations using two simple solution reactions. One fitting example is the $pNPP^{2-}$ dissociation discussed, because DFTB3/MM and B3LYP/MM results have notable differences for both structures and energetics. The other example is the intra-molecular proton transfer in malondiadehyde in solution (a water droplet of 20 Å radius), for which we compare DFTB3/MM and AM1/MM simulations because they lead to rather similar structural properties but very different energetics. The low computational cost of these methods makes it straightforward to compute the reference 1D PMFs using umbrella sampling (see Fig. 5.6a); the reaction coordinate is defined as distance(O_1-H) - distance(O_2-H), 21 windows with an equal spacing of 0.1 Å are used for each PMF, and each window is sampled for 500 ps. We treat DFTB3/MM as the "low-level" theory for both systems, and the aim is to explore whether (or when) the dual-level free energy sampling schemes discussed in Sect.5.2.2.2 (see Fig. 5.2) are able to reproduce the "high-level" PMF in a numerically efficient and robust fashion.

We first discuss the case of proton transfer in malondiadehyde. As shown in Fig. 5.6a, AM1/MM and DFTB3/MM give barrier heights that differ by more than 15 kcal/mol; it is likely that the overestimation at the AM1 level is more significant(133), although this is not the focus of this study. Due to the symmetry of the problem, the barrier location is consistent at the two levels, while the reactant region at the DFTB3/MM level occurs at slightly smaller absolute values of the order parameters ($\sim \pm 0.8$ Å vs. $\sim \pm 1.0$ Å at the AM1 level). In terms of "correction" of the DFTB3/MM results, in addition to comparing the NBB and FEP+BAR routes, we also compare making corrections along the entire range of the order parameter, so as to obtain a corrected PMF, with making corrections only at the reactant/barrier windows so as to obtain only a corrected barrier height. When making corrections along the entire path, we see that the FEP+BAR route appears to lead to a better agreement in the barrier height (by $\sim$ 2 kcal/mol) than the NBB route with the actual AM1/MM umbrella sampling result; overall, however, both schemes lead to substantially "improved" barriers. On the other hand, with both schemes, the "corrected" PMF is narrower than the umbrella sampling result. When the "correction" is applied to the reactant (s_0=-1.0 Å) and barrier (s_n=0.0 Å) windows only, however, the barrier height is in good agreement with the original AM1/MM umbrella sampling (see Fig. 5.6a). Our interpretation is that when correction is applied along the entire path, errors due to the limited overlap between the two levels of theory (DFTB3/MM vs. AM1/MM) accumulate, leading to a larger error in the barrier height. Therefore,



(a) malondiadehyde (b) $pNPP^{2-}$

Figure 5.6: Potential of mean force (PMF) computed via standard umbrella sampling (US), and perturbation of the PMF following the NBB and FEP routes (see Fig. 5.2) for (a) intra-molecular proton transfer in malondiadehyde in solution (b) p-nitrophenyl phosphate ($pNPP^{2-}$) dissociation in solution. For the NBB and FEP calculations, the "low-level" method is taken to be DFTB3/MM for both reactions, while the "high-level" method is AM1/MM for (a) and B3LYP/MM for (b). For the NBB/FEP perturbations at the end-points, only two reactant windows and two transition state windows are used, as shown in Fig. 5.2, leading to "corrected" barrier heights indicated by dashed lines (also see text). For (a), $s_0 = -1.0$ Å, $s_1 = -0.9$ Å, $s_{n-1} = -0.1$ Å, s_n=0.0 Å; for (b), $s_0 = 1.6$ Å, $s_1 = 1.7$ Å, $s_{n-1} = 3.1$ Å, s_n=3.2 Å.



(a) malondiadehyde (b) pNPP^2

Figure 5.7: Corrected free energy barrier height via NBB/FEP using different numbers of data points per window. Convergence plots are shown for (a) intra-molecular proton transfer in malondiadehyde in solution (b) p-nitrophenyl phosphate (pNPP^{2-}) dissociation in solution. NBB/FEP at end points (i.e., reactant and transition state windows) only requires adequate overlaps between s_0-s_1 and s_{n-1}-s_n windows, thus the results are less dependent on the number of data points.

making perturbations only in the key windows seems more robust for practical applications.

In Fig. 5.7a, we further compare the convergence behavior of the "corrected" barrier height with different schemes when applied along the entire path or only to the reactant/barrier windows. Again, it is seen that the latter protocol leads to better convergence, with a variation of < 0.5 kcal/mol in the barrier height when from a few thousands to 10^5 number of configurations are included. This contrasts with the situation when NBB is applied along the entire path, where the variation in the barrier height is about 2 kcal/mol, not to mention the fact that the estimated barrier with 10^5 configurations per window is still 2 kcal/mol too low compared to the reference AM1/MM umbrella sampling result.

For the pNPP^{2-} dissociation reaction in solution, only 5,000 data points per window are used for the reactant and "barrier" windows because the B3LYP/MM calculations are substantially more expensive; the reactant window and the "barrier" window are taken to be $d_{O^{lg}P}$=1.6 and 3.2 Å, respectively. Thus only a correction

Table 5.2: Free energy correction and variance (in parentheses) of multi-level free energy estimators[a]

Perturbations	PT in malondiadehyde[b]			$pNPP^{-2}$ dissociation[c]		
	FEP	NBB	LRA	FEP	NBB	LRA
Reactant[d]	8240.7 (0.012/0.007)	-8241.1 (0.012/0.007)	8241.0	-654205.5 (0.06)	654206.4 (0.06)	-654208.9
Transition state[e]	8260.3 (0.011/0.005)	8260.4 (0.011/0.005)	8260.2	-654227.9 (0.32)	-654228.0 (0.33)	-654227.4
Free energy barrier[f]	22.5	23.0	22.2	21.4[g]	20.7[g]	25.3[g]

[a] All values are in kcal/mol. In FEP and NBB, only low (L) level simulations are conducted and perturbations to the high (H) level are computed following different routes (see Fig. 5.2). In LRA, both L and H trajectories are used in a linear response model. For PT in malondiadehyde, L=DFTB3/MM, H=AM1/MM; for $pNPP^{-2}$ dissociation, L=DFTB3/MM, H=B3LYP/MM. [b] $5000/10^5$ data points are included per window. [c] 5000 data points are included. [d] FEP and LRA calculates the free energy difference $s_0^L \rightarrow s_0^H$, while NBB calculates $s_0^H \rightarrow s_1^L$. For PT in malondiadehyde, $s_0 = -1.0$ Å, $s_1 = -0.9$ Å; for $pNPP^{-2}$ dissociation, $s_0 = 1.6$ Å, $s_1 = 1.7$ Å. [e] FEP and LRA calculates the free energy difference $s_n^L \rightarrow s_n^H$, while NBB calculates $s_{n-1}^L \rightarrow s_n^H$. For PT in malondiadehyde, $s_{n-1} = -0.1$ Å, s_n=0.0 Å; for $pNPP^{-2}$ dissociation, $s_{n-1} = 3.1$ Å, s_n=3.2 Å. [f] Corrected free energy barrier $s_0^H \rightarrow s_n^H$. For PT in malondiadehyde, the reference value (at "high"-level) is 22.9 kcal/mol; for $pNPP^{-2}$ dissociation, the reference value is 24.9 kcal/mol. [g] The corrected barrier height contains the 3 kcal/mol difference between 1.6 and 1.7 Å at the B3LYP/MM level (see text).

in the barrier height through either the NBB or the FEP+BAR route is obtained. Moreover, DFTB3/MM has a shorter equilibrium $d_{O^{lg}P}$ distance than B3LYP/MM (1.6 vs. 1.7 Å); thus the difference of $\sim$3 kcal/mol in the B3LYP/MM PMF at these two $d_{O^{lg}P}$ distances is included in the NBB/FEP corrected barrier height. As shown in Fig. 5.6b, the two routes lead to rather similar barrier heights; they still differ $\sim$ 2 kcal/mol from the B3LYP/MM umbrella sampling result (see below), although the improvement is quite encouraging considering the large ($\sim$ 20 kcal/mol) difference between the original DFTB3/MM and B3LYP/MM results. Examination of the convergence behavior (Fig. 5.7b) suggests that 5,000 data points appears to lead to a reasonable estimate for the barrier height correction, although whether this level of sampling is indeed adequate in general needs to be explored with more systems.

In an attempt to measure the reliability of the NBB/FEP+BAR correction schemes, we have also computed the variance of the corrections using the expression discussed in Sect.5.2.2.2 (e.g., Eq.5.19 for NBB). As shown in Table 5.2, the computed values seem rather small and substantially lower than 1 kcal/mol. These values, however, only report on the estimated statistical uncertainty for the corresponding free energy estimators. As discussed in Ref.(126), there can be substantial bias in the estimated free energy even when the variance is small. To illustrate this point, we plot the distributions for the energy difference between the low- and high-levels of theory (ΔU^{LH}) with configurations sampled from the two levels of theory; this compar-

ison is made for both the reactant and barrier windows for the two example reactions. Also plotted is the product of ΔU^{LH} distribution (sampled with the low-level trajectory) multiplied with the Boltzmann factor for ΔU^{LH}; free energy perturbation type of calculations are expected to be reliable only when this product closely matches the ΔU^{LH} distribution sampled at the high-level(126), considering an identity associated with free energy perturbation,

$$exp(-\beta\Delta U^{LH})P_L(\Delta U^{LH}) = exp[-\beta\Delta A(L\rightarrow H)]P_H(\Delta U^{LH}). \tag{5.25}$$

As shown in Fig. 5.8a-b, good match is indeed observed for the proton transfer in malondiadehyde; this is consistent with the fact both NBB and FEP+BAR schemes lead to satisfactory "corrections" of the barrier height. For the $pNPP^{2-}$ example, however, the DFTB3/MM and B3LYP/MM have broad *and* significantly different ΔU^{LH} distributions (Fig. 5.8c-d) that have almost vanishing overlap; as a result, the product $P_{DFTB3/MM}(\Delta U^{LH})e^{-\beta\Delta U^{LH}}$ deviates substantially from $P_{B3LYP/MM}(\Delta U^{LH})$. Thus, it is rather surprising that the one-step FEP seems to lead to fairly adequate correction of the barrier height as shown in Fig. 5.6b.

We rationalize this finding by the observation that ΔU^{LH} distributions are fairly Gaussian in nature at both the low- and high-levels of theories for the two reactions studied here; although variances of the ΔU^{LH} distributions differ at the low- and high-levels in some cases (e.g., see Fig. 5.8d, the variances are 4.7 and 2.9 kcal/mol at DFTB3/MM and B3LYP/MM levels, respectively), the difference is not large. Since second-order cumulant expansion for the FEP expression is exact for a Gaussian distribution(158), free energy difference can be evaluated based solely on ΔU^{LH} data along the trajectory sampled at either level of theory when the ΔU^{LH} variance is not sensitive to the level of theory:

$$\Delta A(L\rightarrow H) = -\beta^{-1}ln\left\langle e^{-\beta\Delta U^{LH}}\right\rangle_L = \left\langle \Delta U^{LH}\right\rangle_L - \frac{\beta}{2}\sigma^2_{\Delta U^{LH},L}, \tag{5.26}$$

$$\Delta A(L\rightarrow H) = \beta^{-1}ln\left\langle e^{-\beta\Delta U^{HL}}\right\rangle_H = -\left\langle \Delta U^{HL}\right\rangle_H + \frac{\beta}{2}\sigma^2_{\Delta U^{HL},H} = \left\langle \Delta U^{LH}\right\rangle_H + \frac{\beta}{2}\sigma^2_{\Delta U^{LH},H}. \tag{5.27}$$

There is of course no formal reason that the variances are the same at the two levels of theory, thus a better approach would be to compute ΔU^{LH} using trajectories sampled at both levels of theory and then compute the free energy difference using a linear response approximation (LRA), as recommended by Warshel and co-workers(13). Using this protocol would lead to a corrected barrier of 25.3 kcal/mol for $pNPP^{2-}$ dissociation, indeed better than using only the low-level (DFTB3/MM) trajectories with NBB or FEP+BAR (see Fig. 5.6 and Table 5.2). In practical applications, this requires running at least short trajectories with high-level

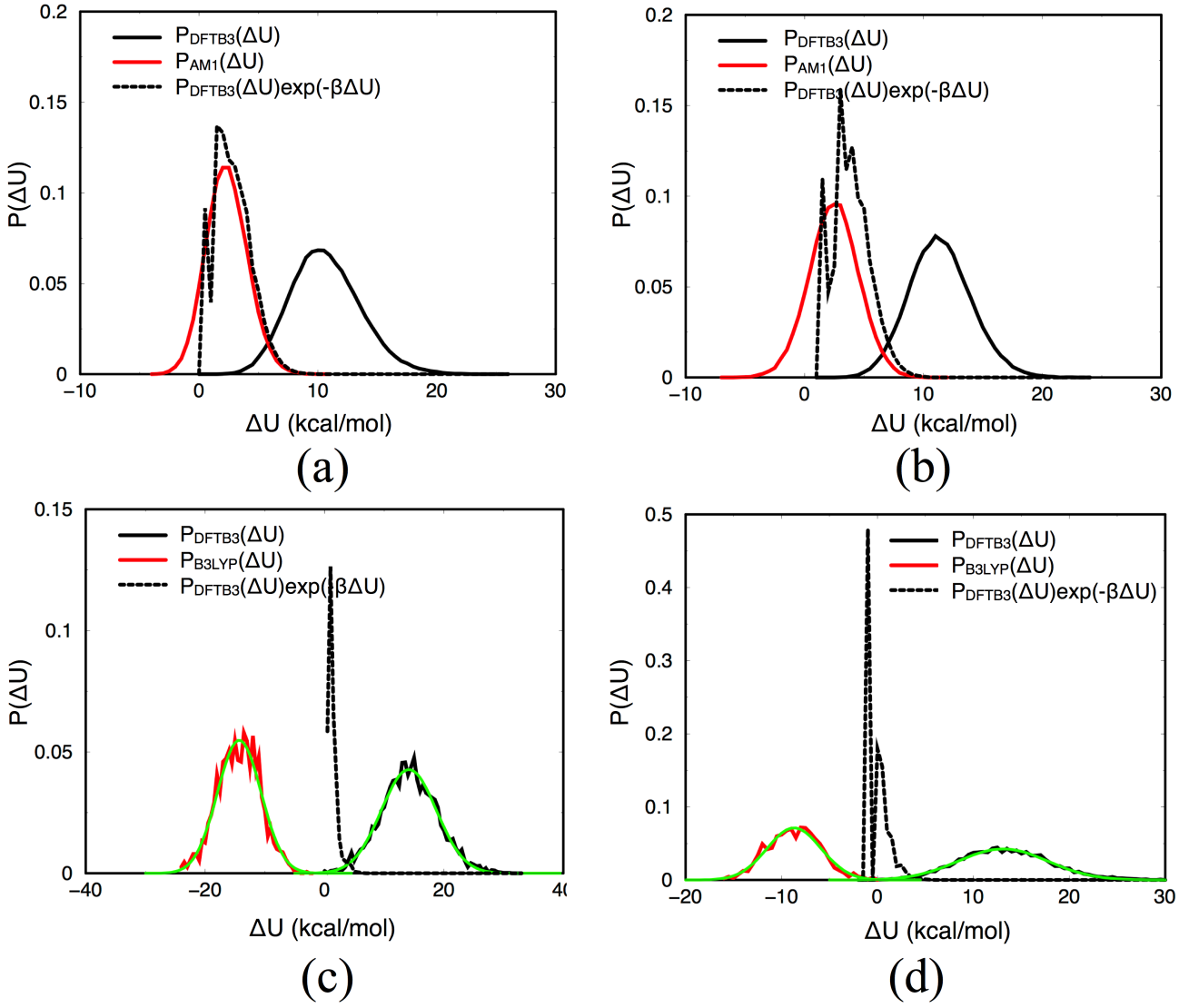


Figure 5.8: Distribution of energy difference between the low- and high-levels of theory, ΔU^{LH}, using trajectories from the two levels of theory; also plotted is the integrand of the FEP expression using the low-level trajectory (i.e., $exp(-\beta\Delta U^{LH})P_L(\Delta U^{LH})$). The ΔU^{LH} data are shifted with a different amount in (a-d) for plotting purposes; see Table 5.2 for absolute free energy perturbation results. The plots are for: (a) reactant in malondiadehyde proton transfer (b) transition state in malondiadehyde proton transfer (c) reactant in $pNPP^{2-}$ dissociation (d) transition state in $pNPP^{2-}$ dissociation. In (a)-(b), $P_{AM1}(\Delta U)$ has a similar shape from $P_{DFTB3}(\Delta U)\exp(-\beta\Delta U)$), thus FEP and NBB calculations yield reliable barrier corrections. In (c)-(d), the ΔU^{LH} distributions are remarkably broader and data at different levels have almost vanishing overlap, thus it seems challenging for FEP and NBB to obtain reliable barrier corrections, although the Gaussian nature of the ΔU^{LH} distributions (fit shown explicitly in green) suggests that a linear response model works well (see Table 5.2).

QM/MM, which may not always be feasible when the QM region is large.

In short, using two simple solution reactions, we show that perturbation schemes following either a NBB or a FEP+BAR route can indeed lead to a notable correction of the reaction free energy. The correction is more robust when perturbations are done only for selected windows, such as for the reactant and transition state windows to obtained an improved barrier height. The NBB approach does not seem to provide any major advantage over the FEP+BAR route. To judge whether the correction is robust, computing the variance of the free energy estimator is unlikely sufficient while examining the Gaussian nature of ΔU^{LH} distributions would be more informative, especially if ΔU^{LH} distributions at both low- and high-levels of theory can be examined. Finally, the correction scheme is expected to be meaningful only when critical structures on the free energy surface (reactant, intermediate, transition state) are similar at different levels of theory, thus determining/confirming these structures at the high-level of theory using, for example the QM/MM-MFEP approach, is expected to be an essential and practical option for realistic applications.

5.4 Concluding Remarks

Reliable free energy simulations are critical to both mechanistic analysis and computational designs. Since QM/MM calculations are generally more expensive compared to pure MM calculations, it is particularly worthwhile to carefully consider the best free energy practices for QM/MM simulations. In this contribution, we briefly review several recent developments in free energy calculations and discuss a number of topics stimulated by those developments using relatively simple processes in water.

When a low-level of QM theory, such as a semi-empirical method, is adequate, both potential of mean force and alchemical free energy simulations can be readily carried out with several established techniques, such as umbrella sampling, metadynamics and free energy perturbation. For complex reactions, minimum free energy path methods have become more prevalent, although we caution that the results, especially free energy profiles, can be sensitive to several parameters in these calculations, such as the choice of the CV space used to span the string and force constants used to estimate the mean force with restrained molecular dynamics. A practical strategy might be to compare the results of string calculations with other techniques for a local segment of the reaction (e.g., a local proton transfer step in a long-range proton transport) so as to calibrate key parameters in the string calculations. Finally, it is also productive to integrate metadynamics with finite temperature string type of calculations to explore the mechanism of complex reactions; metadynamics using

well-chosen CVs can be very effective at identifying new reaction channels and therefore provide good initial guess for string calculations, which are local in nature and therefore rely on well-chosen initial pathways.

The best strategy to employ high-level QM/MM potentials likely depends on the problem in hand. When the QM region is fairly rigid and fluctuations orthogonal to the reaction path are largely harmonic in nature, the QM/MM-MFEP framework pioneered by Yang and co-workers is expected to work well; the necessary mean-field approximation seems to be satisfactory at least for the example analyzed here. When the QM region is highly dynamical, an appropriate alternative is to determine the minimum free energy pathway at a semi-quantitative level using direct sampling with a low-level (semi-empirical) QM/MM method and then correct the energetics for key regions following perturbative schemes (Fig. 5.2). Provided that the key structures on the free energy surface remain similar at different levels of theory, our tests indicate that the perturbative schemes can lead to substantially improved barriers and in favorable cases, only low-level sampling is needed. In general, to make sure that perturbative schemes give meaningful results, however, it is essential to conduct careful analysis of the energy difference distributions and key structural features using trajectories at different levels of theory. In this context, even local structural minimization (e.g., adiabatic PMF mapping) in the QM/MM-MFEP framework again can be highly informative.

Regarding future efforts, we anticipate that continuing developments in the area of reliable low-cost QM methods for diverse elements, robust enhanced sampling methods, efficient ways to include polarization in both low-cost QM method and MM models will be particularly important. Another fruitful direction is the integration of free energy sampling with automated and systematic reaction path search(159, 160), which may find great value in timely applications such as annotation of protein functions(161), understanding of enzyme evolution(162, 163) and design of novel enzymes. Finally, although we focused our discussions on solution and biological systems, the developments are also equally important to chemical processes in other complex environments, such as liquid/solid interfaces that are prevalent in materials science, heterogeneous catalysis and energy/environmental research(164).

References

[1] L. PAULING, *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry, 3rd Ed.*, Cornell University Press, Ithaca, 1960.

[2] J. A. MCCAMMON, B. R. GELIN, and M. KARPLUS, *Nature* **267**, 585 (1977).

[3] A. WARSHEL and M. LEVITT, *J. Mol. Biol.* **103**, 227 (1976).

[4] M. J. FIELD, P. A. BASH, and M. KARPLUS, *J. Comput. Chem.* **11**, 700 (1990).

[5] K. B. LIPKOWITZ and D. B. BOYD, editors, *J. Gao, In Reviews in Computational Chemistry VII*, VCH: New York, 1995.

[6] T. VREVEN and K. MOROKUMA, *J. Comput. Chem.* **21**, 1419 (2000).

[7] R. A. FRIESNER and V. GUALLAR, *Annu. Rev. Phys. Chem.* **56**, 389 (2005).

[8] J. L. GAO, S. H. MA, D. T. MAJOR, K. NAM, J. Z. PU, and D. G. TRUHLAR, *Chem. Rev.* **106**, 3188 (2006).

[9] D. RICCARDI, P. SCHAEFER, Y. YANG, H. YU, N. GHOSH, X. PRAT-RESINA, P. KONIG, G. LI, D. XU, H. GUO, M. ELSTNER, and Q. CUI, *J. Phys. Chem. B* **110**, 6458 (2006).

[10] Y. ZHANG, *Theor. Chem. Acc.* **116**, 43 (2006).

[11] H. HU and W. T. YANG, *Annu. Rev. Phys. Chem.* **59**, 573 (2008).

[12] H. M. SENN and W. THIEL, *Angew. Chem. Int. Ed.* **48**, 1198 (2009).

[13] S. C. L. KAMERLIN, M. HARANCZYK, and A. WARSHEL, *J. Phys. Chem. B* **113**, 1253 (2009).

[14] E. BRUNK and U. ROTHLISBERGER, *Chem. Rev.* **115**, 6217 (2015).

[15] J. L. GAO and D. G. TRUHLAR, *Annu. Rev. Phys. Chem.* **53**, 467 (2002).

[16] G. MONARD and J. K. M. MERZ, *Acc. Chem. Res.* **32**, 904 (1999).

[17] S. SHAIK, S. COHEN, Y. WANG, H. CHEN, D. KUMAR, and W. THIEL, *Chem. Rev.* **110**, 949 (2010).

[18] M. GARCIA-VILOCA, J. GAO, M. KARPLUS, and D. G. TRUHLAR, *Science* **303**, 186 (2004).

[19] V. C. NASHINE, S. HAMMES-SCHIFFER, and S. J. BENKOVIC, *Curr. Opin. Chem. Biol.* **14**, 644 (2010).

[20] M. W. VAN DER KAMP and A. J. MULHOLLAND, *Biochem.* **52**, 2708 (2013).

[21] R. LONSDALE and A. J. MULHOLLAND, *Curr. Top. Med. Chem.* **14**, 1339 (2014).

[22] M. T. REETZ, M. PULS, J. D. CARBALLEIRA, A. VOGEL, K. E. JAEGER, T. EGGERT, W. THIEL, M. BOCOLA, and N. OTTE, *ChemBioChem* **8**, 106 (2007).

[23] M. P. FRUSCHICHEVA, M. J. L. MILLS, P. SCHOPF, M. K. SINGH, R. B. PRASAD, and A. WARSHEL, *Curr. Opin. Chem. Biol.* **21**, 56 (2014).

[24] J. WIETEK, J. S. WIEGERT, N. ADEISHVILI, F. SCHNEIDER, H. WATANABE, S. P. TSUNODA, A. VOGT, M. ELSTNER, T. G. OERTNER, and P. HEGEMANN, *Science* **344**, 409 (2014).

[25] D. A. KRAUT, K. S. CARROLL, and D. HERSCHLAG, *Annu. Rev. Biochem.* **72**, 517 (2003).

[26] J. G. ZALATAN and D. HERSCHLAG, *Nat. Chem. Biol.* **5**, 516 (2009).

[27] A. FERSHT, *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, W.H. Freeman and Company, 1999.

[28] Y. Q. DENG and B. ROUX, *J. Phys. Chem. B* **113**, 2234 (2009).

[29] H. L. SWEENEY and A. HOUDUSSE, *Annu. Rev. Biophys.* **39**, 539 (2010).

[30] Y. YANG, H. YU, and Q. CUI, *J. Mol. Biol.* **381**, 1407 (2008).

[31] C. X. SONG and C. HE, *Trends in Biochem. Sci.* **38**, 480 (2013).

[32] C. YI, G. JIA, G. HOU, Q. DAI, G. ZHENG, X. JIAN, C. G. YANG, Q. CUI, and C. HE, *Nature* **468**, 330 (2010).

[33] C. S. NABEL and R. M. KOHLI, *Science* **333**, 1229 (2011).

[34] V. R. KAILA, M. V. VERKHOVSKY, and M. WIKSTRÖM, *Chem. Rev.* **110**, 7062 (2010).

[35] P. GOYAL, J. LU, S. YANG, M. R. GUNNER, and Q. CUI, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18886 (2013).

[36] P. GOYAL, S. YANG, and Q. CUI, *Chem. Sci.* **6**, 826 (2015).

[37] O. KHERSONSKY, C. ROODVELDT, and D. S. TAWFIK, *Curr. Opin. Chem. Biol.* **10**, 498 (2006).

[38] J. K. LASSILA, J. G. ZALATAN, and D. HERSCHLAG, *Annu. Rev. Biochem.* **80**, 669 (2011).

[39] A. C. BABTIE, S. BANDYOPADHYAY, L. F. OLGUIN, and F. HOLLFELDER, *Angew. Chem. Int. Ed.* **48**, 3692 (2009).

[40] G. H. HOU and Q. CUI, *J. Am. Chem. Soc.* **134**, 229 (2012).

[41] G. H. HOU and Q. CUI, *J. Am. Chem. Soc.* **135**, 10457 (2013).

[42] F. DUARTE, B. A. AMREIN, and S. C. L. KAMERLIN, *Phys. Chem. Chem. Phys.* **15**, 11160 (2013).

[43] T. KOSUGI and S. HAYASHI, *J. Chem. Theory Comput.* **8**, 322 (2012).

[44] T. H. ROD and U. RYDE, *J. Chem. Theory Comput.* **1**, 1240 (2005).

[45] Q. CUI and M. KARPLUS, *J. Am. Chem. Soc.* **124**, 3093 (2002).

[46] K. NAM, Q. CUI, J. GAO, and D. M. YORK, *J. Chem. Theory Comput.* **3**, 486 (2007).

[47] Y. YANG, H. YU, D. YORK, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **4**, 2067 (2008).

[48] Y. ZHOU and J. PU, *J. Chem. Theory Comput.* **10**, 3038 (2014).

[49] N. V. PLOTNIKOV, S. C. L. KAMERLIN, and A. WARSHEL, *J. Phys. Chem. B* **115**, 7950 (2011).

[50] M. ELSTNER, D. POREZAG, G. JUNGNICKEL, J. ELSNER, M. HAUGK, T. FRAUENHEIM, S. SUHAI, and G. SEIFERT, *Phys. Rev. B* **58**, 7260 (1998).

[51] M. ELSTNER, T. FRAUENHEIM, and S. SUHAI, *THEOCHEM* **632**, 29 (2003).

[52] M. GAUS, Q. CUI, and M. ELSTNER, *WIREs Comput. Mol. Sci.* **4**, 49 (2014).

[53] Q. CUI and M. ELSTNER, *Phys. Chem. Chem. Phys.* **16**, 14368 (2014).

[54] P. GOYAL, H. J. QIAN, S. IRLE, X. LU, D. ROSTON, T. MORI, M. ELSTNER, and Q. CUI, *J. Phys. Chem. B* **118**, 11007 (2014).

[55] M. GAUS, Q. CUI, and M. ELSTNER, *J. Chem. Theory Comput.* **7**, 931 (2011).

[56] W. THIEL, *Adv. Chem. Phys.* **93**, 703 (1996).

[57] M. KUBILLUS, T. KUBAŘ, M. GAUS, J. ŘEZÁČ, and M. ELSTNER, *J. Chem. Theory Comput.* **11**, 332 (2015).

[58] X. LU, M. GAUS, M. ELSTNER, and Q. CUI, *J. Phys. Chem. B (Jorgensen Festschrift)* **119**, 1062 (2015).

[59] M. GAUS, H. JIN, D. DEMAPAN, A. S. CHRISTENSEN, P. GOYAL, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **11**, 4205 (2015).

[60] D. G. XU, Q. CUI, and H. GUO, *Int. Rev. Phys. Chem.* **33**, 1 (2014).

[61] D. K. CHAKRAVORTY, B. WANG, C. W. LEE, D. P. GIEDROC, and K. M. MERZ, JR., *J. Am. Chem. Soc.* **134**, 3367 (2012).

[62] V. MLYNSKY, P. BANAS, J. SPONER, M. W. VAN DER KAMP, A. J. MULHOLLAND, and M. OTYEPKA, *J. Chem. Theory Comput.* **10**, 1608 (2014).

[63] M. GAUS, X. LU, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **10**, 1518 (2014).

[64] G. M. TORRIE and J. P. VALLEAU, *J. Comput. Phys.* **23**, 187 (1977).

[65] E. NERIA and M. KARPLUS, *Chem. Phys. Lett.* **267**, 23 (1997).

[66] A. WARSHEL, *Computer Modeling of Chemical Reactions in Enzymes and Solution*, Wiley, New York, 1991.

[67] Y. R. MO and J. L. GAO, *J. Comput. Chem.* **21**, 1458 (2000).

[68] C. BARTELS and M. KARPLUS, *J. Comput. Chem.* **18**, 1450 (1997).

[69] E. DARVE and A. POHORILLE, *J. Chem. Phys.* **115**, 9169 (2001).

[70] M. SPRIK and G. CICCOTTI, *J. Chem. Phys.* **109**, 7737 (1998).

[71] S. KUMAR, D. BOUZIDA, R. H. SWENDSEN, P. A. KOLLMAN, and J. M. ROSENBERG, *J. Comput. Chem.* **13**, 1011 (1992).

[72] D. RICCARDI, P. KOENIG, H. GUO, and Q. CUI, *Biochem.* **47**, 2369 (2008).

[73] W. K. DEN OTTER, *J. Chem. Phys.* **112**, 7283 (2000).

[74] J. HENIN and C. CHIPOT, *J. Chem. Phys.* **121**, 2904 (2004).

[75] E. ROSTA and G. HUMMER, *J. Chem. Theory Comput.* **11**, 276 (2015).

[76] H. WU, A. S. J. S. MEY, E. ROSTA, and F. NOÉ, *J. Chem. Phys.* **141**, 214106 (2014).

[77] Y. SUGITA, A. KITAO, and Y. OKAMOTO, *J. Chem. Phys.* **113**, 6042 (2000).

[78] K. MURATA, Y. SUGITA, and Y. OKAMOTO, *Chem. Phys. Lett.* **385**, 1 (2004).

[79] H. FUKUNISHI, O. WATANABE, and S. TAKADA, *J. Chem. Phys.* **116**, 9058 (2002).

[80] S. PARK, T. KIM, and W. IM, *Phys. Rev. Lett.* **108**, 108102 (2012).

[81] A. LAIO and F. L. GERVASIO, *Rep. Prog. Phys.* **71**, 126601 (2008).

[82] A. BARDUCCI, M. BONOMI, and M. PARRINELLO, *Wiley Inter. Rev. Comp. Mol. Sci.* **1**, 826 (2011).

[83] T. HUBER, A. TORDA, and W. F. VAN GUNSTEREN, *J. Comput. Aided Mol. Des.* **8**, 695 (1994).

[84] J. F. DAMA, M. PARRINELLO, and G. A. VOTH, *Phys. Rev. Lett.* **112**, 240602 (2014).

[85] A. BARDUCCI, G. BUSSI, and M. PARRINELLO, *Phys. Rev. Lett.* **100**, 020603 (2008).

[86] S. PIANA and A. LAIO, *J. Phys. Chem. B* **111**, 4553 (2007).

[87] K. FUKUI, *Acc. Chem. Res.* **14**, 363 (1981).

[88] Y. K. ZHANG, J. KUA, and J. A. MCCAMMON, *J. Phys. Chem. B* **107**, 4459 (2003).

[89] D. J. WALES, *Energy Landscapes*, Cambridge University Press, 2003.

[90] P. G. BOLHUIS, D. CHANDLER, C. DELLAGO, and P. L. GEISSLER, *Annu. Rev. Phys. Chem.* **53**, 291 (2002).

[91] L. MARAGLIANO, A. FISCHER, E. VANDEN-EIJNDEN, and G. CICCOTTI, *J. Chem. Phys.* **125**, 024106 (2006).

[92] C. DELLAGO, P. BOLHUIS, and P. GEISSLER, *Adv. Chem. Phys.* **123**, 1 (2003).

[93] W. E, W. REN, and E. VANDEN-EIJNDEN, *Chem. Phys. Lett.* **413**, 242 (2005).

[94] R. BEST and G. HUMMER, *Proc. Natl. Acad. Sci. USA* **102**, 6732 (2005).

[95] B. PETERS and B. TROUT, *J. Chem. Phys.* **125**, 054108 (2006).

[96] S. V. KRIVOV and M. KARPLUS, *J. Phys. Chem. B* **110**, 12689 (2006).

[97] R. ELBER and M. KARPLUS, *Chem. Phys. Lett.* **139**, 375 (1987).

[98] G. HENKELMAN, B. UBERGUAGA, and H. JÓNSSON, *J. Chem. Phys.* **113**, 9978 (2000).

[99] W. E, W. REN, and E. VANDEN-EIJNDEN, *J. Chem. Phys.* **126**, 164103 (2007).

[100] V. OVCHINNIKOV, M. KARPLUS, and E. VANDEN-EIJNDEN, *J. Chem. Phys.* **134**, 085103 (2011).

[101] L. MARAGLIANO and E. VANDEN-EIJNDEN, *Chem. Phys. Lett.* **446**, 182 (2007).

[102] L. MARAGLIANO, B. ROUX, and E. VANDEN-EIJNDEN, *J. Chem. Theory Comput.* **10**, 524 (2014).

[103] A. C. PAN, D. SEZER, and B. ROUX, *J. Phys. Chem. B* **112**, 3432 (2008).

[104] M. SANCHEZ-MARTINEZ, M.FIELD, and R. CREHUET, *J. Phys. Chem. B* **119**, 1103 (2014).

[105] H. HU, Z. Y. LU, and W. T. YANG, *J. Chem. Theory Comput.* **3**, 390 (2007).

[106] H. HU, Z. Y. LU, J. M. PARKS, S. K. BURGER, and W. T. YANG, *J. Chem. Phys.* **128**, 034105 (2008).

[107] E. M. MULLER, A. DE MEIJERE, and H. GRUBMULLER, *J. Chem. Phys.* **116**, 897 (2002).

[108] L. J. YANG and Y. Q. GAO, *J. Chem. Phys.* **131**, 214109 (2009).

[109] Y. Q. GAO, *J. Chem. Phys.* **128**, 064105 (2008).

[110] L. WANG, R. A. FRIESNER, and B. J. BERNE, *J. Phys. Chem. B* **115**, 9431 (2011).

[111] J. ZHANG, Y. I. YANG, L. J. YANG, and Y. Q. GAO, *J. Phys. Chem. B* **119**, 5518 (2015).

[112] T. P. STRAATSMA and J. A. MCCAMMON, *Annu. Rev. Phys. Chem.* **43**, 407 (1992).

[113] P. A. KOLLMAN, *Chem. Rev.* **93**, 2395 (1993).

[114] T. SIMONSON, G. ARCHONTIS, and M. KARPLUS, *Acc. Chem. Res.* **35**, 430 (2002).

[115] T. MORI, R. J. HAMERS, J. A. PEDERSEN, and Q. CUI, *J. Phys. Chem. B* **118**, 8210 (2014).

[116] C. D. CHRIST and W. F. VAN GUNSTEREN, *J. Chem. Phys.* **126**, 184110 (2007).

[117] S. MARTI, V. MOLINER, and I. TUÑÓN, *J. Chem. Theory Comput.* **1**, 1008 (2005).

[118] F. CLAEYSSENS, J. N. HARVEY, F. R. MANBY, R. A. MATA, A. J. MULHOLLAND, K. E. RANAGHAN, M. SCHUTZ, S. THIEL, W. THIEL, and H. J. WERNER, *Angew. Chem. Int. Ed.* **45**, 6856 (2006).

[119] C. J. WOODS, F. R. MANBY, and A. J. MULHOLLAND, *J. Chem. Phys.* **128**, 014109 (2008).

[120] M. RETEGAN, M. MARTINS-COSTA, and M. F. RUIZ-LOPEZ, *J. Chem. Phys.* **133**, 064103 (2010).

[121] I. POLYAK, T. BENIGHAUS, E. BOULANGER, and W. THIEL, *J. Chem. Phys.* **139**, 064105 (2013).

[122] G. KÖNIG, P. S. HUDSON, S. BORESCH, and H. L. WOODCOCK, *J. Chem. Theory Comput.* **10**, 1406 (2014).

[123] J. HEIMDAL and U. RYDE, *Phys. Chem. Chem. Phys.* **14**, 12592 (2012).

[124] N. V. PLOTNIKOV, *J. Chem. Theory Comput.* **10**, 2987 (2014).

[125] K. NAM, *J. Chem. Theory Comput.* **10**, 4175 (2014).

[126] A. POHORILLE, C. JARZYNSKI, and C. CHIPOT, *J. Phys. Chem. B* **114**, 10235 (2010).

[127] X. Y. JIA, M. T. WANG, Y. H. SHAO, G. KÖNIG, B. R. BROOKS, J. Z. H. ZHANG, and Y. MEI, *J. Chem. Theory Comput.* **Submitted** (2015).

[128] B. R. BROOKS, C. L. B. III, A. D. MACKERELL, L. NILSSON, R. J. PETRELLA, B. ROUX, Y. WON, G. ARCHONTIS, C. BARTELS, S. BORESCH, A. CAFLISCH, L. CAVES, Q. CUI, A. R. DINNER, M. FEIG, S. FISCHER, J. GAO, M. HODOSCEK, W. IM, K. KUCZERA, T. LAZARIDIS, J. MA, V. OVCHINNIKOV, E. PACI, R. W. PASTOR, C. B. POST, J. Z. PU, M. SCHAEFER, B. TIDOR, R. M. VENABLE, H. L. WOODCOCK, X. WU, W. YANG, D. M. YORK, and M. KARPLUS, *J. Comput. Chem.* **30**, 1545 (2009).

[129] P. KOENIG, N. GHOSH, M. HOFFMAN, M. ELSTNER, E. TAJKHORSHID, T. FRAUENHEIM, and Q. CUI, *J. Phys. Chem. A* **110**, 548 (2006).

[130] J. ZIENAU and Q. CUI, *J. Phys. Chem. B* **116**, 12522 (2012).

[131] M. K. PETERSEN, S. S. IYENGAR, T. J. F. DAY, and G. A. VOTH, *J. Phys. Chem. B* **108**, 14804 (2004).

[132] T. MORI, R. J. HAMERS, J. A. PEDERSEN, and Q. CUI, *J. Chem. Theory Comput.* **9**, 5059 (2013).

[133] M. GAUS, A. GOEZ, and M. ELSTNER, *J. Chem. Theory Comput.* **9**, 338 (2012).

[134] Y. X. PENG, J. M. J. SWANSON, S. G. KANG, R. H. ZHOU, and G. A. VOTH, *J. Phys. Chem. B* **119**, 9212 (2015).

[135] D. N. SILVERMAN and R. MCKENNA, *Acc. Chem. Res.* **40**, 669 (2007).

[136] Y. Y. SHAM, I. MUEGGE, and A. WARSHEL, *Proteins: Struct., Funct., & Bioinf.* **36**, 484 (1999).

[137] M. KATO, A. V. PISLIAKOV, and A. WARSHEL, *Proteins: Struct. Funct. Bioinfor.* **64**, 829 (2006).

[138] W. E and E. VANDEN-EIJNDEN, *Annu. Rev. Phys. Chem.* **61**, 391 (2010).

[139] A. BECKE, *J. Chem. Phys.* **98**, 5648 (1993).

[140] C. LEE, W. YANG, and R. G. PARR, *Phys. Rev. B* **37**, 785 (1988).

[141] C. L. BROOKS III and M. KARPLUS, *J. Chem. Phys.* **79**, 6312 (1983).

[142] M. KARPLUS and J. N. KUSCHICK, *Macromol.* **14**, 325 (1981).

[143] J. SCHLITTER, *Chem. Phys. Lett.* **215**, 617 (1993).

[144] A. D. MACKERELL JR., D. BASHFORD, M. BELLOTT, R. L. DUNBRACK JR., J. D. EVENSECK, M. J. FIELD, S. FISCHER, J. GAO, H. GUO, S. HA, D. JOSEPH-MCCARTHY, L. KUCHNIR, K. KUCZERA, F. T. K. LAU, C. MATTOS, S. MICHNICK, T. NGO, D. T. NGUYEN, B. PRODHOM, W. E. REIHER III, B. ROUX, M. SCHLENKRICH, J. C. SMITH, R. STOTE, J. STRAUB, M. WATANABE, J. WIÓRKIEWICZ-KUCZERA, D. YIN, and M. KARPLUS, *J. Phys. Chem. B* **102**, 3586 (1998).

[145] D. SHIVAKUMAR, Y. Q. DENG, and B. ROUX, *J. Chem. Theory Comput.* **5**, 919 (2009).

[146] Y. Q. DENG and B. ROUX, *J. Phys. Chem. B* **108**, 16567 (2004).

[147] Y. LIN, A. ALEKSANDROV, T. SIMONSON, and B. ROUX, *J. Chem. Theory Comput.* **10**, 2690 (2014).

[148] X. LU and Q. CUI, *J. Phys. Chem. B* **117**, 2005 (2013).

[149] M. A. KASTENHOLZ and P. H. HÜNENBERGER, *J. Chem. Phys.* **124**, 224501 1 (2006).

[150] D. RICCARDI, G. LI, and Q. CUI, *J. Phys. Chem. B* **108**, 6467 (2004).

[151] Q. CUI, M. ELSTNER, E. KAXIRAS, T. FRAUENHEIM, and M. KARPLUS, *J. Phys. Chem. B* **105**, 569 (2001).

[152] G. HOU, X. ZHU, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **8**, 4293 (2012).

[153] T. J. GIESE and D. M. YORK, *J. Chem. Phys.* **127**, 194101 (2007).

[154] W. IM, S. BERNECHE, and B. ROUX, *J. Chem. Phys.* **114**, 2924 (2001).

[155] P. SCHAEFER, D. RICCARDI, and Q. CUI, *J. Chem. Phys.* **123**, Art. No. 014905 (2005).

[156] A. V. MARENICH, C. P. KELLY, J. D. THOMPSON, G. D. HAWKINS, C. C. CHAMBERS, D. J. GIESEN, P. WINGET, C. J. CRAMER, and D. G. TRUHLAR, *Minnesota Solvation Database – version 2012* (2012).

[157] C. N. ROWLEY and B. ROUX, *J. Chem. Theory Comput.* **8**, 3526 (2012).

[158] R. KUBO, M. TODA, and N. HASHITSUME, *Statistical Physics II: Nonequilibrium Statistical Mechanics*, Springer, 2003.

[159] S. MAEDA, K. OHNO, and K. MOROKUMA, *Phys. Chem. Chem. Phys.* **15**, 3683 (2013).

[160] S. HABERSHON, *J. Chem. Phys.* **143**, 094106 (2015).

[161] M. P. JACOBSON, C. KALYANARAMAN, S. W. ZHAO, and B. X. TIAN, *Trends Biochem. Sci.* **39**, 363 (2014).

[162] D. S. TAWFIK, *Curr. Opin. Chem. Biol.* **21**, 73 (2014).

[163] A. TOTH-PETROCZY and D. S. TAWFIK, *Curr. Opin. Struct. Biol.* **26**, 131 (2014).

[164] C. J. MURPHY, A. M. VARTANIAN, F. M. GEIGER, R. J. HAMERS, J. PEDERSON, Q. CUI, C. L. HAYNES, E. E. CARLSON, R. HERNANDEZ, R. D. KLAPER, G. ORR, and Z. ROSENZWEIG, *ACS Central* **1**, 117 (2015).

Chapter 6

Regulation and Plasticity of Catalysis in Enzymes: Insights from Analysis of Mechanochemical Coupling in Myosin[1]

6.1 Introduction

Rational design of highly efficient catalysts is one of the grand challenges of modern science. Guided by different theoretical and computational analyses, many successful examples of rational catalyst design have been reported, with applications in electrocatalysts,(1) inorganic materials,(2) nanoporous metal-organic-frameworks(3) and artificial enzymes(4, 5). It remains difficult to achieve a consistently high degree of catalytic proficiency, especially for structurally flexible systems. The rational design process requires not only a deep understanding the physicochemical properties of the active site (e.g., flexibility and hydration)(6) but also of the allosteric coupling throughout the system; multiple studies show that even mutations that are remote from the active site make significant contributions to catalysis(7–9). Consequently, despite ongoing progress, most designed enzymes have a rather low level of catalytic activity, and need to be further refined through rounds of directed evolution(4, 5, 10, 11).

Further progress in the design of efficient catalysts will require a precise understanding of how the immediate coordination shell of active site and the nearby residues and water molecules affect catalytic efficiency. This issue goes beyond artificial enzymes. For example, in the design of nano/meso-porous materials, an

[1]Reprinted (adapted) with permission from "Regulation and Plasticity of Catalysis in Enzymes: Insights from Analysis of Mechanochemical Coupling in Myosin", X. Lu, V. Ovchinnikov, D. Demapan, D. Roston and Q. Cui, *Biochemistry* **56**, 1482 (2017). Copyright 2017 American Chemical Society.

open question is whether the catalysis can be fine-tuned by adjusting the polarity and solvation properties of intrazeolite void environments(12).

With these considerations in mind, in this study we focus on ATP hydrolysis in the Myosin II molecular motor. Myosins are remarkable biomolecular machines that utilize the chemical energy of ATP hydrolysis to perform diverse biological functions, such as cellular transport, signaling and muscle contraction(13). In normal myosin function, ATP hydrolysis is tightly coupled to the conformational state of the motor domain. The underpinnings of this "mechanochemical coupling" in ATPases have been studied by both experiment(14–19) and computations(20–37) for several systems, but the mechanism is not completely understood.

Myosin II (henceforth referred to as myosin) is chosen for this study because its functional cycle is well-characterized, and because several high-resolution Xray structures are available. We investigate ATP hydrolysis in the pre-powerstroke(PPS)(38) and in the post-rigor(PR)(39) states of motor domain.

The PR state binds ATP tightly but is incapable of ATP hydrolysis(17, 40). In the PPS state Switch II motif (SwII) in the active site (Fig. 6.1), moves to interact with the Switch I motif (SwI), forming a salt-bridge between Arg238 in SwI and Glu459 in SwII (the numbering of residues follows myosin II from D. *discoideum*), and a hydrogen bond between the backbone amide of Gly457 of SwII and the γ phosphate of ATP. As a result of these changes, ATP is hydrolyzed efficiently in the PPS state, (in fact, X-ray structure contains the hydrolysis transition state analog ADP·VO_4^- in the active site).

The two X-ray structures have additional differences that span different scales (Fig. 6.1). The most striking structural change is that the C-terminal converter undergoes a 60° rotation(41) (Fig. 6.1a). The converter rotation in the PR→PPS transition 'primes' the converter for the subsequent power stroke, which is triggered by interactions with actin, and is accompanied by the sequential release of hydrolysis products, phosphate and ADP. The power stroke is the main motion responsible for muscle contraction(17). The communication between the nucleotide pocket and the converter is facilitated by a "Relay helix"(17, 18) (Fig. 6.1a), which adopts a locally unwound or "kinked" conformation in the PPS state. The Relay immediately follows SwII in sequence, and thus translates the 'chemical' signals of ATP binding and hydrolysis in the active site into mechanical motions of the converter.

In this study, we focus on identifying the main features of the PPS conformation that facilitate ATP hydrolysis. Previously, the importance of aligning the lytic water molecule in an "near-in-line-attack" config-

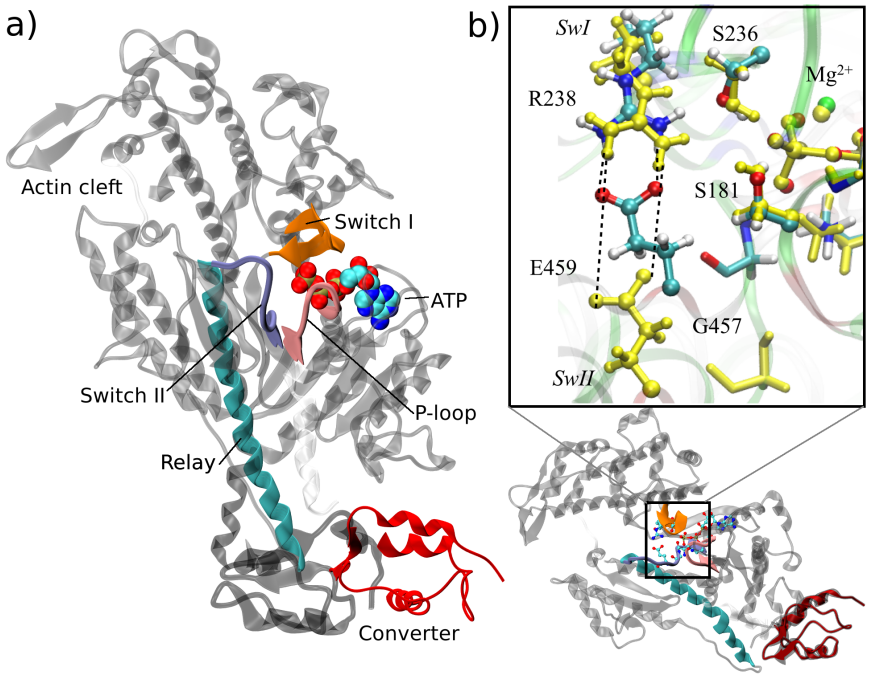


Figure 6.1: (a) Overview of the myosin motor. The main structural elements involved in the mechanochemical coupling between ATP hydrolysis and mechanical motion are shown for the PPS conformation in color ribbon; these are Switch I, Switch II and P-loop in the nucleotide binding site, and the Relay helix connected to the Converter; the rest of myosin is drawn in light gray; ATP is drawn as color spheres. The displacements of the converter and the Relay in the transition to the PR structure are indicated in dark gray. (b) Expanded view of the nucleotide pocket centered on the QM region. Local transition in the active site that largely involves the displacement of the SwII motif, leading to the break of the salt-bridge between Glu459 and Arg238 as well as displacement of Gly457 away from the γ-phosphate of ATP. Switch I and P-loop remain largely intact between the PR and PPS states.

uration has been emphasized(42). Active site water molecules in PR and PPS states feature different average positions(43) and dynamics(31). However, the free energy penalty associated with aligning the lytic water molecule at room temperature was estimated through classical molecular dynamics to be only 2-3 kcal/mol(32). This corresponds to a hydrolysis rate difference of $\sim$30-150 fold according to transition state theory, and is unlikely to be a major factor considering the large difference in the ATPase activity between the post-rigor and pre-powerstroke states(17, 43).

Previous studies focused on polar interactions involving the γ phosphate to activate hydrolysis. Applying minimum energy path (MEP) analysis and QM/MM calculations to myosin, Li et al.(31) and Kiani et al.(28) highlighted the importance of phosphate stabilization by polar interactions unique to the PPS active site, e.g. via the aforementioned Gly457 backbone amide. In the study of F_1-ATPase(29, 44) and GTPases(34, 35, 45, 46), the role of an "Arginine finger" in activating hydrolysis in specific conformational states is established. In combined DFT/MM and experimental IR study of GTPase activity in ras and ras-GAP, Gerwert and co-workers(47) noted the importance of electrostatic stabilization of the γ phosphate, including a change of relative orientation of the β and γ phosphates that promotes further polarization of the scissile P_γ-$O_{3\beta}$ bond(45, 48). However, a calculation to characterize the explicit contribution of this effect to the GTP hydrolysis barrier was not performed.

During the hydrolysis reaction, one proton of the lytic water is transferred to a nearby base, which can be either a protein side chain or the γ phosphate of the ATP (or GTP). Therefore, proton transfer is an important aspect of the hydrolysis reaction, and motifs that facilitate such transfers are likely to be part of the regulatory mechanism. In myosin, the highly conserved SwII residue Glu459 establishes a water-mediated interaction with the γ phosphate only in the PPS state, and is therefore likely to play a crucial role in the hydrolysis reaction. Consistently, mutation of either Arg238 or Glu459 leads to significant decrease in the ATPase activity, while swapping the position of the two residues preserves normal hydrolysis activity(49, 50). It is unclear whether Glu459 is involved explicitly by serving as a general base, or merely provides stabilization for the proton transfer. The former scenario was observed in a QM/MM-MEP study of myosin by Grigorenko et al.(51), and was also proposed for several other NTPases including F_1-ATPase,(29, 44) kinesin(30) and actin(52) based on QM/MM and free energy simulations.

In contrast, Kiani and co-workers(28) found that the energy barrier for the pathway in which Glu459 is

explicitly protonated is higher than in mechanisms in which it is involved indirectly. However, the barriers were computed along a path of minimum energy (rather than minimum *free* energy), and therefore exclude contributions due to the fluctuations in the local protein environment and the hydrogen bonding network(53–55). Consistently, an earlier study from our group(33) showed that, for the hydrolysis-competent PPS state, the computed MEP barriers were substantially higher ($\sim$12 kcal/mol) than the experimental measurements, which could be explained by a lack of entropic compensation. At minimum, a thermodynamically rigorous computation of the hydrolysis rate should involve an average over water and protein configurations sampled from the canonical ensemble, which would include realistic enthalpic and entropic corrections to the barrier energies.

To understand the influence of water molecules beyond the immediate coordination shell of the phosphate on the catalysis rate, Yang et al.(33) performed ATP hydrolysis calculations starting from several different snapshots obtained from molecular dynamics simulations of a model myosin structure in which the active site adopts the structure of PPS while the rest in the PR conformation. The barriers along the MEPs were found to be highly variable ($>$ 10 kcal/mol). The differences were mainly due to the higher mobility of water molecules in the active site, which prevented them from being persistently engaged in the stabilization of the hydrolysis transition state. These observations underscored the importance of sampling over multiple configurations of water molecules in the active site that follow the canonical distribution.

The above discussion above makes clear that free energy simulations of ATP hydrolysis are needed. In the current study, we perform DFTB3/MM(56–59) free energy simulations of ATP hydrolysis with the myosin head modeled in different conformational states. The DFTB3/MM approach was parameterized recently for phosphoryl transfer reactions,(58) and found to provide a semi-quantitative description of phosphoryl transfer transition states in alkaline phosphatase, including trends in the transition state structure for a series of substrates,(60) and also primary and secondary kinetic isotope effects(61). Therefore, we expect that the approach is able to capture key trends in ATP hydrolysis in different conformational states of myosin. The high computational efficiency of DFTB3 makes it possible to conduct much more thorough conformational sampling (by a factor of $\sim$100) than previous DFT/MM studies(30, 52). The DFTB3/MM method is combined with the metadynamics(62) and the string method in collective variables(63, 64) to compute minimum free energy pathways. The string method was used previously in classical MD simulations compute transition

pathways of converter rearrangement in myosin VI(64, 65). Here, the unique combination of DFTB3/MM with metadynamics and the string method (see Computational Methods) allows us to identify multiple hydrolysis reaction pathways; comparison of energetics of these pathways provides further insights into factors that regulate the hydrolysis activity.

In the following, we first summarize the computational methodology and system setup. Next, we present optimized reaction pathways and corresponding free energy profiles for ATP hydrolysis in different conformational states of myosin; these include the PPS and PR states based on the corresponding X-ray structures, as well as the closed PR model similar to that constructed in Ref. 33. The results are discussed first specially for myosin and then in the broader context of catalytic activity regulation in enzymes. We end by summarizing a few conclusions.

6.2 Computational Methods

6.2.1 Basic QM/MM setup

The enzyme model is constructed based on high-resolution (1.9 Å) X-ray structures of the myosin motor domain from D. *discoideum*. Three systems are studied: PPS with ATP bound (PDB code 1VOM(38)) at 1.9 Å resolution, PR with ATP bound (PDB code 1FMW(39)) at 2.15 Å resolution, and the closed PR model with ATP bound (see below). QM/MM calculations are carried out to obtain the reaction pathways of ATP hydrolysis and the corresponding free energy profiles. The QM region includes the tri-phosphate and part of the ribose group of ATP, Ser181, Ser236, Glu459, Arg238, the full coordination shell of the Mg^{2+} ion and any water molecules that potentially participate in the reaction. Only side chains of protein residues are included in the QM region and link atoms are introduced to saturate the valence of the QM boundary atoms using the DIV scheme(66). The QM regions are identical in all simulations except for the water molecules, because the post-rigor structure has a significantly more open active site (Fig. 6.1b). Pre-powerstroke and closed post-rigor simulations include 3 QM waters and post-rigor simulations contain 9 water molecules in the QM region. To ensure that the QM water molecules are always inside the active-site pocket in the post-rigor state simulations, the Flexible Inner Region Ensemble Separator (FIRES)(67) restraining potential is imposed. ATP is assumed to be fully deprotonated and all titratable sidechains are kept in their normal protonation states which are consistent with simple estimate of pK_a values using Poisson-Boltzmann (PB) calculations(68).

Because of the considerable computational cost associated with QM/MM free energy calculations (with more than 80 QM atoms), our approach is to use generalized solvent boundary potential (GSBP)(69, 70) in a DFTB3/MM framework(56). For metalloenzyme applications, recent DFTB3 developments have led to encouraging results in predicting reliable structural and semi-quantitative energetic properties,(71–73) including trends in phosphoryl transfer transition state properties in alkaline phosphatase(60, 61). Nevertheless, we conduct calibration calculations for DFTB3 using B3LYP calculations for both model and enzyme systems (see **Supporting Information**), and the results support the general trends emerged from the DFTB3/MM simulations. In the GSBP setup, the system is partitioned into a 24 Å spherical inner region centered at the Mg^{2+} ion, with the remaining portion of the system in the outer region. Newtonian equations-of-motion are solved for the MD region (within 20 Å), and Langevin equations-of-motion are solved for the buffer region (20-24 Å) with a temperature bath of 300 K. Protein atoms in the buffer region are harmonically constrained with force constants determined from the crystallographic B-factors(74). Protein atoms in the MM region are described by the all-atom CHARMM22 force field(75) and water molecules are described with the TIP3P model(75). All water molecules in the inner region are subject to a weak GEO type of restraining potential to keep them inside the inner sphere. The static field due to outer-region atoms, $\phi_s^{(io)}$, is evaluated with the linear PB equations. The reaction field matrix **M** is evaluated with 625 spherical harmonics. In the PB calculations, the protein dielectric constant is set to 1, the water dielectric constant is set to 80, and the salt concentration is set to 0. To be consistent with the GSBP protocol, the extended electrostatic model(76) is used to treat the electrostatic interactions among inner region atoms where interactions beyond 12 Å are calculated with multipolar expansions. Our previous studies(77, 78) indicated that the GSBP protocol is reliable for a site well shielded from the bulk solvent. Prior to the free energy simulations, we run QM/MM simulations to first equilibrate the structures for 500 ps with a time step of 0.5 fs.

6.2.2 A Closed PR Structural Model

The closed PR model is a conformation in which SwII is displaced to form a closed active site in the otherwise PR structure; it was first established by Yang et al.(33) to probe the contribution of residues beyond the "immediate coordination shell" of the γ phosphate to the ATP hydrolysis activity. Starting from the post-rigor X-ray structure, a 1D RMSD restraint is applied to selected atoms in the active site (backbones of SwI 232-238, SwII 454-459 and P-loop 179-186 and side chain of residue 238, 459, 461) with respect to the PPS

crystal structure. The force constant is 500 kcal·mol^{-1}·Å^{-2} and an offset is set to be 0.5 Å with respect to the reference structure. The model is carefully equilibrated with 500 ps before extensive QM/MM string free energy simulations. As discussed in the **Supporting Information**, the active site of the closed PR model remains close to that of the PPS state; we observed, however, that SwII is slightly more flexible, and the salt-bridge between Glu459 and Arg238 sidechains has a minor tendency to deviate from planarity.

6.2.3 Metadynamics

We explore ATP hydrolysis mechanisms in myosin by a combination of metadynamics(62, 79) simulations and string methods(63). Metadynamics is performed by the PLUMED(80) package interfaced to CHARMM(81). The collective variables (CVs) for metadynamics include the anti-symmetric stretch involving the lytic water and the scissile P-O bond ($P_\gamma O^{ATP}_{3\beta} - P_\gamma O^{W1}$) and the ζ coordinate(82) that describes proton transfers in the active site. The proton donor, acceptor and residues that potentially participate in the proton transfer for various pathways are summarized in the **Supporting Information**. For each pathway, SHAKE(83) is applied to bonds involving hydrogens that do not participate proton transfers. The well-tempered metadynamics method(84) is utilized with a bias factor of 20$\sim$30 depending on the estimated barrier height. Gaussian potentials are placed every 50 fs; the Gaussian height is initially set to 0.1 kcal/mol, and the Gaussian width for the antisymmetrical stretch and ζ coordinate are set to 0.1 Å and 0.05, respectively. The simulations ran for 2 ns for each mechanism and recrossing events between two free energy minima are observed two to three times for each pathway before the simulation is terminated.

The use of the ζ CV is advantageous because neither the identity nor the number of the groups that facilitate the proton transfer need to be specified prior to the simulation(85). However, a disadvantage is that the ζ CV cannot not separate distinct proton transfer pathways, e.g., different configurations that correspond to physically distinct pathways can be mapped to the same value of ζ, which also implies that ζ is not generally a kinetically meaningful quantity. To overcome this limitation, metadynamics is used here only to explore the ensemble of configurations sampled by various proton transfer pathways. These configurations are subsequently used to generate initial pathways for free energy simulations using string method, discussed below.

6.2.4 String Method

Exploring a high-dimensional free energy surface of a biomolecular system is typically computationally prohibitive, and an alternative approach is to search for a minimum free energy path (MFEP) with the string method(63). In this case, a one-dimensional curve connecting two wells is the target, regardless of the number of CVs to represent the path, thus the computational cost scales much more favorably compared to multi-dimensional umbrella sampling/metadynamics. We define the CV space spanned by a set of order parameters most relevant to the reaction process $\vec{\xi}(\mathbf{X}) = (\xi_1(\mathbf{X}), \xi_2(\mathbf{X}), \cdots, \xi_n(\mathbf{X}))$ and the MFEP in this space is a string defined by the condition that the projection of the mean force perpendicular to the path vanishes,

$$\left[\mathbf{M}(\vec{\xi}) \cdot \nabla W(\vec{\xi})\right]^{\perp} = 0 \tag{6.1}$$

where $\mathbf{M}_{i,j} = \langle \nabla_{X^*} \vec{\xi}_i \cdot \nabla_{X^*} \vec{\xi}_j \rangle_{\vec{\xi}=\vec{\xi}}$ is the metric tensor associated with the transformation to the space of collective variables, and $W(\vec{\xi})$ is the potential of mean force. In practice, the MFEP is optimized with the chain-of-states methods(86, 87). The string is discretized into a set of N equidistant images $\vec{\xi}_i$, with $i \in [1 \ldots N]$. A separate simulation system is assigned to each image. Starting from an initial condition for the string, the images are updated iteratively until a discretized version of Eq. 6.1 is satisfied. A straightforward iterative method is to first advance the string using steepest descent:

$$\vec{\xi}_i^* = \vec{\xi}_i^m - \frac{\Delta t}{\gamma} \mathbf{M}(\vec{\xi}_i^m) \nabla W(\vec{\xi}_i^m), \tag{6.2}$$

followed by reparametrization. $\mathbf{M}$ and ∇W in Eq. 6.2 are computed from simulations with harmonic restraints $k \times [\vec{\xi}(\mathbf{X}) - \vec{\xi}_i]^2$, and the parameters Δt and γ are adjusted empirically to accelerate convergence to the MFEP(64).

The free energy profile is the integral along the converged path

$$\Delta W(\alpha) = W(\vec{\xi}(\alpha)) - W(\vec{\xi}(0)) = \int_0^{\alpha} \sum_{i=1}^{n} \frac{d\xi_i(\alpha')}{d\alpha'} \frac{\partial W(\vec{\xi}(\alpha'))}{\partial \xi_i} d\alpha', \tag{6.3}$$

in which α is used to parametrize the string curve.

In the present study, 32 string images are used for each reaction pathway; the initial guess of the string images is constructed from the approximate reaction path obtained in metadynamics simulations using the $O^{W1}P_{\gamma}O^{ATP}_{3\beta}$ anti-symmetric stretch and ζ as the collective variables as described above, while the CV set used to span each string is much more expanded and involves 13$\sim$19 active site distances, as summarized in the

Supporting Information. The values employed for the harmonic force constant and friction in the string method are $k = 400$ kcal·mol^{-1}·Å^{-2} and $\gamma = 20$ ps^{-1} (determined by trial and error). 1.5 ns MD simulation at 300 K with a time step of 0.5 fs is performed for each string image. For the last 500 ps, image centers are fixed and free energy gradients are calculated and stored for data analysis. String evolution and reparametrization are performed every 0.5 ps.

6.2.5 Average Electrostatic Potential Analysis

For each image of interest (see Sect. "ATP hydrolysis in the closed post-rigor model"), which corresponds to either the reactant or transition state in the hydrolysis reaction, 2,000 configurations are taken from the corresponding trajectories during the final stage of string simulations. In the PB calculations, the protein is assigned a dielectric constant of 1, and all explicit water molecules are retained and regarded as part of the solute (thus is assigned to a dielectric constant of 1); the implicit bulk solvent has a dielectric constant of 80. As in the GSBP calculations, PB calculations with a smooth dielectric boundary are carried out. Charges on all atoms in the QM region and all water molecules are turned off, and the average electrostatic potentials at grid points within the van der Waals surface of the sites of interest are recorded.

6.3 Results

6.3.1 The Reactant and Product Structures in the Pre-powerstroke State

The first issue concerns the reactant and product of the ATP hydrolysis. It is well accepted that a water molecule in the active site is the nucleophile and the γ-phosphate serves as the ultimate base for its own hydrolysis. The end-state structures of the hydrolysis reaction in myosin are depicted in Fig.6.2. In the reactant state (Fig. 6.2a), W1 and W2 water molecules form a hydrogen-bonding network that connects the γ-phosphate of ATP and Glu459. In the presence of such a network, the lytic water W1 adopts a favorable position to conduct an in-line nucleophilic attack. O^{W1} – P_{γ}^{ATP} distance is stable during the MD simulations around 3.2 Å, and the O^{W1} – P_{γ}^{ATP} – $O_{3\beta}^{ATP}$ angle fluctuates around 170°. The active site undergoes small structural changes when the ATP molecule is hydrolyzed to ADP·Pi in the product state, as indicated in Fig. 6.2b. The inorganic γ-phosphate retains the H-bond to the W2 water and makes a H-bond to the β-phosphate group of the ADP. In the present calculations, the proton between the β- and the γ-phosphate is bonded to the β-phosphate rather than the γ-phosphate. However, the energies of ADP·$H_2PO_4^-$ and ADPH·HPO_4^{2-} are may

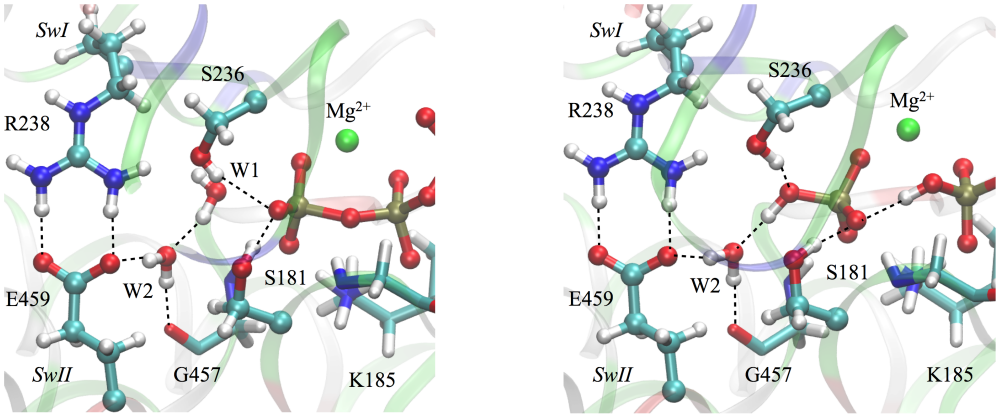


Figure 6.2: The reactant and product states of ATP hydrolysis in myosin simulated by DFTB3/MM (a) Reactant ATP. (b) Product ADP·P_i. Ligands of Mg^{2+} are omitted for clarity. Key hydrogen bonds are highlighted with dotted lines. Note that these structures do not represent energy-optimized stationary points but rather selected structures in the trajectory.

be too close in energy (e.g., ∼2-3 kcal/mol as found in Ref.28), to be distinguished reliably by DFTB3, which has limited accuracy predicting proton affinity in phosphate species(56, 58). However, the precise location of the proton is not expected to have a significant effect on the reaction mechanisms described below, since the geometries of the two possible end structures are very close.

The positions of other catalytically relevant residues such as Lys185, Ser236, Ser181, Gly457 and P-loop backbones remain almost identical between the reactant and the product (the RMSD of residues listed above is 0.36 Å between the average reactant/product structures). Moreover, the salt bridge between the SwI and SwII regions is intact; the coordination shell of Mg^{2+} in the product also remains the same as in the reactant(59). The similarity between reactant and product structures is consistent with the reversibility of ATP hydrolysis in myosin(88, 89).

6.3.2 Reaction Pathways in the Pre-powerstroke State

We have observed several distinct reaction pathways for ATP hydrolysis in the pre-powerstroke state. A common feature among these pathways is that the P_γ-$O_{3\beta}$ bond breaks prior to complete bond formation between P_γ and the lytic water (W1); i.e., the reaction is concerted in nature with a loose transition state, since

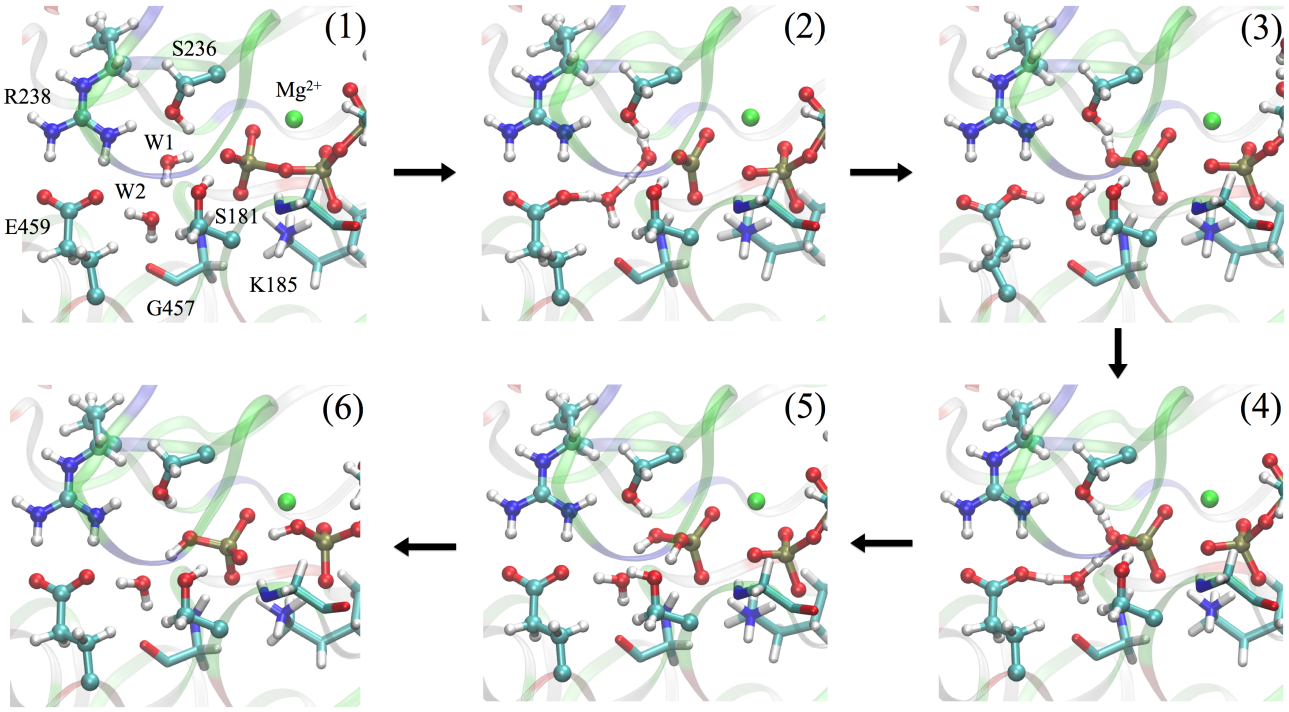


Figure 6.3: Representative structures along the MFEP obtained from the W2-E459 mechanism. The structures are numbered according to the locations identified in Fig. 6.5.

the metaphosphate species does not correspond to a stable intermediate(90). The lytic water readily associates with the metaphosphate-like species and the free energy bottleneck is largely associated with the transfer of the proton in the lytic water that ultimately ends up on one of the dissociation products (either the inorganic phosphate or ADP); the proton transfer can be mediated with different polar groups in the active site, leading to distinct pathways with different free energy profiles. We note that the different pathways show similar free energy differences for the overall hydrolysis reaction, indicating that the computed free energy profiles are converged. In the following, we first discuss several low-energy pathways; several higher-energy pathways are then discussed in the next subsection.

The pathway with lowest free energy profile involves Glu459 and W2 directly in the proton transfers; we refer to this pathway as the W2-E459 mechanism. Representative structures along the MFEP are illustrated in Fig. 6.3. Immediately following the P_{γ}-$O_{3\beta}$ bond cleavage, proton transfer from the lytic water to the carboxylate of Glu459 proceeds through a mediating water molecule W2 (Fig. 6.3 (2)), leading to a high-

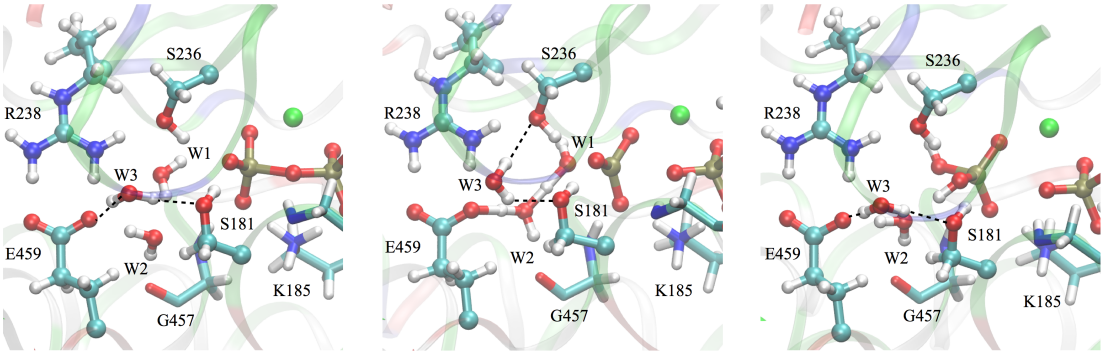


Figure 6.4: Rearrangement of hydrogen-bond associated with W3 in the W2-E459 mechanism. Key hydrogen bonds are highlighted with dotted lines.

energy quasi-stable intermediate state (Fig. 6.3 (3)). During the same transition, W1 becomes a OH^- and covalently bonds to P_γ, leading to the formation of an inorganic phosphate. Next, another proton transfer proceeds from the carboxylic acid of Glu459 to the inorganic phosphate via the same mediating water molecule W2 (Fig. 6.3 (4)), generating the doubly protonated inorganic phosphate (Fig. 6.3 (5)), which ultimately shares the proton with the $O_{3\beta}$ of ADP, leading to the final hydrolysis products (Fig. 6.3 (6)).

We note that the H-bond network in the active site is not optimal prior to the first proton transfer, because another active site water molecule W3 donates a H-bond to the Glu459 side chain (Fig. 6.4a). Therefore, a rearrangement of H-bond network is required to facilitate the reaction: W3 breaks its H-bond with Glu459 and rotates to become H-bonded to Ser236 instead (Fig. 6.4b). After Glu459 delivers the proton to the γ-phosphate, the H-bond between W3 and Glu459 is restored (Fig. 6.4c). We note that such H-bond reorganization is difficult to take into account using only a 'zero-temperature' minimum-energy path (MEP) analysis, because at least some thermal activation may be required to overcome barriers to W3 rotation, which is orthogonal to the (reaction) coordinate of the proton transfer. This may explain why the W2-E459 pathway was found to have higher barriers in the study of Kiani and Fischer(28).

The one-dimensional free energy profile is depicted in Fig. 6.5a. The first barrier corresponds to the proton transfer between the lytic water to the side chain of Glu459, while the second barrier is for the proton transfer from Glu459 to the inorganic phosphate; the relatively flat plateau between the two peaks corresponds to the

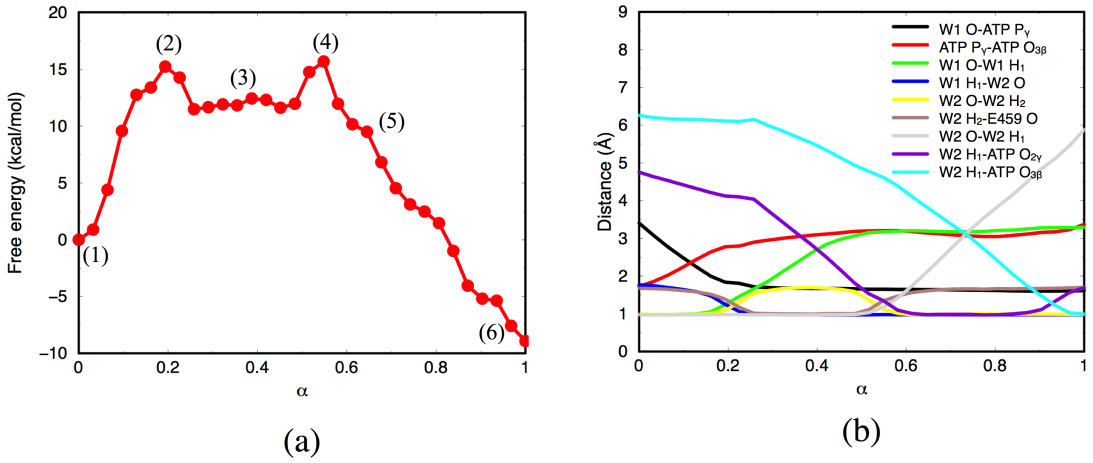


Figure 6.5: (a) The 1D free energy profile along the MFEP obtained from W2-E459 mechanism. Each dot corresponds to an image along the string. (b) Values of the important distances along the MFEP.

intermediate structure where the carboxylate of Glu459 is transiently protonated. Fig. 6.5b shows the changes in important bond distances along the MFEP. In the early stage of the reaction, d(O^{W1} - P_{γ}^{ATP}) decreases and d(P_{γ}^{ATP} - $O_{3\beta}^{ATP}$) increases. Dissociation of the $P_{\gamma} - O_{3\beta}$ bond is closely correlated with the distance of the new P – O bond with the nucleophile. Subsequently, movements along d(O^{W1} - H_1^{W1}), d(H_1^{W1} - O^{W2}), d(O^{W2} - H_2^{W2}), d(H_2^{W2} - O^{E459}) are synchronous, indicating a concerted proton transfer process.

Our simulations predict the activation free energy of 15-16 kcal/mol. Experimental measurements obtained ATPase rate constants(91–93) that range from 1 to 100 s^{-1}, which correspond to free energy barriers in the range of 15-17 kcal/mol (converted based on transition state theory and a prefactor of kT/h at room temperature). Therefore, the computational and experimental barrier heights are in good agreement, although single point B3LYP calculations along MEP indicate that the DFTB3 barrier appears systematically too low by at least 2∼3 kcal/mol (see **Supporting Information**). For the exothermicity, the DFTB3/MM value is about 10 kcal/mol. Using the difference between DFTB3 and B3LYP/aug-cc-pVTZ for the exothermicity of a model hydrolysis reaction, which was 9 kcal/mol(see **Supporting Information**), the corrected value is range of 1∼2 kcal/mol, which is in rough agreement with the experimental finding that ATP hydrolysis in the pre-powerstroke state is nearly thermoneutral(40, 91). This small amount of free energy drop reflects that force generation in myosin is mainly driven by ATP/actin binding and the hydrolysis step mainly functions as a switch to move the system forward into the next stages of the functional cycle(20, 32, 94).

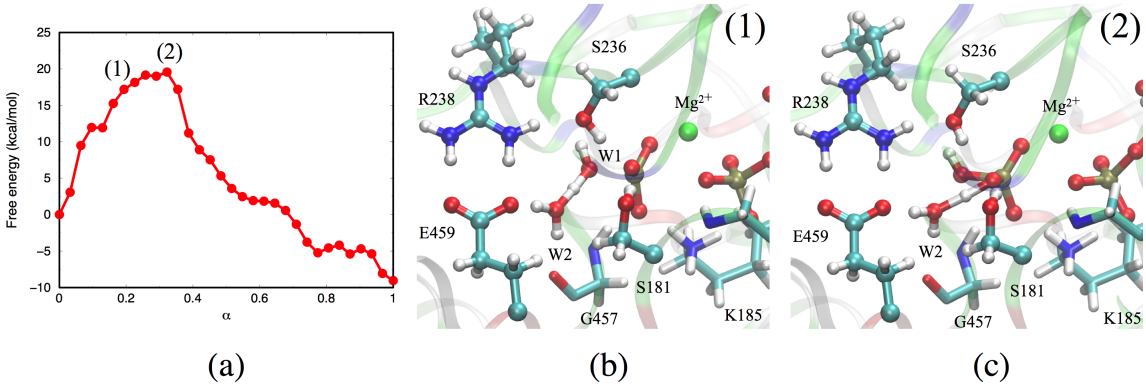


Figure 6.6: (a) Free energy profile along the MFEP obtained from W2 mechanism. Each dot corresponds to an image along the string. (b) Proton transfer from W1 to W2. (c) Proton transfer from H_3O^+ to $HP_\gamma O_4^{2-}$. Note that these structures do not represent energy-optimized stationary points but selected structures in the specific string image.

A pathway that does not explicitly protonates Glu459 involves a proton transfer through W1→W2→$P_\gamma O_3^-$, and it is named here the W2 mechanism. Here W2 transiently becomes a H_3O^+ hydronium (Fig. 6.6b), which is stabilized by the negatively charged Glu459 sidechain; slight tilting of the hydronium allows it to transfer a proton back to $P_\gamma O_3^-$ (Fig. 6.6c). This pathway has a barrier of 19 kcal/mol (Fig. 6.6a), which is slightly higher than the W2-E459 mechanism. The competition between the W2 mechanism and the W2-E459 mechanism can be traced back to the pK_b difference between Glu459 and W2 water in the enzyme environment. Single point calculations of potential energy barriers at the B3LYP level (see **Supporting Information**) further supports that the W2 mechanism is less favorable than the W2-E459 pathway.

Another route to transfer proton from W1 to $P_\gamma O_3^-$ involves the intervening water W2 and Ser181: W1→W2→Ser181→$P_\gamma O_3^-$. The W2-S181 mechanism was considered most favorable by the B3LYP/MM MEP studies by Kiani et al.(28) The pathway requires W2 to first break H-bond with the carbonyl group in Gly457 and to form a new H-bond with the side chain of Ser181. The next event of the reaction is similar to the W2 mechanism: formation of a "metaphosphate" and proton transfer occurs from W1 to W2; the proton is further transferred to the hydroxyl oxygen of Ser181 then the inorganic phosphate and finally ADP. The most important intermediates are shown in Fig. 6.7b-c, and the overall free energy barrier is 22 kcal/mol. The higher free energy barrier

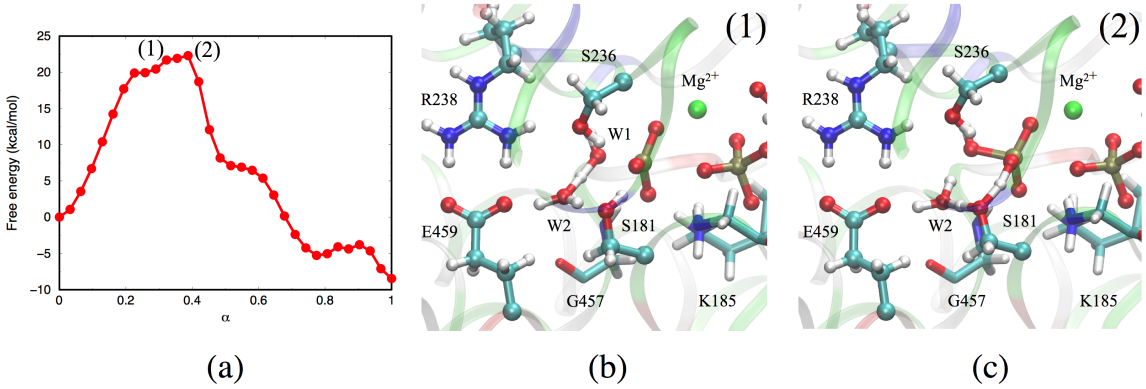


Figure 6.7: (a) Free energy profile along the MFEP obtained from the W2-S181 mechanism. Each dot corresponds to an image along the string. (b) Proton transfer from W1 to W2. (c) Concerted proton transfer from H_3O^+ to S181 and from S181 to $HP_\gamma O_4^{2-}$. Note that these structures do not represent energy-optimized stationary points but selected structures in the MD trajectories.

compared to the W2-E459 mechanism is likely due to the extra penalty for the initial rearrangement of H-bond network associated with the key W2 water. However, single-point energy calculations using QM/MM with B3LYP suggest that the barrier energies for W2-E459 are underestimated with DFTB3/MM (see **Supporting Information**), and therefore the W2-S181 and W2-E459 should be considered competitive.

6.3.3 Higher free energy pathways in the pre-powerstroke state

Highly conserved active site serine residues such as Ser236 and Ser181 were proposed(31, 33) to serve as proton relays during the hydrolysis process, allowing better stereochemistry for the proton transfers(42, 95). In the S236 mechanism, for example, the lytic water W1 first transfers one of its protons to Ser236, which in turn transfers its hydroxyl proton to the γ-phosphate through a six-membered H-bond ring (Fig. 6.8b). The rate-limiting barrier along the free energy profile (Fig. 6.8a) is 25 kcal/mol. Alternatively, a "cycling" pathway, referred to as the S236-W3-S181 mechanism, is also observed; proton transfer occurs along a path W1→Ser236→W3→S181→$P_\gamma O_3^-$ (Fig. 6.9b). During this transition, $P_\gamma \cdots O^{W1}$–$H^{W1} \cdots O^{S236}$–$H^{S236} \cdots O^{W3}$–$H^{W3} \cdots O^{S181}$–$H^{S181} \cdots$O-P_γ switches to P_γ–$O^{W1} \cdots H^{W1}$–$O^{S236} \cdots H^{S236}$–$O^{W3} \cdots H^{W3}$–$O^{S181} \cdots H^{S181}$–O–P_γ. This involves crossing a barrier of 24 kcal/mol (Fig. 6.9a). In the above two mechanisms, the barriers

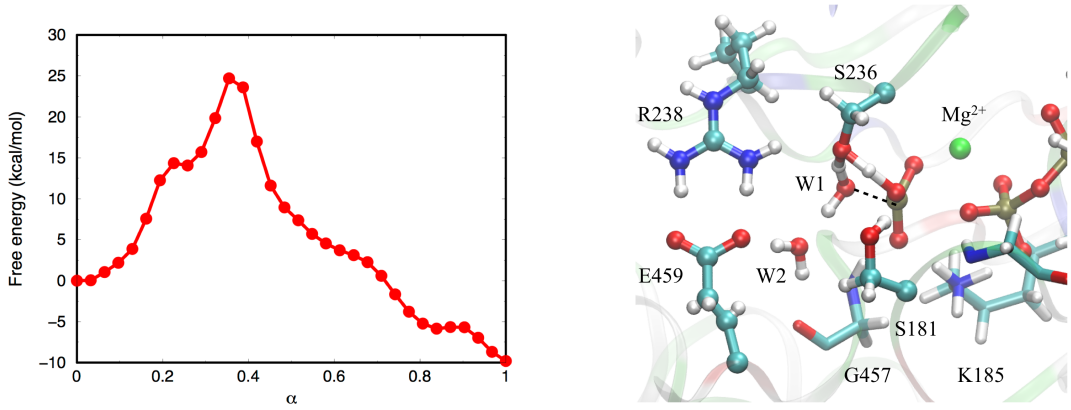


Figure 6.8: (a) The 1D free energy profile along the MFEP obtained from the S236 mechanism. Each dot corresponds to an image along the string. (b) Transition structure in which W1 transfers one of its protons to Ser236, which in turn transfers its hydroxyl proton to the γ-phosphate

are higher than in the other mechanisms because the center of excess charge in the proton wire lies further away from W2 and Glu459 (see further discussions below).

6.3.4 ATP Hydrolysis in the Post-rigor State

In the PR state of myosin, SwII moves away from the active site and the salt bridge between Glu459 and Arg238 is broken (Fig. 6.1b). This has at least two consequences to the hydrolysis activity. First, both Glu459 and Gly457 are further away from ATP and the backbone amide of Gly457 can no longer make an H-bond to the γ-phosphate. This implies that Glu459 is too far away from the reaction site to participate in the proton transfer. Second, the active site is more accessible to bulk water and water molecules in the active site exhibit significantly larger motions(31). Because of these motions, there are no water molecules that can maintain stable contacts with the lytic water. Therefore below we only discuss the S236 mechanism.

In the transition state of the S236 pathway in the PR state, similar to the situation in the PPS state (Fig. 6.8b), the lytic water proton is shared between the donor O^{W1} and the acceptor O^{S236}, while the hydroxyl proton in Ser236 is transferred to the γ-phosphate oxygen (Fig. 6.10b). The free energy profiles (compare Fig. 6.8a and 6.10a), however, are rather different. The hydrolysis reaction is substantially less exothermic in the post-rigor state and the barrier is $\sim$ 40 kcal/mol, which is nearly 25 kcal/mol above the lowest free energy

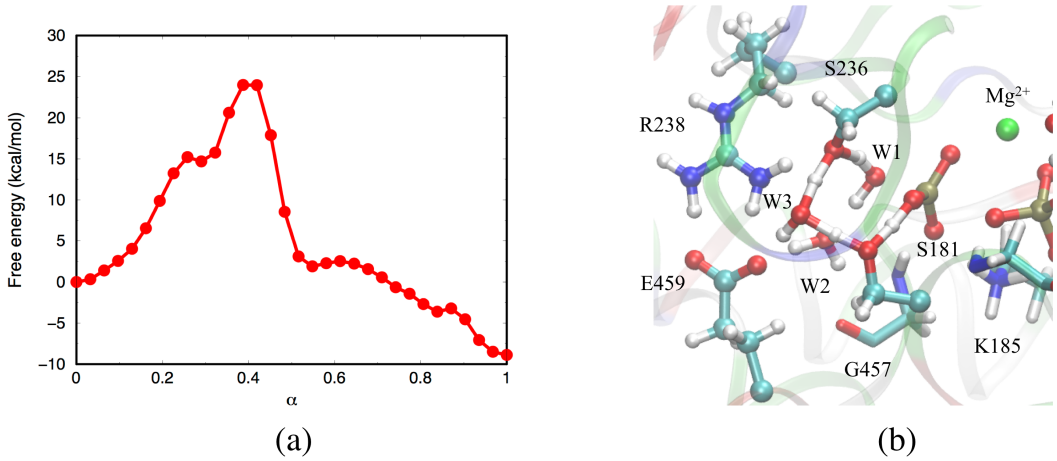


Figure 6.9: (a) Free energy profile along the MFEP obtained from the S236-W3-S181 mechanism. (b) Transition structure in which proton transfers along the proton wire W1→Ser236→W3→S181→$P_\gamma O_3^-$.

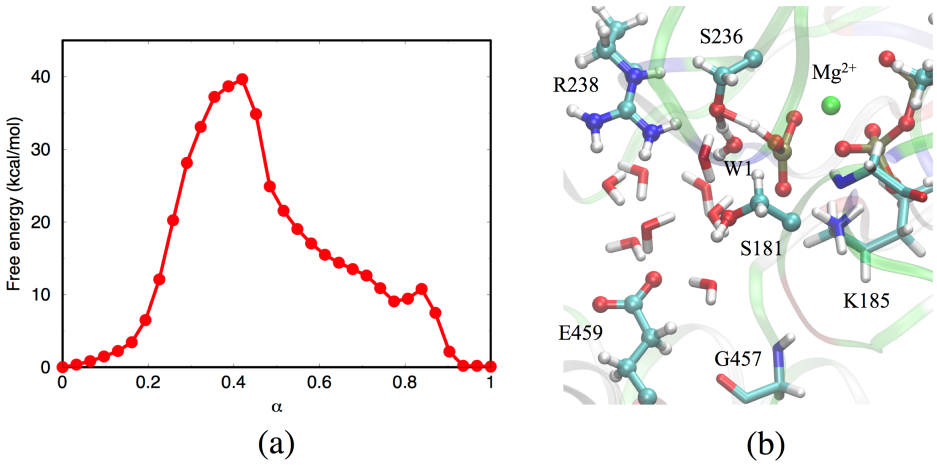


Figure 6.10: (a) The 1D free energy profile along the MFEP obtained from the S236 mechanism in the post-rigor state. Each dot corresponds to an image along the string. (b) Transition structure in which proton transfers between W1 and S236, and S236 and γ-phosphate.

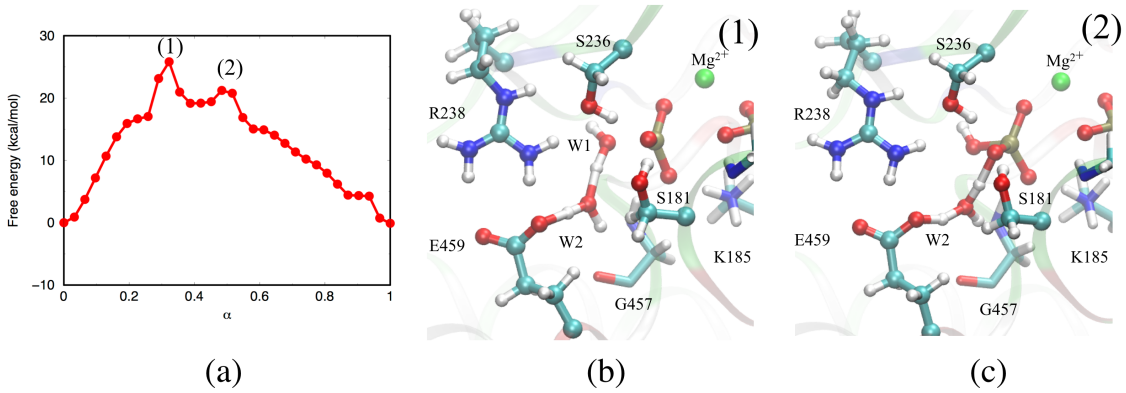


Figure 6.11: (a) Free energy profile for the W2-E459 mechanism in the closed post-rigor structural model. (b) Proton transfer from W1 to the carboxylate group in E459 via W2. (c) Proton transfer from -COOH in E459 to $H_\gamma PO_4^{2-}$ through W2. The structures are taken from MD simulation trajectories.

barrier in PPS state and ∼15 kcal/mol higher than the same S236 pathway in the pre-powerstroke state; in fact, it is about 5-10 kcal/mol above the barrier for uncatalyzed ATP hydrolysis in aqueous solution(96). Such a high barrier is qualitatively consistent with the lack of hydrolysis activity in PR state, an important feature that avoids futile hydrolysis in the functional cycle of myosin(17, 40).

6.3.5 ATP hydrolysis in the closed post-rigor model

An important question is the extent to which ATPase activity is influenced by groups beyond the immediate coordination shell of the γ-phosphate. To address this question, we study the hydrolysis reaction in the closed post-rigor structural model(33), in which the immediate environment of the γ-phosphate resembles that in PPS, while the rest adopts the PR conformation.

As shown in the **Supporting Information**, most structural features of the active site in the closed PR model resemble those in PPS in MD simulations. Therefore, we focus on the W2-E459 mechanism, which is found to be most favorable in the PPS. The approximate transition states (shown in Fig. 6.11b, c) are rather similar to the ones for PPS (see Fig. 6.3(2,4)). However, somewhat surprisingly, the free energy barrier is substantially higher in the closed PR model, by about 9 kcal/mol, and the reaction is also notably less exothermic by approximately the same amount (compare Fig. 6.11a and 6.5a).

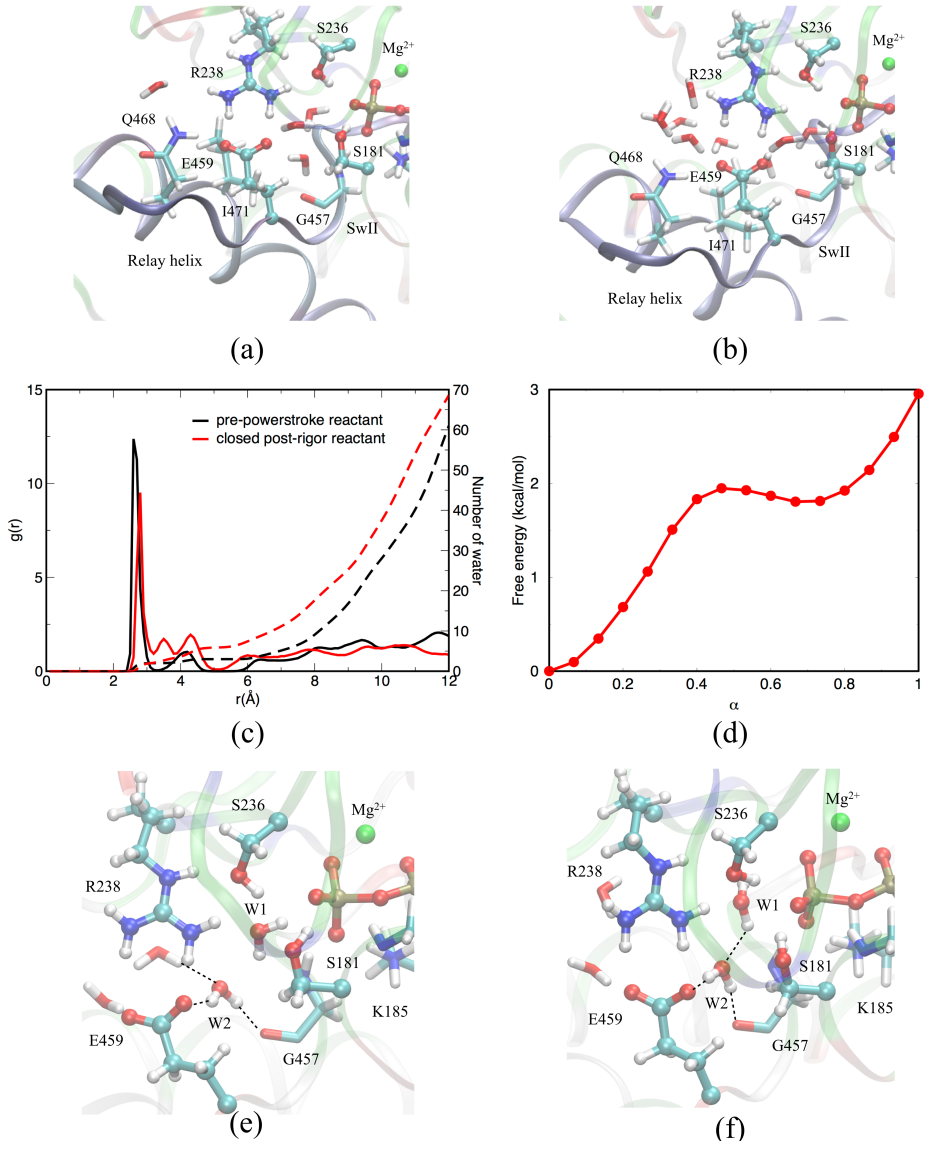


Figure 6.12: (a) Comparison of water distribution around carboxylate oxygen of Glu459 (the one closer to water W1) in the PPS and closed PR structures. (b) Free energy profile for water reorganization in the closed PR state. (c) Initial state of water orientation. (d) W1 and W2 reorient their dipoles before hydrolysis.

The differences are due in part to distinct water structures in the closed post-rigor model compared to the PPS state. Even though the active site in the model has been closed (Fig.6.12a), it still is more accessible by the solvent than in the PPS structure. This is essentially because residues in the N-terminus of the relay helix, which is immediately C-terminal to SwII, remain in the PR state.

W2 is located further away from the γ-phosphate and receives additional stabilization from bulk water (see Fig. 6.12c). As the result, W1 and W2 need to reorganize to form the proper H-bond structure required to facilitate the hydrolysis reaction through the W2-E459 mechanism (see Fig. 6.12d); e.g., the distance between O^{W2} and P_γ changes from 5.7 Å to 4.5 Å during this transition. To estimate the free energy cost associated with this "water reorganization" step, we conduct separate string calculations for the transition between structures shown in Fig. 6.12c and d. As shown in Fig. 6.12b, the free energy change is ∼3 kcal/mol, similar to our previous findings(32) .

To understand the origin for the remaining ∼6 kcal/mol difference in the hydrolysis barriers in the pre-powerstroke and closed post-rigor states, we analyze electrostatic potentials in the active site with Poisson-Boltzmann calculations. Since the barriers in the W2-E459 pathway mainly involve proton transfers among W1, W2 and E459, we focus on the electrostatic potential differences on the relevant oxygen atoms (Table 6.1). To explicitly evaluate contributions from the protein environment, charges of QM atoms and all water molecules are set to zero; including water contributions does not alter the qualitative trends.

The electrostatic potential differences summarized in Table 6.1 reflect to what degree the protein environment favors the proton transfer among these sites. In PPS, the electrostatic potential difference between O^{W1} and O^{W2} ($\Delta\phi_{W1;W2}$) is ∼25 kcal·mol^{-1}·e^{-1} in both reactant and TS1. The equality of the potential suggests that the protein environment is preorganized for the proton transfer between W1 and W2; e.g. significant dipolar reorganization is not needed to reach the transition state from the reactants. Similar trend is found for the difference between O^{W2} and O^{E459} ($\Delta\phi_{W2;E459}$). In the closed PR model, $\Delta\phi_{W1;W2}$ is also ∼25 kcal·mol^{-1}·e^{-1} in the reactant configuration (following the completion of the water reorganization shown in Fig. 6.12c-d); however, the value drops substantially to ∼13 kcal·mol^{-1}·e^{-1} in TS1. For $\Delta\phi_{W2;E459}$, the TS1 value in the closed PR model is also smaller by 50% compared to PPS. Therefore, compared to PPS, the transferring proton in the closed PR structure experiences a smaller electrostatic driving force going from W1 to W2 (and from W2 to E459) in the transition state; moreover, the large difference in $\Delta\phi_{W1;W2}$ between the reactant and

Table 6.1: The average electrostatic potential difference ($\Delta\phi$ in kcal·mol^{-1}·e^{-1}) between sites of interest due to the protein environment (excluding QM atoms and water molecules) in the reactant and transition state configurations for the pre-powerstroke and closed post-rigor structures

Structural model	Reactive configuration	Residues excluded	$\Delta\phi_{W1;W2}$ (O^{W1} - O^{W2})	$\Delta\phi_{W2;E459}$ (O^{W2} - O^{Glu459})
pre-powerstroke	Reactant[b]	None	25.2 (1.7[a])	9.1 (1.3)
pre-powerstroke	TS1[c]	None	24.8 (2.2)	9.9 (2.6)
pre-powerstroke	Reactant	G457	7.6 (1.7)	17.8 (1.3)
pre-powerstroke	TS1	G457	6.7 (2.3)	17.4 (2.6)
pre-powerstroke	Reactant	S456	18.9 (1.7)	-1.4 (1.1)
pre-powerstroke	TS1	S456	20.2 (2.1)	-0.8 (2.6)
pre-powerstroke	Reactant	$\leq$ 5 Å of O^{W1}	-12.1 (1.7)	9.2 (1.3)
pre-powerstroke	TS1	$\leq$ 5 Å of O^{W1}	-12.8 (2.1)	7.7 (2.5)
"closed" post-rigor	Reactant[d]	None	25.6 (1.4)	2.8 (1.5)
"closed" post-rigor	TS1[e]	None	12.6 (2.9)	4.7 (2.6)
"closed" post-rigor	Reactant	G457	6.9 (1.2)	11.1 (1.4)
"closed" post-rigor	TS1	G457	-2.3 (2.8)	7.8 (2.6)
"closed" post-rigor	Reactant	S456	20.5 (1.4)	-6.6 (1.4)
"closed" post-rigor	TS1	S456	11.4 (2.3)	5.2 (2.1)
"closed" post-rigor	Reactant	$\leq$ 5 Å of O^{W1}	1.3 (0.9)	10.0 (1.3)
"closed" post-rigor	TS1	$\leq$ 5 Å of O^{W1}	-0.6 (1.3)	11.4 (1.0)

[a] The values in parentheses are statistical errors calculated from block average; [b] Image 3 of the string in the pre-powerstroke state; [c] Image 6 of the string in the pre-powerstroke state; [d] Image 7 of the string in the closed post-rigor structural model (i.e., active site water reorientations described in Fig. 6.12 has been completed); [e] Image 10 of the string in the "closed" post-rigor structural model;

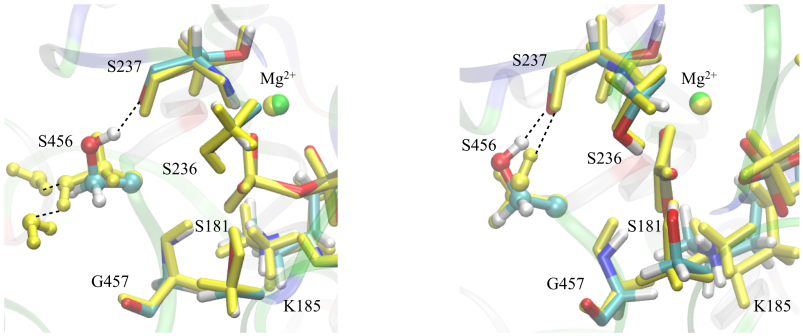
TS1 in the closed PR model indicates that the protein environment undergoes more significant dipolar reorganization compared to the pre-powerstroke state. These two factors (less electrostatic driving force and larger reorganization) combine to yield a higher barrier in the closed post-rigor model compared to the pre-power stroke state.

In an attempt to identify residues that dictate the drop in driving force and increase in reorganization in the closed PR model, perturbative analysis is conducted in which the partial charges of specific residues are set to zero and $\Delta\phi$ values are re-evaluated. In light of the discussion above, we focus on the value of $\Delta\phi^{TS1}_{W1;W2}$ and the value of $\Delta\Delta\phi_{W1,W2}$ between reactant and TS1. As shown in Table 6.1, when the contributions from residues within 5 Å from O^{W1} are excluded, $\Delta\Delta\phi_{W1,W2}$ is near zero for both closed PR and PPS structures, and $\Delta\phi^{TS1}_{W1;W2}$ is in fact more unfavorable (i.e., more negative) in the pre-powerstroke state. These observations suggest that the larger electrostatic reorganization and lower electrostatic driving force for the closed post-rigor model is dictated by nearby ($\leq$ 5 Å from W1) residues. Further looking into contributions of these residues, the perturbation results reveal major contributions from Gly457 and Ser456. In PPS, removing Gly457 and Ser456 contributions decreases $\Delta\phi^{TS1}_{W1;W2}$ by $\sim$18 and $\sim$4 kcal·mol^{-1}·e^{-1}, respectively; the corresponding effects on $\Delta\Delta\phi_{W1,W2}$ are small and close to be $\pm$1 kcal·mol^{-1}·e^{-1}. In the closed PR model, removing Gly457 and Ser456 contributions decreases $\Delta\phi^{TS1}_{W1;W2}$ by $\sim$15 and $\sim$ 1 kcal·mol^{-1}·e^{-1}, respectively; the corresponding effects on $\Delta\Delta\phi_{W1,W2}$ are $\sim$ 4 kcal·mol^{-1}·e^{-1}. Therefore, compared to PPS, Gly457 and Ser456 contribute less to the electrostatic driving force and more to the electrostatic reorganization in the closed post-rigor structural model. Analysis of conformations sampled in different MD simulations suggests that, at least for Ser456, the difference is due to different orientations of the side chain atoms in the PPS vs. closed PR structures due to differences in the local solvation environment (see Fig. 6.13).

6.4 Discussion

6.4.1 Mechanism of ATP hydrolysis in the pre-powerstroke state of myosin

The mechanism of ATP/GTP hyrdolysis in various molecular motors, GTPases and actin has been discussed extensively in recent years. For a number of insightful and recent review articles, see Refs.(35, 37, 97). In this study, we performed free energy simulations of ATP hydrolysis in different conformation of myosin II using a combination of QM/MM and enhanced sampling methods.



(a) Reactant (b) TS

Figure 6.13: Superposition of average positions of several active site residues in the (a) reactant and (b) TS windows in the PPS and closed PR (yellow) simulations. The most visible difference is that Ser456 adopts different orientations in the reactant state window due to differences in the solvation environment.

The hydrolysis pathways characterized here are broadly consistent with previous studies, and in particular, with the recent analysis of Kiani and Fischer (28) who found that P_γ-$O_{3\beta}$ bond cleavage is facile and occurs early during the reaction. Proton transfer is found to be an important part of the rate-limiting transition state, as also found for other molecular motors, e.g. kinesin and F1-ATPase using kinetic isotope effect measurements.

Differences from previous studies involve identity of the general base, and whether the number of water molecules involved in mediating proton transfer is important to the hydrolysis barrier. Although glutamate has been suggested to be a viable base for ATP hydrolysis in several systems(16, 30, 52), Glu459 in myosin forms a salt-bridge with Arg238 in the pre-powerstroke state; this interaction is expected to substantially lower the pK_a of Glu459, making it a weaker base. Nevertheless, the present QM/MM simulations indicate that the W2-E459 pathway is likely; in fact, DFTB3/MM free energy simulations predict that the W2-E459 pathway has the lowest free energy barrier; however, single point energy calculations using higher-level B3LYP/MM DFT, performed along minimized structures from DFTB3/MM reaction paths (see **Supporting Information**), indicate that the W2-E459 and W2-S181 pathways are both equally likely. (Free energy simulations using B3LYP/MM DFT are currently computationally prohibitive for large systems). The free energy profile in Fig. 6.5 indicates that, although protonating Glu459 is indeed energetically costly, the formation of the hydroxide as a strong nucleophile ensures that the energetics of the intermediate species (3) to be in an acceptable range.

Rearrangements in hydrogen bonding networks (Fig. 6.4) also facilitate the changes in the protonation of

Glu459 in the W2-E459 pathway; such rearrangements can be missed in an MEP type of analysis, which may explain the observation of a higher energy barrier for the W2-E459 pathway in MEP studies;(28) further,since Glu459 is expected to have a low pK_a due to interaction with the positivey-charged Arg238, the deprotonation of Glu459 is unlikely to involve a high barrier.

We note that experiments are unlikely to shed light on the relative importance of the pathways involving Glu459 or Ser181 as the base; for example, both W2-E459 and W2-S181 pathways involve the transfer of multiple protons transition state, and it is unclear whether proton inventory experiments(98) would be able to distinguish between the two mechanisms.

It is noteworthy that the reaction barrier cannot be estimated on the basis of the number of water involved in mediating the proton transfer. For example, both W2-S181 and S236 pathways involve one water and one hydroxyl as the proton relay groups, yet the free energy barrier for W2-S181 is notably lower (compare Figs. 6.7 and 6.8). The S236-W3-S181 has a significantly longer proton relay pathway, yet its free energy barrier is close to that for the S236 pathway (compare Figs. 6.8 and 6.9). As discussed further below, the free energy barrier is more sensitive to the proximity of the excess charge in the proton transfer transition state to the stabilizing charge group, Glu459 in the current case.

6.4.2 Factors that regulate ATPase activity in myosin

The main goal of our work is to reveal factors that regulate the ATPase activity in myosin. As summarized in the Introduction, several factors have been discussed in previous studies, which include: (1) orientation of the lytic water along the "in-line" attack configuration(42); (2) polarization of the γ phosphate via nearby polar/charged interactions(28, 31, 99) or affecting β-γ phosphate orientation(45); (3) length of water wire that bridges the general base and the γ phosphate(52); (4) stability of hydrogen bonding network in the active site, which may rely on the interaction network surrounding the active site and therefore structural state of the distal region(33). These factors are discussed in turn below in light of the present results.

Calculation of the free energy cost for the water reorganization in the closed post-rigor model (Fig. 6.12b) using the present DFTB3/MM simulations is in accord with the previous classical MD results of 2-3 kcal/mol. In kinetic terms, this corresponds to a factor of 30-150 at room temperature according to transition state theory, which is quite significant; it is, nevertheless, a modest component of the free energy barrier difference between the fully active and inactive (or uncatalyzed) ATPase states. Water reorganization is insufficient to

explain differences in the hydrolysis rates corresponding to different structures or pathways.

On the role of γ phosphate polarization by nearby polar residues, our analysis of $\Delta\phi_{W1,W2}$ (Table 6.1) supports the importance of Gly457 backbone amide as proposed by several previous studies(28, 31, 99); the substantially higher (by ∼15 kcal/mol, compare Figs.6.10 and 6.8) free energy barrier for the S236 pathway in the post-rigor state than in the pre-powerstroke state further quantifies the importance of electrostatic polarization of the γ phosphate, although the contribution likely comes from multiple residues, including not only Gly457 but also water-mediated interaction with Glu459. Other polar interactions that involve SwI and the P-loop,(28, 31, 99) e.g. residues Ser181, Lys185 and Asn234, are also important to the polarization of ATP; however, they are similar the PPS and PR states.

Gerwert and co-workers observed that when Mg^{2+}·GTP binds to Ras or Ras·GAP, the β-γ phosphate moiety changes its orientation from staggered to eclipsed(45). This conformational change could potentially contribute to the polarization of the scissile bond.

In the PR and PPS states of myosin, the β-γ phosphate orientation remains similar with a $O_{1\gamma}$-P_γ-P_β-$O_{1\beta}$ dihedral angle of -25° degrees. Therefore, change of β-γ phosphate orientation is not a significant factor in the regulation of ATPase activity in myosin. Time-resolved infrared spectroscopic studies could provide additional information on how ATP is polarized in the different conformational states of the myosin motor domain, as was done for the analysis of GTP hydrolysis in Ras/Rab systems(45, 46).

We highlight, however, that although analysis of vibrational spectra and charge transfer effects in the ground state structures are highly informative, it remains difficult to evaluate the importance of these factors to catalysis without explicitly computing the transition states and comparing barrier heights(100).

It is interesting to understand why various pathways have substantially different barrier heights. In the study of ATP hydrolysis in actin, McCullagh et al.(52) proposed that the hydrolysis barrier in G-actin is substantially higher than in F-actin because proton transfer to the base in G-actin proceeds through a longer water wire. As discussed in the last subsection, the comparison of different pathways in the PPS state indicates that the length of the proton transfer wire *per se* is unlikely the critical factor. Instead, the free energy barrier is lower when the center of excess charge in the transition state is stabilized by charged residue(s) in the active site. This occurs in the three low-barrier pathways: W2-E459, W2 and W2-S181, in which the hydronium-like species in the transition state is directly stabilized by the carboxylate of Glu459. In the two high-barrier

pathways, S236 and S236-W3-S181, the hydronium-like species is further away from Glu459 and therefore receives much less stabilization. This trend can be captured by a simple gas-phase cluster model (see **Supporting Information**), which also shows that the effect is adequately reproduced when the carboxylate is described by a MM model, suggesting that electrostatic polarization is most important and charge transfer is not crucial. Finally, as mentioned above, the importance of Glu459 electrostatic polarization is also reflected by the much higher barrier for the S236 pathway in PR than in PPS.

A comparison of hydrolysis free energy profiles for the PPS and closed PR model provides insights on the influence of residues beyond the immediate coordination shell of the γ-phosphate. Despite the general similarity in the active site protein structures (see **Supporting Information**), hydrolysis in the closed PR model is substantially less (∼9 kcal/mol) endothermic and has a higher barrier than in PPS state (compare Figs. 6.11a and 6.5a). A modest part (2-3 kcal/mol) of this difference comes from water reorganization (Fig. 6.12b), while the rest is mostly due to residues in close proximity ($\leq$ 5 Å) of the γ phosphate; compared to in PPS, these nearby residues (e.g., Ser456 and Gly457) in the closed post-rigor model provide less electrostatic driving force and undergo more significant dipolar reorganization for the proton transfers during hydrolysis. Evidently, efficient hydrolysis explicitly requires precise positioning of only the nearby residues, which control both electrostatic driving force (stabilization) and reorganization contributions.

This observation does not rule out the importance of residues further away. As discussed in Ref.26, 33, maintaining a stable and precise positioning of active site residues relies on structural organizations of distal regions; in the particular case of myosin, this mostly involves the N-terminus of the relay helix and relay loops. Without the structural support of these motifs, the active site fluctuates between the open and closed configurations with a rather flat free energy profile;(32)mutations to evaluate the contribution of these distal residues on the ATPase activity have been proposed(33).

6.4.3 Broader Implications: importance of positioning, electrostatics and plasticity in catalysis

It is instructive to discuss the present results in the broader context of enzyme catalysis and enzyme design in general. In their study of ketosteroid isomerase, Fried et al.(101) observed a strong correlation between catalytic efficiency and the magnitude of electric field in the active site, as measured through the Stark effect, highlighting the electrostatic contribution of catalysis. Herschlag and co-workers(102, 103), on the other hand, emphasized that the electrostatic contribution will not be effective unless the key motifs are properly positioned

in the active site. In the case of the design enzyme for Kemp elimination,(104) Mayo and co-workers showed that the low level of catalytic power could be enhanced 6×10^8 fold with directed evolution; comparison of the crystal structures for the low- and high-activity enzymes suggested that the low activity was due to suboptimal orientation of the substrate, which weakened the stabilization of the transition state for proton abstraction.

In the case of ATP hydrolysis in myosin, we have shown that properly positioning of water molecules and nearby polar/charged groups (Gly457, Glu459) are clearly important. Disruption of the water network contributes to an amount of 2-3 kcal/mol in free energy barrier, which is modest in energetic terms but appreciable in kinetic terms. Displacement of SwII from the active site, as in the the post-rigor state, weakens stabilizing electrostatic interactions from Gly457 and Glu459, and leads to a much higher hydrolysis barrier. Further, simulations using the closed post-rigor model structure show a substantially higher hydrolysis barrier, compared to simulations that use the PPS structure. This was somewhat surprising, because in the model structure the active site was modified to be in a very similar conformation to that of the PPS state. We therefore conclude that residues beyond the active site also have a significant effect on the catalysis rate. Analysis of electrostatic potentials in the active site further suggests that proper positioning of residues in the active site is important for both enhancement of electrostatic driving force and minimization of dipolar reorganization (or maximization of pre-organization).Therefore, it is fruitful to consider both electrostatic stabilization and pre-organization effects in catalyst design(105–107).

Another interesting feature observed in myosin is the existence of multiple hydrolysis pathways that have somewhat similar free energy barriers in the pre-powerstroke state; the barriers for the five pathways span a range of less than 10 kcal/mol, with the three low-barrier pathways featuring even smaller differences considering the range of accuracy associated with the semiempiracle QM/MM approach adopted here. Together with the discussion of catalytic promiscuity in enzymes,(60, 108–111) the existence of multiple, near-degenerate pathways makes it clear that enzyme catalysis can be much more plastic in nature than what the traditional "one-enzyme-one-mechanism" picture implies. It will be interesting to systematically explore whether multiple competitive pathways coexist in other enzymes and if so, the implication to biological function and evolution of enzymes(112). For instance, the existence of multiple electron transfer pathways and inference among them has been highlighted(113, 114), and multiple polar groups may mediate proton transfers(115–118). Does this mean the existence of multiple pathways is mostly limited to enzymes that catalyze electron and proton

transfer reactions? Is the existence of multiple pathways an accidental feature or actually taken advantage of by evolution, since one may argue that the existence of multiple pathways improves robustness of catalysis and thus perhaps the evolvability of the enzyme. With progress in automated reaction pathway exploration methodologies(119, 120) and integration with enzyme evolution data(121), these fascinating questions can be explored at a large scale.

6.5 Concluding Remarks

One of the major hallmarks of natural enzymes is that their activities are tightly regulated. This feature is particularly important to enzymes involved in energy and signal transductions, for which a tight regulation of the chemical event (e.g., NTP hydrolysis) is key to the efficiency of the energy transduction(20, 23, 32, 94, 122) and timing of signal transduction(123). Therefore, understanding how catalytic activities in enzymes are modulated by structural and dynamical properties of the active site and its surrounding is essential to the mechanism of these fascinating systems. On the practical side, such understanding is valuable to the design of effective catalysts.

By combining QM/MM simulations with metadynamics and the string method, we have analyzed in detail the mechanism of ATP hydrolysis activity in the myosin domain. Focusing mainly on the hydrolysis-competent pre-powerstroke state, we identified multiple reaction pathways of ATP hydrolysis and computed the associated free energy profiles. The scissile bond in the γ-phosphate of ATP breaks early in the reaction; proton transfer from the lytic water to the γ phosphate through active site residues is an important part of the kinetic bottleneck of the hydrolysis. Different hydrolysis pathways with similar free energy barriers are found, indicating plasticity in the hydrolysis mechanism.

To understand the factors that determine the catalytic efficiency, we also analyzed hydrolysis pathways in PR (in which hydrolysis is disfavored), and in a hybrid model that has active site features of PPS state, with the overall conformation of the PR state. The results suggest that the ATPase activity depends directly on the positioning of nearby water molecules to conduct efficient in-line attack, but also, indirectly, on residues beyond the active site. From the functional perspective of avoiding futile hydrolysis, the apparent sensitivity of ATPase activity to the positioning of active site motifs is clearly essential, provided that structural transitions in the active site and distal regions are highly co-operative in nature(94). In the more general context, the results suggest that optimization of residues beyond the active site for electrostatic stabilization and pre-organization

is likely important to rational enzyme design.

6.6 Supporting Information

6.6.1 Collective Variables in the Metadynamics Simulations

Collective variables used in the metadynamics simulations include the anti-symmetric stretch, $P_\gamma O_{3\beta}$ – $P_\gamma O^{W1}$, and the ζ coordinate(82) that describes the proton transfers. Due to the collective nature of ζ, it is capable of describing proton transfers without explicitly specifying the number and identity of the groups that mediate proton transfer. Therefore, it is well suited for discovering proton transfer pathways in a nearly automated fashion. In certain cases, to turn off undesirable proton transfer pathways, specific sidechains are treated by a MM rather than a QM model. In particular, for the W2 and the S181-W2 mechanisms, Glu459 and Arg238 are treated as MM to avoid Glu459 from accepting a proton. The identity of proton donor, acceptor and potential proton shuttling group used in the definition of ζ for different pathways are listed in Table 6.2.

Table 6.2: Identity of proton donor, acceptor and potential proton shuttling group used in the definition of ζ in the metadynamics simulations of different reaction pathways.

Structural model	Pathway	Proton Donor	Proton Acceptor	Proton Shuttling groups
pre-powerstroke	W2-E459	P_β	Glu459 OE1	W1, W2, E459, ATP
pre-powerstroke	W2	P_β	Glu459 OE1	W1, W2, ATP
pre-powerstroke	W2-S181	P_β	Glu459 OE1	W1, W2, S181, ATP
pre-powerstroke	S236	P_β	Arg238 CZ	W1, S236, ATP
pre-powerstroke	S236-W3-S181	P_β	Glu459 OE1	W1, W3, S236, S181, ATP
post-rigor	S236	P_β	Arg238 CZ	W1, S236, ATP
"closed" post-rigor	W2-E459	P_β	Glu459 OE1	W1, W2, E459, ATP

6.6.2 The Projection Problem in Metadynamics Simulations for Myosin

The advantage of using the ζ coordinate, which is highly collective in nature, is that metadynamics simulations can be used to automatically discover reaction pathways with little prior input. On the other hand, the collective nature of ζ also introduces uncertainty in the computed 2D free energy surface, since rather different reactive configurations (which likely have very different energies) can be mapped onto the same location on the low-dimensional manifold. In this section, we use the S236 pathway in the pre-powerstroke state as an example to illustrate such projection problem in the metadynamics simulations; this type of problem is ob-

served for all pathways studied here, while the S236 pathway is chosen here due to its simplicity in the nature of transition state.

Figs. 6.14c-d show two examples, which correspond to a structure close to the final hydrolysis product and transition state, respectively. They share similar values for the set of collective coordinates: the $P_\gamma O_{3\beta}$ – $P_\gamma O^{W1}$ is 0.8 $\sim$ 1.0 Å and ζ is $\sim$ 0.0. Energetically, the configure close to the product is far more stable than the transition state, so Gaussian potentials keep being deposited to this region during metadynamics simulations, skewing the free energy surface. Indeed, the free energy barrier found by metadynamics is $\sim$ 30 kcal/mol, which is 5 kcal/mol higher than the value from the string calculations.

Moreover, as Gaussian potentials accumulate in the region close to the product region, the system is driven away from the minimum free energy pathway during metadynamics simulations. In Fig. 6.14a, the free energy surface is superimposed with a reactive trajectory observed in metadynamics simulation; in Fig. 6.14b, the same free energy surface is superimposed with the converged string projected onto the two-dimensional space. Clearly, the two pathways are distinct: while P_γ-$O_{3\beta}$ breaks prior to proton transfer in the converged string, the metadynamics simulation leads to a reactive trajectory that follows the opposite sequence of event and transverses the high energy region of the free energy surface.

These observations highlight that free energy calculations using metadynamics should be treated with great care when projecting high-dimensional processes onto a low-dimensional space.

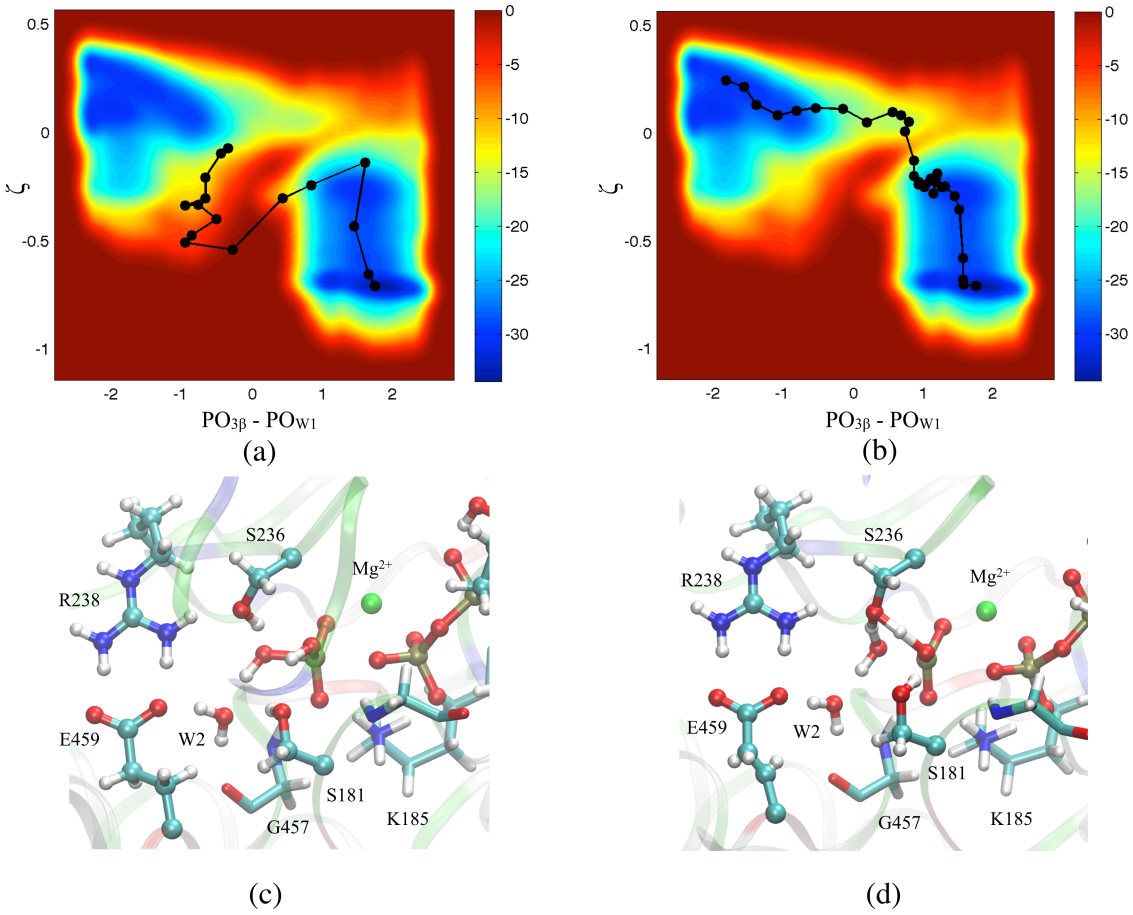


Figure 6.14: The 2D free energy surface shown in (a)(b) is obtained from metadynamics for S236 pathway in the pre-powerstroke state. (a) The reactive trajectory is projected on the surface as the solid black line. The structures in the reactive trajectory are used as the initial guess in the following string calculation. (b) The converged string is illustrated on the surface as the solid black line. Each circle corresponds to an image along the string. The color scale denotes free energy in the unit of kcal/mol. (c) The representative pre-product structure in which hydroxyl group in the inorganic phosphate is about to rotate towards to ADP. (d) The desired transition state discovered in the MFEP in the string method.

6.6.3 Collective Variables in the String Calculations

Collective variables used in the string calculations are summarized in Table 6.3.

Table 6.3: Definition of the collective variables in the string method

System	Mechanism	Collective variable
pre-powerstroke	W2-E459	d(W1 O-ATP P_γ), d(ATP P_γ-ATP $O_{3\beta}$), d(W2 H_1-ATP $O_{2\gamma}$), d(W2 H_1-ATP $O_{3\beta}$), d(W1 O-W1 H_1), d(W1 O-W1 H_2), d(W2 O-W2 H_1), d(W2 O-W2 H_2), d(W1 H_1-W2 O), d(W1 O-W2 O), d(W2 H_2-E459 OE1), d(ATP P_β-ATP $O_{3\beta}$), d(Mg-ATP $O_{1\gamma}$), d(Mg-ATP $O_{1\beta}$), d(E459 OE1-R238 HH11), d(E459 OE2-R238 HH21)
pre-powerstroke	W2[a]	d(W1 O-ATP P_γ), d(ATP P_γ-ATP $O_{3\beta}$), d(W1 H_1-ATP $O_{2\gamma}$), d(W1 H_1-ATP $O_{3\beta}$), d(W1 O-W1 H_1), d(W1 O-W1 H_2), d(W2 O-W2 H_1), d(W2 O-W2 H_2), d(W1 H_1-W2 O), d(W1 O-W2 O), d(W2 H_2-E459 OE1), d(ATP P_β-ATP $O_{3\beta}$), d(Mg-ATP $O_{1\gamma}$), d(Mg-ATP $O_{1\beta}$), d(E459 OE1-R238 HH11), d(E459 OE2-R238 HH21)
pre-powerstroke	W2-S181	d(W1 O-ATP P_γ), d(ATP P_γ-ATP $O_{3\beta}$), d(S181 HG1-ATP $O_{2\gamma}$), d(S181 HG1-ATP $O_{3\beta}$), d(W1 O-W1 H_1), d(W1 O-W1 H_2), d(W2 O-W2 H_1), d(W2 O-W2 H_2), d(W1 H_1-W2 O), d(W1 O-W2 O), d(W2 H_2-E459 OE1), d(W2 H_1-G457 O), d(W2 H_1-S181 OG), d(S181 OG-S181 HG1), d(ATP P_β-ATP $O_{3\beta}$), d(Mg-ATP $O_{1\gamma}$), d(Mg-ATP $O_{1\beta}$), d(E459 OE1-R238 HH11), d(E459 OE2-R238 HH21)
pre-powerstroke	S236	d(W1 O-ATP P_γ), d(ATP P_γ-ATP $O_{3\beta}$), d(S236 HG1-ATP $O_{2\gamma}$), d(S236 HG1-ATP $O_{3\beta}$), d(W1 O-W1 H_1), d(W1 O-W1 H_2), d(W1 O-W2 O), d(W1 H_2-W2 O), d(W1 H_1-S236 OG), d(S236 OG-S236 HG1), d(ATP P_β-ATP $O_{3\beta}$), d(Mg-ATP $O_{1\gamma}$), d(Mg-ATP $O_{1\beta}$)
pre-powerstroke	S236-W3-S181	d(W1 O-ATP P_γ), d(ATP P_γ-ATP $O_{3\beta}$), d(S181 HG1-ATP $O_{2\gamma}$), d(S181 HG1-ATP $O_{3\beta}$), d(W1 O-W1 H_1), d(W1 O-W1 H_2), d(W1 O-W2 O), d(W1 H_2-W2 O), d(W1 H_1-S236 OG), d(S236 OG-S236 HG1), d(S236 HG1-W3 O), d(W3 O-W3 H_1), d(W3 O-W3 H_2), d(W3 H_2-S181 OG), d(S181 OG-S181 HG1), d(ATP P_β-ATP $O_{3\beta}$), d(Mg-ATP $O_{1\gamma}$), d(Mg-ATP $O_{1\beta}$)
post-rigor	S236	d(W1 O-ATP P_γ), d(ATP P_γ-ATP $O_{3\beta}$), d(S236 HG1-ATP $O_{2\gamma}$), d(S236 HG1-ATP $O_{3\beta}$), d(W1 O-W1 H_1), d(W1 O-W1 H_2), d(W1 H_1-S236 OG), d(S236 OG-S236 HG1), d(ATP P_β-ATP $O_{3\beta}$), d(Mg-ATP $O_{1\gamma}$), d(Mg-ATP $O_{1\beta}$)
"closed" post-rigor	W2-E459	d(W1 O-ATP P_γ), d(ATP P_γ-ATP $O_{3\beta}$), d(W1 H_1-ATP $O_{2\gamma}$), d(W1 H_1-ATP $O_{3\beta}$), d(W1 O-W1 H_1), d(W1 O-W1 H_2), d(W2 O-W2 H_1), d(W2 O-W2 H_2), d(W1 H_1-W2 O), d(W1 O-W2 O), d(W2 H_1-E459 OE1), d(W2 O-E459 CD), d(ATP P_β-ATP $O_{3\beta}$), d(Mg-ATP $O_{1\gamma}$), d(Mg-ATP $O_{1\beta}$), d(E459 OE1-R238 HH11), d(E459 OE2-R238 HH21)
"closed" post-rigor	water reorganization	d(W1 O-ATP P_γ), d(W2 O-ATP P_γ), d(W1 O-W2 O), d(W2 H_1-E459 OE1), d(W2 O-E459 CD), d(W2 H_2-G457 O)

[a] a NOE constraint is applied to the distance between W2 water and Glu459 to prevent a proton from being too close to the Glu459.

6.6.4 Comparison between Metadynamics and String Calculations for the W2-E459 Pathway in the Pre-powerstroke State

The minimum free energy path (MFEP) computed by string method in this work does not capture contributions due to the thermal fluctuations of the CVs orthogonal to the string. Therefore, too few number of CVs in MFEP calculations do not properly describe the reactive process, while too many CVs might potentially freeze out the relevant degrees of freedom. In this section, we use the W2-E459 pathway to demonstrate that the string method arrives at the same pathway and free energy barrier with metadynamics simulations if the projection problem in metadynamics discussed in Sect. 6.6.2 is properly handled. To remove the contamination due to the product-like structures shown in Fig. 6.14c, a strong harmonic restraint is applied to the ζ coordinate to enforce it to be positive; this helps prevents metadynamics from sampling the pre-product configuration (with a positive ζ, the excess charge does not reach γ-phosphate). As seen in Fig. 6.15, the converged string projected onto the 2D space closely follows the free energy valley computed by metadynamics simulations. The two approaches also give similar free energy barrier of $\sim$17 kcal/mol. Therefore, entropic effects associated with the fluctuations of CVs orthogonal to the string do not appear to play a major role in the current problem.

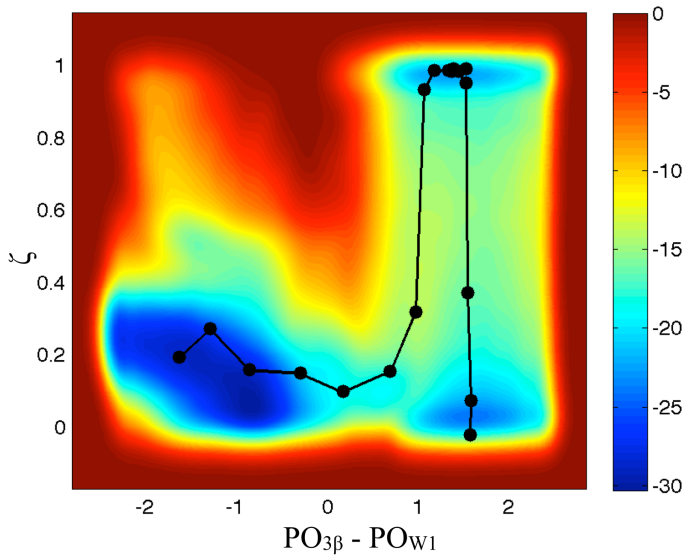


Figure 6.15: The 2D free energy surface obtained from metadynamics for W2-E459 pathway in the pre-powerstroke state. The converged string is illustrated on the surface as the solid black line. Results obtained from metadynamics and string match each other well regarding to the mechanism and free energy barrier. Each circle corresponds to an image along the string. The color scale denotes free energy in the unit of kcal/mol.

6.7 Electrostatic Polarization due to Glu459 is Captured by a Simple Model

To illustrate the impact of electrostatic polarization due to Glu459 on the proton transfers, we construct a simple gas phase model (Fig. 6.16). The model consists of one acetate group and two water molecules (referred to as the two-water model) or three water molecules (referred to as the three-water model), and we study the proton transfer from the first water (W1) to the second water (W2). The positions of the water molecules and acetate are adjusted to allow hydrogen bonding interactions of typical distances; a fixed MM point charge is placed next to W1 to stabilize the hydroxide in the product state. The proton transfer is studied by conducting one-dimensional adiabatic mapping at the DFTB3 level; the reaction coordinate is the antisymmetric stretch involving the breaking O^{W1}-H bond and forming O^{W2}-H bond. The transition state is approximated by the highest energy structure along the adiabatic mapping profile. Single point energy calculations at the B3LYP/aug-cc-pVTZ level are also carried out for comparison.

At the DFTB3 level, the barrier height is 4.8 kcal/mol for the two-water model and substantially higher

(14.1 kcal/mol) for the three-water model; the exothermicity of the reaction is similar, -1.0 kcal/mol and +2.3 kcal/mol for the two-water and three-water model, respectively. The values are similar at the B3LYP level; for example, the barrier heights are 6.5 kcal/mol and 16.0 kcal/mol, respectively. Therefore, it is evident that direct stabilization of the hydronium-like species by acetate in the two-water model reduces the proton transfer barrier significantly; i.e., the stabilization effect of the acetate is quenched significantly when a water molecule separates the acetate and the hydronium-like species. Moreover, if the acetate is treated classically using a MM model, the barrier height in the two-water model is increased by less than 1 kcal/mol, demonstrating that stabilization of the transition state is not due to charge transfer.

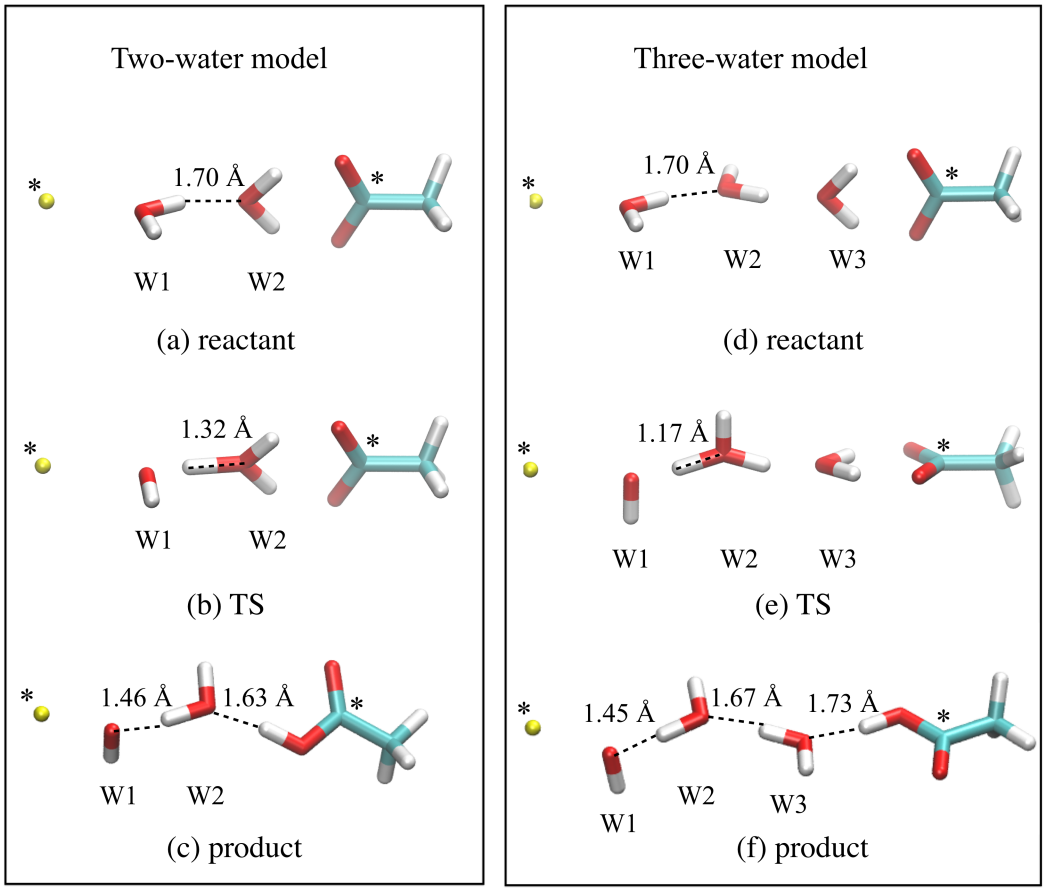


Figure 6.16: Proton transfer in a model system consisting of acetate and water molecules to illustrate the impact of polarization effects. The acetate and water molecules are colored by atom type, and the MM point charge is colored in yellow. Atoms fixed are denoted by asterisk. (a) - (c) Snapshots for the reactant, transition state and product in the two-water model, respectively. (d) - (f) Snapshots for the reactant, transition state and product in the three-water model, respectively.

6.7.1 Calibration of DFTB3 with B3LYP calculations

Previous benchmark calculations(58) and recent applications(60, 61) indicate that the DFTB3 approach provides a semi-quantitative description for phosphoryl transfer reactions, including key structural, energetic and vibrational features of phosphoryl transition states. Nevertheless, the DFTB3 model is not quantitatively accurate for complex reactions that involve highly charged species like ATP hydrolysis. To ensure that key

trends observed in the free energy profiles are meaningful, we conduct B3LYP and B3LYP/MM calculations for model reactions and the real enzyme system, respectively.

6.7.2 Exothermicity of Hydrolysis

The free energy simulations indicate that the hydrolysis reaction in the pre-powerstroke state is exothermic by about 10 kcal/mol, while experimental measurements indicate that the equilibrium constant for the hydrolysis step in this functional state is close to 1. Therefore, it is likely that the DFTB3 approach overestimates the exothermicity of the hydrolysis reaction by a substantial amount.

To verify this hypothesis, we study a model hydrolysis reaction in the gas phase; the corresponding reactant and product molecules are illustrated in Fig. 6.17. The reaction energy is evaluated by full geometry optimizations with DFTB3/3OB-OPhyd and the result is compared to B3LYP/aug-cc-pVTZ calculations (optimized at the respective levels). As shown in Table 6.4, the DFTB3/3OB-OPhyd model indeed overestimates the hydrolysis exothermicity by about 9 kcal/mol.

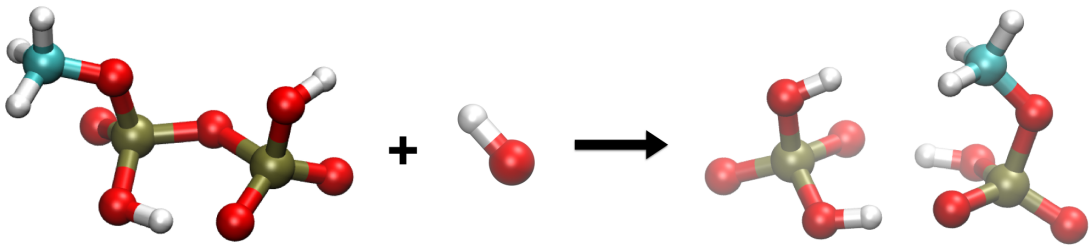


Figure 6.17: Model ADP hydrolysis reaction in the gas phase

Table 6.4: Energy difference between the product and the reactant in the model-ADP system obtained at DFTB3 and B3LYP/aug-cc-pvtz level in the unit of kcal/mol

Component	DFTB3/OPhyd	B3LYP/aug-cc-pvtz
OH^-	-2305.83	-47587.98
Model ADP^-	-19361.40	-785033.86
$(CH_3O)HPO_3^- + H_2PO_4^-$	-21681.65	-832627.19
Energy difference	-14.42	-5.35

6.7.3 Transition State Energetics in the Enzyme

As an attempt to calibrate the computed barrier heights in different pathways, we first conduct minimum energy path calculations at the DFTB3/MM level, and then compute single point energy calculations at the B3LYP/MM level for the barrier heights. Although these are not free energy barrier comparisons (which would be difficult to do at the B3LYP/MM level), the comparison provides qualitative indications regarding whether trends in the DFTB3/MM free energy barriers are meaningful.

As discussed extensively in previous studies(53, 124–126), MEP calculations in enzymes are sensitive to the protein environment used to initiate such calculations. Therefore, independent starting protein structures are selected for the MEP calculations; three snapshots from the reactant, TS and product regions of the W2-E459 pathway are used. The same starting protein structures are used to initiate MEP calculations for the different hydrolysis mechanisms, so that a fair comparison for the barrier heights can be made; otherwise, fluctuations in the MEP barrier height are likely too large to make the comparison meaningful. The MEP calculations are conducted using the zero-temperature string method implemented in CHARMM, and the transition state is further refined using conjugate peak refinement(127) to obtain a precise estimate of the barrier height. The B3LYP/MM single point energy calculations are done with the 6-31+G(d,p) basis set, which was shown in previous studies(128) to be of adequate accuracy for the current purpose.

As shown in Table 6.5, potential energy barrier heights from the MEP calculations at the DFTB3/MM level largely follow the same trend as the free energy barriers computed using the string method. With the same starting protein environment, the mechanisms that involve W2 have lower barriers than those involve S236. For the same mechanism, the potential barrier is lower when the protein environment is selected from MD simulations for the transition state; this is because the protein environment has reorganized to preferentially stabilize charge distributions characteristic of transition state structures(53).

The barrier heights computed at the B3LYP/MM level are generally higher and the difference from DFTB3/MM ranges from $\sim$2 kcal/mol to $\sim$10 kcal/mol; generally, the correction appears to be particularly large for the W2 mechanism. Therefore, it is evident that the DFTB3/MM approach is not yet quantitatively accurate for studying ATP hydrolysis in a complex enzyme like myosin. The encouraging aspect, however, is that the key trends are essentially maintained with the B3LYP/MM barrier heights: the W2-E459 and W2-S181 pathways feature barriers close to 20 kcal/mol, while the S236 and S236-W3-S181 pathways feature barriers close to be

30 kcal/mol; considering the large difference between DFTB3/MM and B3LYP/MM results for the W2 mechanism, it is difficult to conclude whether it should be considered as a low-barrier mechanism. Calculations at the B3LYP/MM level with more extensive sampling are required to better rank the W2 mechanism against others.

Table 6.5: Potential Energy Barriers from DFTB3/MM and B3LYP/6-31+G(d,p)/MM calculations for different reaction mechanisms in the pre-powerstroke state

Mechanism	Protein Environment		
	Reactant	Transition State	Product
W2-E459	15.4 (17.6), 17.5 (23.0)	11.0 (15.5), 10.0 (20.2)	14.8 (22.5)
W2-S181	19.1 (23.3)	15.7 (24.2)	17.5 (22.9)
W2	21.8 (32.2)	14.5 (26.9)	13.3 (24.7)
S236	25.0 (28.4)	20.9 (27.7)	23.8 (28.9)
S236-W3-S181	23.3 (30.1)	19.1 (28.6)	23.4 (29.3)

[a] The numbers without parentheses are at DFTB3 level, while the numbers with parentheses are single-point correction from B3LYP/6-31+G(d,p) on top of DFTB3 geometries.

6.7.4 Comparison of active site structure and fluctuations between the pre-powerstroke state and the closed post-rigor model

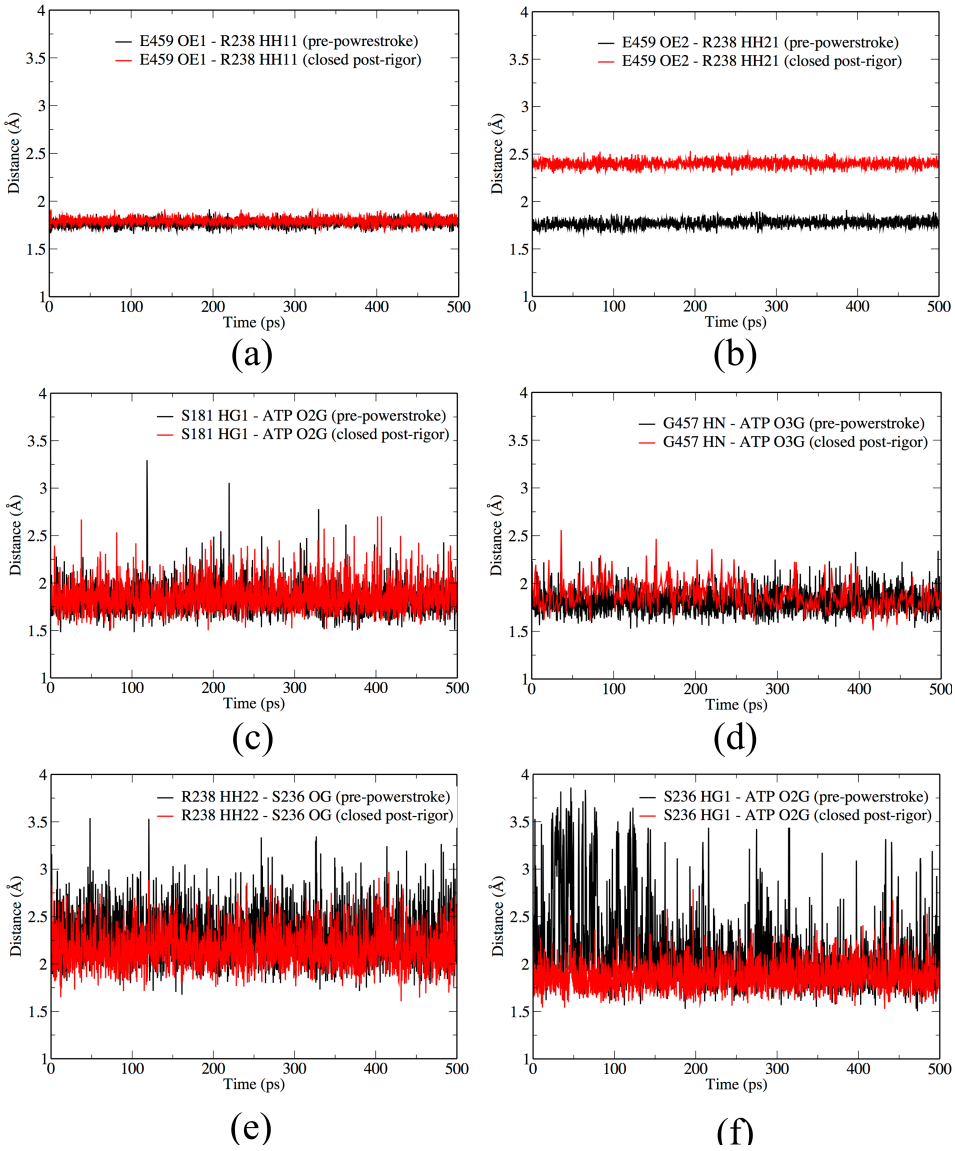


Figure 6.18: Distances between key residues in the reactant state

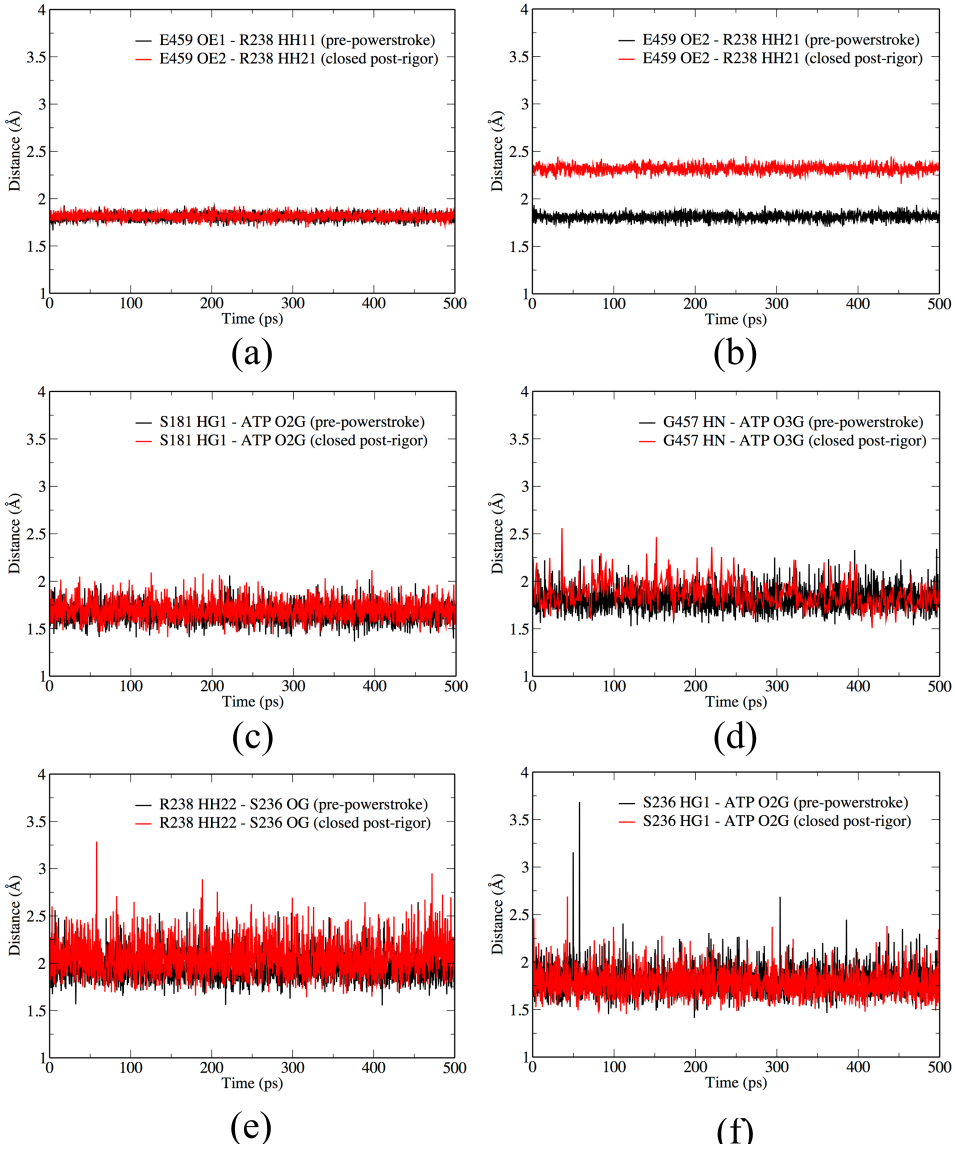


Figure 6.19: Distances between key residues in the transition state

6.7.5 RMSD between reaction pathways in the pre-powerstroke state and the closed post-rigor model

To better compare the reaction pathways in the two structural states, we calculate the RMSD between images along the optimized strings in the pre-powerstroke state and the closed post-rigor model. Both strings describe the W2-E459 mechanism, although there are structural (water) reorganizations in the initial stage for the closed post-rigor model (see discussion in main text), as reflect by the slightly larger RMSD values

between these images and the intermediate images in the pre-powerstroke state.

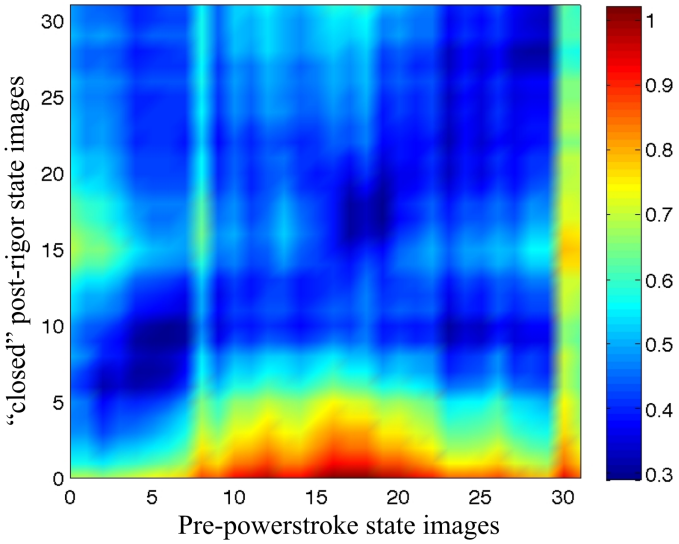


Figure 6.20: 2D RMSD between the pre-powerstroke state images and the "closed" post-rigor state images

6.7.6 W2 Pathway in the Pre-powerstroke State

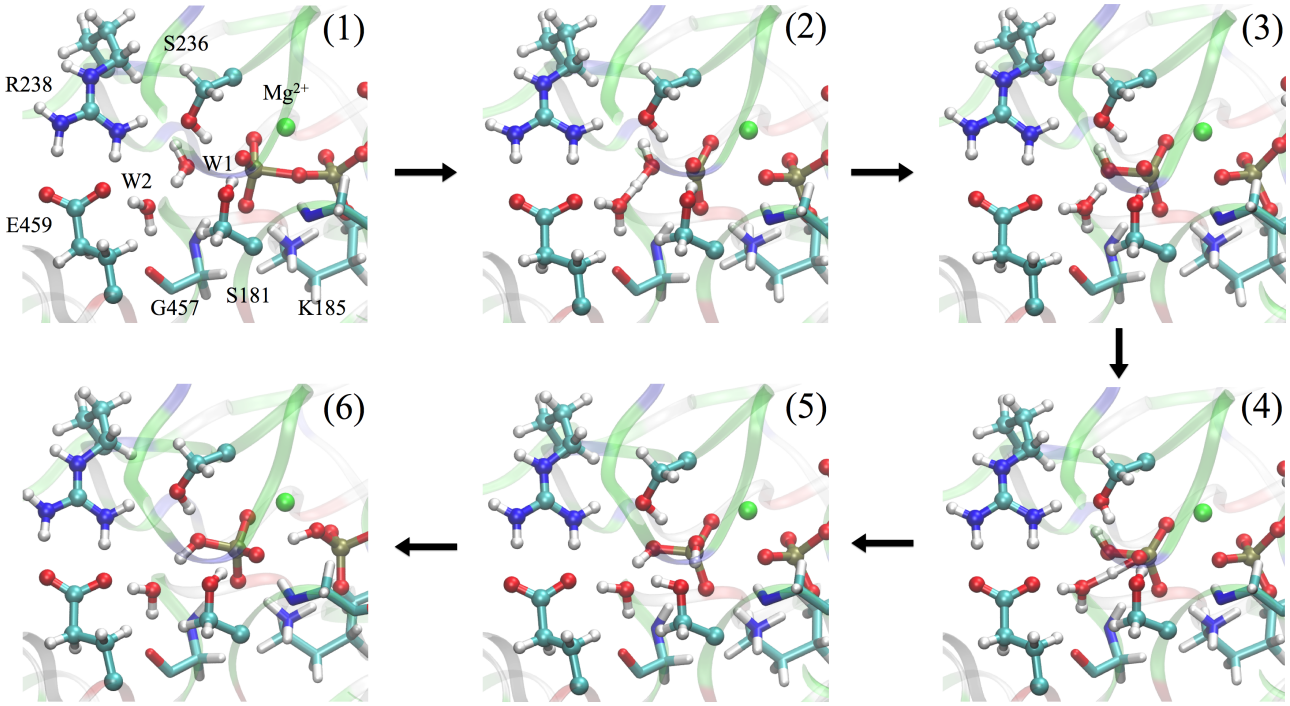


Figure 6.21: Representative structures along the MFEP obtained from the W2 mechanism in the pre-powerstroke state.

6.7.7 W2-S181 Pathway in the Pre-powerstroke State

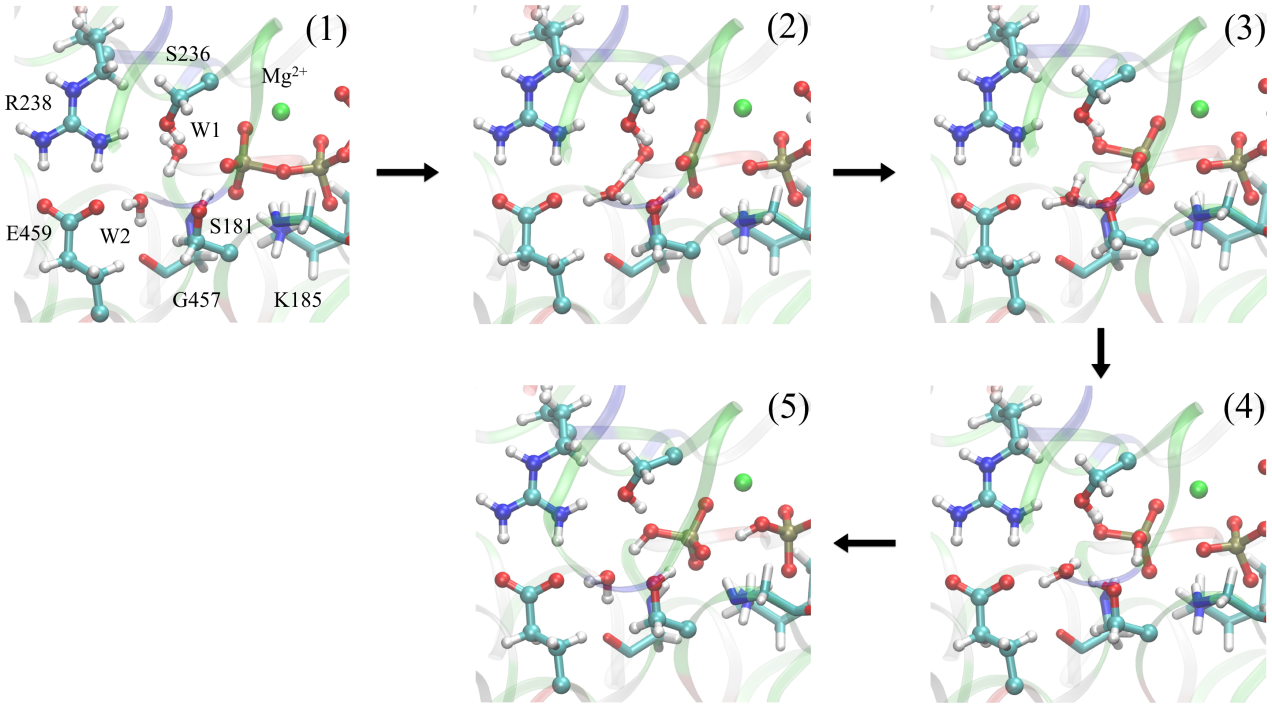


Figure 6.22: Representative structures along the MFEP obtained from the W2-S181 mechanism in the pre-powerstroke state.

6.7.8 S236 pathway in the Pre-powerstroke State

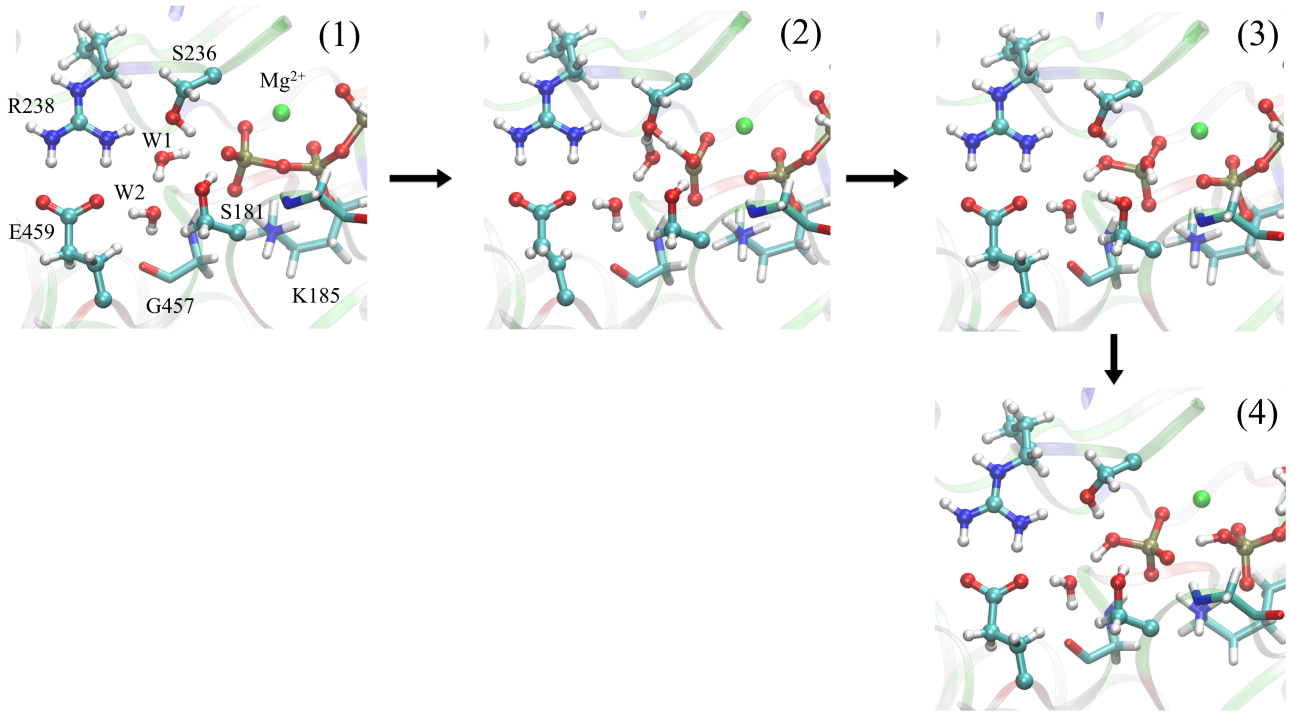


Figure 6.23: Representative structures along the MFEP obtained from the S236 mechanism in the pre-powerstroke state.

6.7.9 S236-W3-S181 pathway in the Pre-powerstroke State

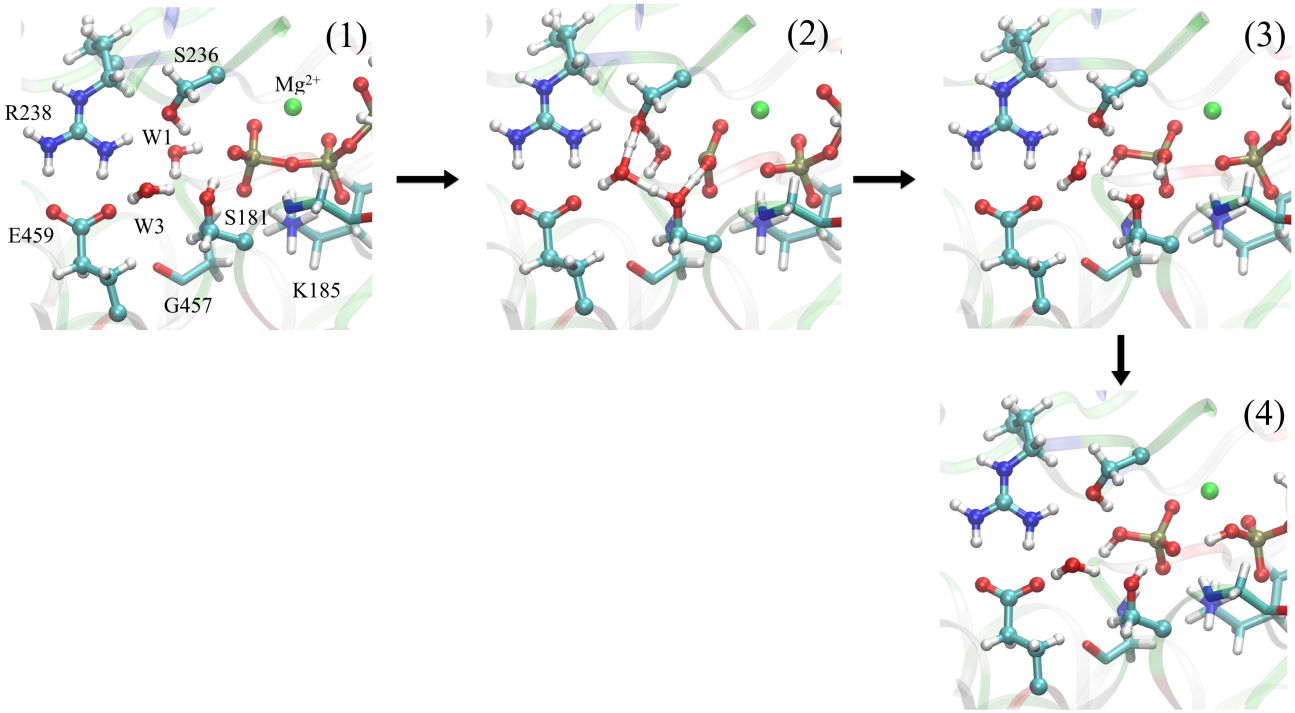


Figure 6.24: Representative structures along the MFEP obtained from the S236-W3-S181 mechanism in the pre-powerstroke state.

6.7.10 S236 pathway in the Post-rigor State

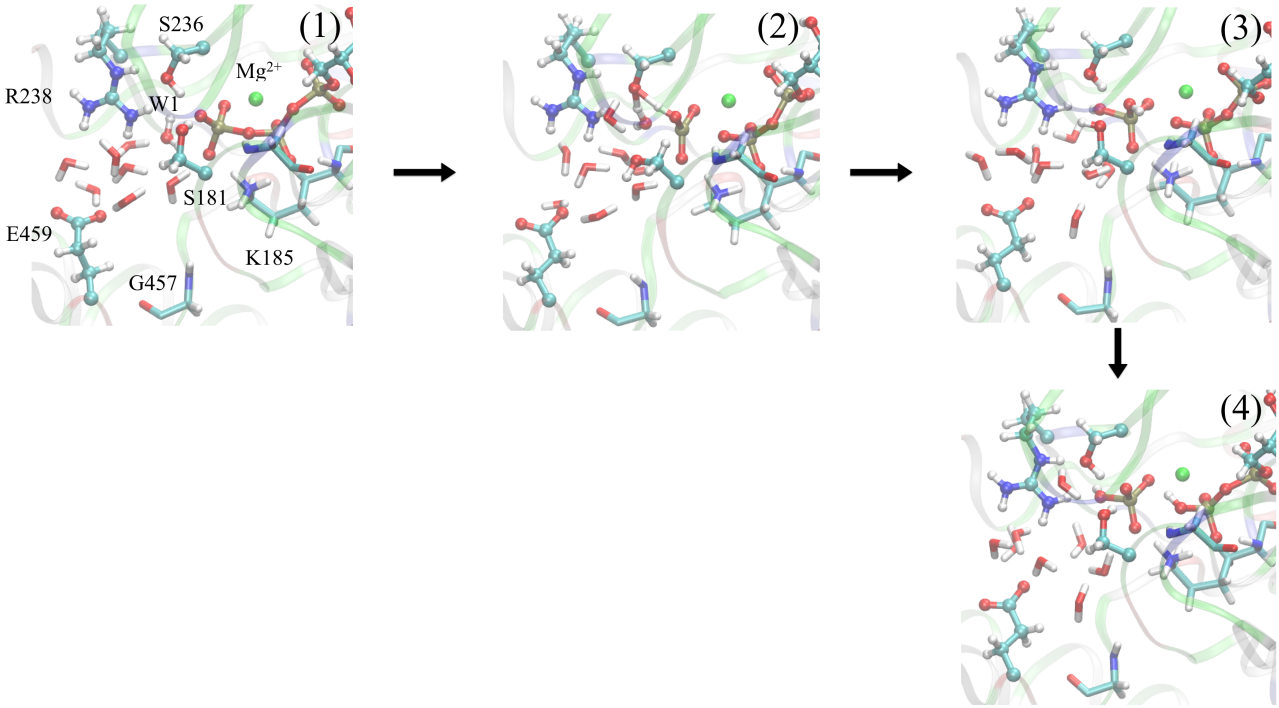


Figure 6.25: Representative structures along the MFEP obtained from the S236 mechanism in the post-rigor state.

6.7.11 W2-E459 pathway in the ”closed” Post-rigor State

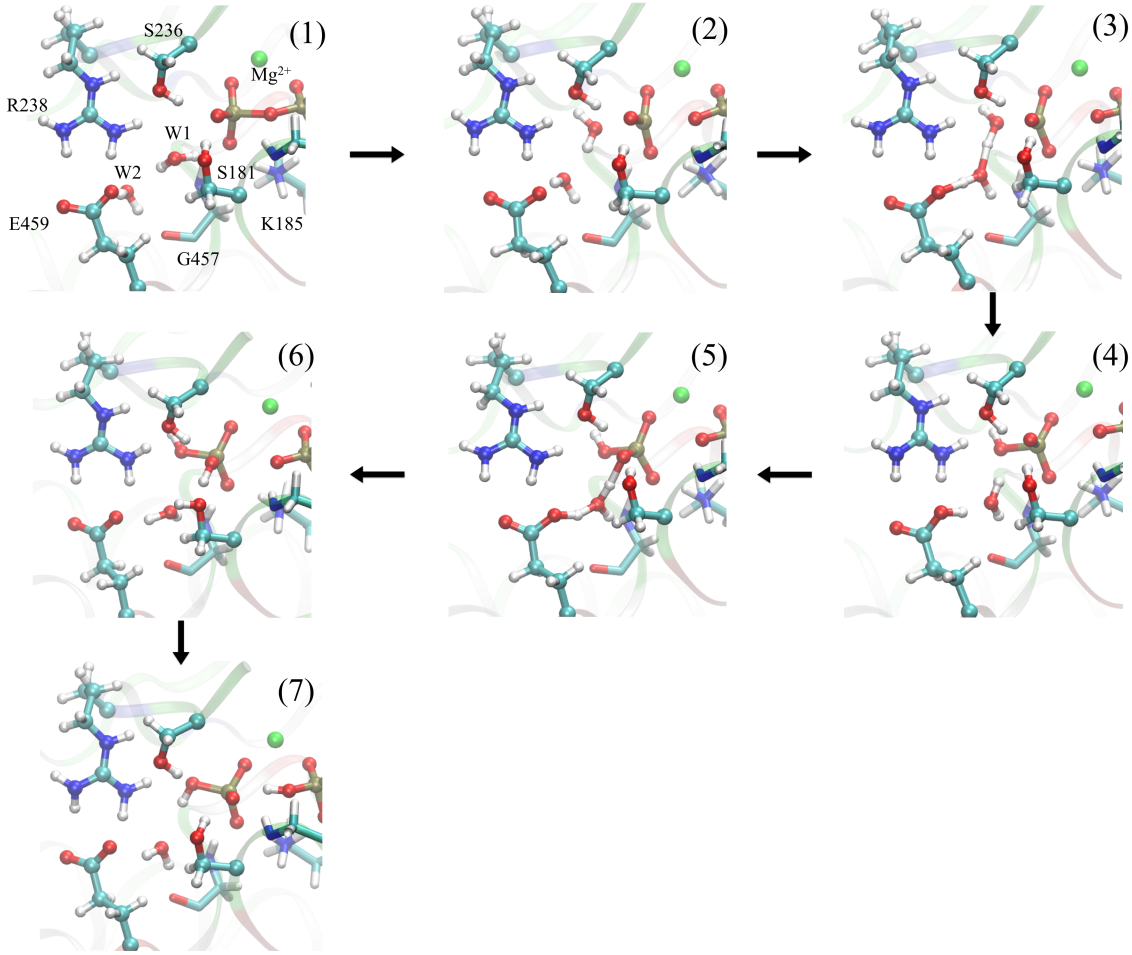


Figure 6.26: Representative structures along the MFEP obtained from the W2-E459 mechanism in the “closed” post-rigor state.

References

[1] B. H. SOLIS and S. HAMMES-SCHIFFER, *Inorg. Chem.* **53**, 6427 (2014).

[2] T. A. BETLEY, Q. WU, T. VAN VOORHIS, and D. G. NOCERA, *Inorg. Chem.* **47**, 1849 (2008).

[3] J. E. MONDLOCH, M. J. KATZ, W. C. ISLEY, III, P. GHOSH, P. LIAO, W. BURY, G. WAGNER, M. G. HALL, J. B. DECOSTE, G. W. PETERSON, R. Q. SNURR, C. J. CRAMER, J. T. HUPP, and O. K. FARHA, *Nat. Mater.* **14**, 512 (2015).

[4] G. KISS, N. CELEBI-OELCUEM, R. MORETTI, D. BAKER, and K. N. HOUK, *Angew. Chem., Int. Ed.* **52**, 5700 (2013).

[5] D. HILVERT, Design of Protein Catalysts, in *ANNUAL REVIEW OF BIOCHEMISTRY, VOL 82*, edited by KORNBERG, RD, volume 82 of *Annu. Rev. Biochem.*, pp. 447–470, 2013.

[6] J. K. LASSILA, D. BAKER, and D. HERSCHLAG, *Proc. Natl. Acad. Sci. USA* **107**, 4937 (2010).

[7] L. WANG, N. M. GOODEY, S. J. BENKOVIC, and A. KOHEN, *Proc. Natl. Acad. Sci. USA* **103**, 15753 (2006).

[8] C. SCHULENBURG, Y. STARK, M. KUENZLE, and D. HILVERT, *J. Biol. Chem.* **290**, 9310 (2015).

[9] N. M. GOODEY and S. J. BENKOVIC, *Nat. Chem. Biol.* **4**, 474 (2008).

[10] C. A. TRACEWELL and F. H. ARNOLD, *Curr. Opin. Chem. Biol.* **13**, 3 (2009).

[11] C. JAECKEL, P. KAST, and D. HILVERT, *Annu. Rev. Biophys.* **37**, 153 (2008).

[12] R. GOUNDER and M. E. DAVIS, *AIChE J.* **59**, 3349 (2013).

[13] B. ALBERTS, J. BRAY, D ABD LEWIS, M. RAFF, K. ROBERTS, and J. D. WATSON, *Molecular Biology of the Cell*, Garland Publishing, Inc, 1994.

[14] P. BOYER, *Annu. Rev. Biochem.* **66**, 717 (1997).

[15] H. NOJI and M. YOSHIDA, *J. Biol. Chem.* **276**, 1665 (2001).

[16] T. UCHIHASHI, R. IINO, T. ANDO, and H. NOJI, *Science* **333**, 755 (2011).

[17] M. GEEVES and K. HOLMES, The molecular mechanism of muscle contraction, volume 71 of *Adv. Protein Chem.*, p. 161, 2005.

[18] H. L. SWEENEY and A. HOUDUSSE, Structural and Functional Insights into the Myosin Motor Mechanism, in *ANNUAL REVIEW OF BIOPHYSICS, VOL 39*, edited by REES, DC AND DILL, KA AND WILLIAMSON, JR, volume 39 of *Annu. Rev. Biophys.*, pp. 539–557, 2010.

[19] A. MALNASI-CSIZMADIA and M. KOVACS, *Trends Biochem. Sci.* **35**, 684 (2010).

[20] C. BUSTAMANTE, D. KELLER, and G. OSTER, *Acc. Chem. Res.* **34**, 412 (2001).

[21] P. C. WHITFORD, K. Y. SANBONMATSU, and J. N. ONUCHIC, *Rep. Prog. Phys.* **75** (2012).

[22] Y. GAO, W. YANG, and M. KARPLUS, *Cell* **123**, 195 (2005).

[23] M. KARPLUS and Y. GAO, *Curr. Opin. Struct. Biol.* **14**, 250 (2004).

[24] K. NAM, J. PU, and M. KARPLUS, *Proc. Natl. Acad. Sci. USA* **111**, 17851 (2014).

[25] K.-I. OKAZAKI and G. HUMMER, *Proc. Natl. Acad. Sci. USA* **110**, 16468 (2013).

[26] S. KOPPOLE, J. C. SMITH, and S. FISCHER, *Structure* **15**, 825 (2007).

[27] S. KOPPOLE, J. C. SMITH, and S. FISCHER, *J. Mol. Biol.* **361**, 604 (2006).

[28] F. A. KIANI and S. FISCHER, *Proc. Natl. Acad. Sci. USA* **111**, E2947 (2014).

[29] S. HAYASHI, H. UENO, A. R. SHAIKH, M. UMEMURA, M. KAMIYA, Y. ITO, M. IKEGUCHI, Y. KOMORIYA, R. IINO, and H. NOJI, *J. Am. Chem. Soc* **134**, 8447 (2012).

[30] M. J. MCGRATH, I. F. W. KUO, S. HAYASHI, and S. TAKADA, *J. Am. Chem. Soc* **135**, 8908 (2013).

[31] G. LI and Q. CUI, *J. Phys. Chem. B* **108**, 3342 (2004).

[32] H. YU, L. MA, Y. YANG, and Q. CUI, *PLoS Comput. Biol.* **3**, 0199 (2007).

[33] Y. YANG, H. YU, and Q. CUI, *J. Mol. Biol.* **381**, 1407 (2008).

[34] T. GLENNON, J. VILLA, and A. WARSHEL, *Biochemistry* **39**, 9641 (2000).

[35] R. B. PRASAD, N. V. PLOTNIKOV, J. LAMEIRA, and A. WARSHEL, *Proc. Natl. Acad. Sci. USA* **110**, 20509 (2013).

[36] A. J. ADAMCZYK and A. WARSHEL, *Proc. Natl. Acad. Sci. USA* **108**, 9827 (2011).

[37] J. AQVIST and S. C. L. KAMERLIN, *ACS Cata.* **6**, 1737 (2016).

[38] C. A. SMITH and I. RAYMENT, *Biochemistry* **35**, 5404 (1996).

[39] C. BAUER, H. HOLDEN, J. THODEN, R. SMITH, and I. RAYMENT, *J. Biol. Chem.* **275**, 38494 (2000).

[40] M. A. GEEVES and K. HOLMES, *Annu. Rev. Biochem.* **68**, 687 (1999).

[41] S. FISCHER, B. WINDSHUGEL, D. HORAK, K. HOLMES, and J. SMITH, *Proc. Natl. Acad. Sci. USA* **102**, 6873 (2005).

[42] W. W. CLELAND, , and A. C. HENGGE, *Chem. Rev.* **106**, 3252 (2006).

[43] I. RAYMENT, *J. Biol. Chem.* **271**, 15850 (1996).

[44] M. DITTRICH, S. HAYASHI, and K. SCHULTEN, *Biophys. J.* **85**, 2253 (2003).

[45] T. RUDACK, F. XIA, J. SCHLITTER, C. KOETTING, and K. GERWERT, *Proc. Natl. Acad. Sci. USA* **109**, 15295 (2012).

[46] K. GAVRILJUK, E.-M. GAZDAG, A. ITZEN, C. KOETTING, R. S. GOODY, and K. GERWERT, *Proc. Natl. Acad. Sci. USA* **109**, 21348 (2012).

[47] C. KOETTING and K. GERWERT, *Biol. Chem.* **396**, 131 (2015).

[48] M. KLAHN, J. SCHLITTER, and K. GERWERT, *Biophys. J.* **88**, 3829 (2005).

[49] H. ONISHI, S. KOJIMA, K. KATOH, K. FUJIWARA, H. MARTINEZ, and M. MORALES, *Proc. Natl. Acad. Sci. USA* **95**, 6653 (1998).

[50] H. ONISHI, T. OHKI, N. MOCHIZUKI, and M. MORALES, *Proc. Natl. Acad. Sci. USA* **99**, 15339 (2002).

[51] B. L. GRIGORENKO, A. V. ROGOV, I. A. TOPOL, S. K. BURT, H. M. MARTINEZ, and A. V. NEMUKHIN, *Proc. Natl. Acad. Sci. USA* **104**, 7057 (2007).

[52] M. MCCULLAGH, M. G. SAUNDERS, and G. A. VOTH, *J. Am. Chem. Soc.* **136**, 13053 (2014).

[53] D. RICCARDI, P. SCHAEFER, Y. YANG, H. YU, N. GHOSH, X. PRAT-RESINA, P. KONIG, G. LI, D. XU, H. GUO, M. ELSTNER, and Q. CUI, *J. Phys. Chem. B* **110**, 6458 (2006).

[54] Q. CUI and M. KARPLUS, *J. Phys. Chem. B* **106**, 7927 (2002).

[55] Q. CUI and M. KARPLUS, *J. Am. Chem. Soc* **124**, 3093 (2002).

[56] M. GAUS, Q. CUI, and M. ELSTNER, *J. Chem. Theory Comput.* **7**, 931 (2011).

[57] M. GAUS, A. GOEZ, and M. ELSTNER, *J. Chem. Theory Comput.* **9**, 338 (2013).

[58] M. GAUS, X. LU, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **10**, 1518 (2014).

[59] X. LU, M. GAUS, M. ELSTNER, and Q. CUI, *J. Phys. Chem. B* **119**, 1062 (2015).

[60] D. ROSTON, D. DEMAPAN, and Q. CUI, *J. Am. Chem. Soc* **138**, 7386 (2016).

[61] D. ROSTON and Q. CUI, *J. Am. Chem. Soc* **138**, 11946 (2016).

[62] A. LAIO and F. L. GERVASIO, *Rep. Prog. Phys.* **71** (2008).

[63] L. MARAGLIANO, A. FISCHER, E. VANDEN-EIJNDEN, and G. CICCOTTI, *J. Chem. Phys.* **125** (2006).

[64] V. OVCHINNIKOV, M. KARPLUS, and E. VANDEN-EIJNDEN, *J. Chem. Phys.* **134** (2011).

[65] V. OVCHINNIKOV, M. CECCHINI, E. VANDEN-EIJNDEN, and M. KARPLUS, *Biophys. J.* **101**, 2436 (2011).

[66] P. KONIG, M. HOFFMANN, T. FRAUENHEIM, and Q. CUI, *J. Phys. Chem. B* **109**, 9082 (2005).

[67] C. N. ROWLEY and B. ROUX, *J. Chem. Theory Comput.* **8**, 3526 (2012).

[68] D. BASHFORD and M. KARPLUS, *Biochemistry* **29**, 10219 (1990).

[69] W. IM, S. BERNÈCHE, and B. ROUX, *J. Chem. Phys.* **114**, 2924 (2001).

[70] P. SCHAEFER, D. RICCARDI, and Q. CUI, *J. Chem. Phys.* **123**, 014905 1 (2005).

[71] Q. CUI and M. ELSTNER, *Phys. Chem. Chem. Phys.* **16**, 14368 (2014).

[72] M. GAUS, Q. CUI, and M. ELSTNER, *WIREs Comput. Mol. Sci.* **4**, 49 (2014).

[73] A. S. CHRISTENSEN, T. KUBAR, Q. CUI, and M. ELSTNER, *Chem. Rev.* **116**, 5301 (2016).

[74] C. BROOKS and M. KARPLUS, *J. Mol. Biol.* **208**, 159 (1989).

[75] A. D. MACKERELL, D. BASHFORD, BELLOTT, R. L. DUNBRACK, J. D. EVANSECK, M. J. FIELD, S. FISCHER, J. GAO, H. GUO, S. HA, D. JOSEPH-MCCARTHY, L. KUCHNIR, K. KUCZERA, F. T. K. LAU, C. MATTOS, S. MICHNICK, T. NGO, D. T. NGUYEN, B. PRODHOM, W. E. REIHER, B. ROUX, M. SCHLENKRICH, J. C. SMITH, R. STOTE, J. STRAUB, M. WATANABE, J. WIÓRKIEWICZ-KUCZERA, D. YIN, and M. KARPLUS, *J. Phys. Chem. B* **102**, 3586 (1998).

[76] R. H. STOTE, D. J. STATES, and M. KARPLUS, *J. Chim. Phys.* **88**, 2419 (1991).

[77] D. RICCARDI and Q. CUI, *J. Phys. Chem. A* **111**, 5703 (2007).

[78] X. LU and Q. CUI, *J. Phys. Chem. B* **117**, 2005 (2013).

[79] A. LAIO and M. PARRINELLO, *Proc. Natl. Acad. Sci. USA* **99**, 12562 (2002).

[80] M. BONOMI, D. BRANDUARDI, G. BUSSI, C. CAMILLONI, D. PROVASI, P. RAITERI, D. DONADIO, F. MARINELLI, F. PIETRUCCI, R. A. BROGLIA, and M. PARRINELLO, *Comput. Phys. Commun.* **180**, 1961 (2009).

[81] B. R. BROOKS, C. L. BROOKS, III, A. D. MACKERELL, JR., L. NILSSON, R. J. PETRELLA, B. ROUX, Y. WON, G. ARCHONTIS, C. BARTELS, S. BORESCH, A. CAFLISCH, L. CAVES, Q. CUI, A. R. DINNER, M. FEIG, S. FISCHER, J. GAO, M. HODOSCEK, W. IM, K. KUCZERA, T. LAZARIDIS, J. MA, V. OVCHINNIKOV, E. PACI, R. W. PASTOR, C. B. POST, J. Z. PU, M. SCHAEFER, B. TIDOR,

R. M. VENABLE, H. L. WOODCOCK, X. WU, W. YANG, D. M. YORK, and M. KARPLUS, *J. Comput. Chem.* **30**, 1545 (2009).

[82] P. H. KNIG, N. GHOSH, M. HOFFMANN, M. ELSTNER, E. TAJKHORSHID, T. FRAUENHEIM, and Q. CUI, *J. Phys. Chem. A* **110**, 548 (2006).

[83] J.-P. RYCKAERT, G. CICCOTTI, and H. J. BERENDSEN, *J. Comp. Phys.* **23**, 327 (1977).

[84] A. BARDUCCI, G. BUSSI, and M. PARRINELLO, *Phys. Rev. Lett.* **100** (2008).

[85] P. KONIG, N. GHOSH, M. HOFFMANN, M. ELSTNER, E. TAJKHORSHID, T. FRAUENHEIM, and Q. CUI, *J. Phys. Chem. A* **110**, 548 (2006).

[86] R. ELBER and M. KARPLUS, *Chem. Phys. Lett.* **139**, 375 (1987).

[87] G. HENKELMAN, B. UBERUAGA, and H. JONSSON, *J. Chem. Phys.* **113**, 9901 (2000).

[88] C. BAGSHAW, D. TRENTHAM, R. WOLCOTT, and P. BOYER, *Proc. Natl. Acad. Sci. USA* **72**, 2592 (1975).

[89] M. WEBB and D. TRENTHAM, *J. Biol. Chem.* **256**, 910 (1981).

[90] J. K. LASSILA, J. G. ZALATAN, and D. HERSCHLAG, Biological Phosphoryl-Transfer Reactions: Understanding Mechanism and Catalysis, in *ANNUAL REVIEW OF BIOCHEMISTRY, VOL 80*, edited by KORNBERG, RD AND RAETZ, CRH AND ROTHMAN, JE AND THORNER, JW, volume 80 of *Annu. Rev. Biochem.*, pp. 669–702, 2011.

[91] D. TRENTHAM, J. ECCLESTON, and C. BAGSHAW, *Q. ReV. Biophys.* **9**, 217 (1976).

[92] C. MURPHY, R. ROCK, and J. SPUDICH, *Nat. Cell Biol.* **3**, 311 (2001).

[93] A. MALNASI-CSIZMADIA, D. PEARSON, M. KOVACS, R. WOOLLEY, M. GEEVES, and C. BAGSHAW, *Biochemistry* **40**, 12727 (2001).

[94] M. D. DAILY, H. YU, G. N. PHILLIPS, JR., and Q. CUI, Allosteric Activation Transitions in Enzymes and Biomolecular Motors: Insights from Atomistic and Coarse-Grained Simulations, in *DYNAMICS IN ENZYME CATALYSIS*, edited by KLINMAN, J AND HAMMESSCHIFFER, S, volume 337 of *Topics in Curr. Chem.*, pp. 139–164, 2013.

[95] Y. YANG and Q. CUI, *J. Phys. Chem. B* **113**, 4930 (2009).

[96] D. MILLER and F. WESTHEIM, *J. Am. Chem. Soc* **88**, 1507 (1966).

[97] F. A. KIANI and S. FISCHER, *Curr. Opin. Struct. Biol.* **31**, 115 (2015).

[98] K. VENKATASUBBAN and R. SCHOWEN, *CRC Crit Rev. Biochm.* **17**, 1 (1984).

[99] S. SCHWARZL, J. SMITH, and S. FISCHER, *Biochemistry* **45**, 5830 (2006).

[100] Y. YANG and Q. CUI, *J. Phys. Chem. A* **113**, 12439 (2009).

[101] S. D. FRIED, S. BAGCHI, and S. G. BOXER, *Science* **346**, 1510 (2014).

[102] S. D. FRIED and S. G. BOXER, *Science* **349** (2015).

[103] A. NATARAJAN, F. YABUKARSKI, V. LAMBA, J. P. SCHWANS, F. SUNDEN, and D. HERSCHLAG, *Science* **349** (2015).

[104] H. K. PRIVETT, G. KISS, T. M. LEE, R. BLOMBERG, R. A. CHICA, L. M. THOMAS, D. HILVERT, K. N. HOUK, and S. L. MAYO, *Proc. Natl. Acad. Sci. USA* **109**, 3790 (2012).

[105] M. P. FRUSHICHEVA, J. CAO, Z. T. CHU, and A. WARSHEL, *Proc. Natl. Acad. Sci. USA* **107**, 16869 (2010).

[106] M. P. FRUSHICHEVA, M. J. L. MILLS, P. SCHOPF, M. K. SINGH, R. B. PRASAD, and A. WARSHEL, *Curr. Opin. Chem. Biol.* **21**, 56 (2014).

[107] K. SWIDEREK, I. TUNON, V. MOLINER, and J. BERTRAN, *ACS Cata.* **5**, 2587 (2015).

[108] P. O'BRIEN and D. HERSCHLAG, *Chem. Biol.* **6**, R91 (1999).

[109] O. KHERSONSKY, C. ROODVELDT, and D. S. TAWFIK, *Curr. Opin. Chem. Biol.* **10**, 498 (2006).

[110] G. HOU and Q. CUI, *J. Am. Chem. Soc.* **135**, 10457 (2013).

[111] A. PABIS and S. C. L. KAMERLIN, *Curr. Opin. Struct. Biol.* **37**, 14 (2016).

[112] D. S. TAWFIK, *Curr. Opin. Chem. Biol.* **21**, 73 (2014).

[113] I. BALABIN and J. ONUCHIC, *Science* **290**, 114 (2000).

[114] J. REGAN, S. RISSER, D. BERATAN, and J. ONUCHIC, *J. Phys. Chem.* **97**, 13083 (1993).

[115] Y. SHAM, I. MUEGGE, and A. WARSHEL, *Proteins: Struct., Funct., Bioinf.* **36**, 484 (1999).

[116] D. RICCARDI, S. YANG, and Q. CUI, *Biochim. Biophys. Acta* **1804**, 342 (2010).

[117] J. M. J. SWANSON, C. M. MAUPIN, H. CHEN, M. K. PETERSEN, J. XU, Y. WU, and G. A. VOTH, *J. Phys. Chem. B* **111**, 4300 (2007).

[118] O. P. ERNST, D. T. LODOWSKI, M. ELSTNER, P. HEGEMANN, L. S. BROWN, and H. KANDORI, *Chem. Rev.* **114**, 126 (2014).

[119] S. MAEDA, K. OHNO, and K. MOROKUMA, *Phys. Chem. Chem. Phys.* **15**, 3683 (2013).

[120] P. M. ZIMMERMAN, *J. Comp. Chem.* **34**, 1385 (2013).

[121] M. P. JACOBSON, C. KALYANARAMAN, S. ZHAO, and B. TIAN, *Trends Biochem. Sci.* **39**, 363 (2014).

[122] E. EISENBERG and T. HILL, *Science* **227**, 999 (1985).

[123] J. CAMPISI and F. D. DI FAGAGNA, *Nat. Rev. Mol. Cell Biol.* **8**, 729 (2007).

[124] H. M. SENN and W. THIEL, *Angew. Chem., Int. Ed.* **48**, 1198 (2009).

[125] Q. CUI, , and M. KARPLUS, *J. Am. Chem. Soc* **124**, 3093 (2002).

[126] Y. ZHANG, J. KUA, and J. A. MCCAMMON, *J. Phys. Chem. B* **107**, 4459 (2003).

[127] S. FISCHER and M. KARPLUS, *Chem. Phys. Lett.* **194**, 252 (1992).

[128] Y. YANG, H. YU, D. YORK, M. ELSTNER, and Q. CUI, *J. Chem. Theory Comput.* **4**, 2067 (2008).

Chapter 7

Concluding Remarks

In Chapter 2, we extend the parameterization of DFTB3 to phosphorus and sulfur considering the importance of phosphorus and sulfur in (bio)chemistry. The parameterization is carried out in a framework consistent with the DFTB3/3OB set for O, N, C and H, thus the resulting parameters can be used to describe a broad set of organic and biologically relevant P/S-containing molecules. The 3*d* orbitals, which are known to be polarization functions essential to the proper description of structure and energetics of P/S-containing molecules, are included in the parameterization, in contrast to a few popular wave function based semi-empirical methods such as PM3/PDDG. The parameters are tested with a fairly diverse set of molecules of chemical and biological relevance. We focus on the geometries, reaction energies, proton affinities and hydrogen bonding interactions of these molecules. These properties calculated at the DFTB3/3OB level are compared to results of DFT (B3LYP and PBE), *ab initio* (MP2, G3B3) and several popular semi-empirical methods (PM6 and PDDG), as well as predictions of DFTB3 with the older parameterization (the MIO set). In general, DFTB3/3OB is a major improvement over the previous parameterization (DFTB3/MIO), and for the majority cases tested here, it also outperforms PM6 and PDDG. Compared to high-level calculations, DFTB3/3OB is most successful at predicting geometries; the energetics are largely on par with results of DFT (especially PBE, which is the functional used in DFTB3 for deriving the relevant matrix elements) using a medium basis set (e.g., 6-31G(d)), although larger errors are seen more often with DFTB3. The extensive tests also indicate that there are remaining cases for which the current DFTB3 methodology has trouble properly describing. One limitation is that DFTB3/3OB is still incapable of providing reliable energetics for phosphate hydrolysis reactions that involve a change in the coordination number of the phosphorus. As a temporary solution, we develop a specific pa-

rameterization, 3OB/OPhyd, where an overbinding is introduced to O-P interactions to empirically minimize the errors for a set of reference reactions. Although this is clearly not a satisfactory solution for the long-term, the 3OB/OPhyd can serve as a useful tool for exploring the energy landscape of phosphoryl transfer reactions in biological systems, especially for structural properties. Alternatively, the 3OB structures remain reliable for these challenging cases and one may still use the standard 3OB parameters to obtain structures and improve energetics with higher level methods.

In Chapter 3, we carry out similar parameterization procedure to magnesium and zinc, because magnesium and zinc play many important roles in metalloenzymes and various chemical systems. Benchmark calculations in the gas phase and in the condensed phase with a QM/MM framework demonstrate that the current parameterization generally leads to reliable structures and semiquantitative energetics (with typically a MAD of $\sim$3-5 kcal/mol compared to high level *ab initio* calculations). In the gas phase, DFTB3/3OB are tested with a diverse range of molecules that reflect the typical coordination environment of magnesium and zinc in biomolecules, including di-nuclear metal sites. DFTB3/MM calculations are expected to be effective for studies in condensed phase systems. For example, our solution benchmark study for the magnesium, zinc ion have reproduced experimental solvation structures and relative pK_a of metal bound water, especially when both first and second solvation shells are treated as DFTB3/3OB. For the enzyme benchmark, DFTB3/3OB are successful at producing active site structures in myosin and AP relative to available crystal structures; for AP, DFTB3/3OB also leads to semi-quantitative energetics for an active site model.

The energetic properties from DFTB3 calculations should be taken as semi-quantitative in nature and can be improved by combining with *ab initio* QM/MM calculations. For example, the fact that single point high-level calculations using DFTB3/3OB geometries often exhibit small errors suggest that, DFTB3/3OB is likely effective as the low-level QM approach that drives the sampling for reliable free energy simulations in the framework of dual-level QM/MM calculation. Finally, the remaining limitations highlight that it is also worthwhile continuing to improve the DFTB3 methodology, such as by including multipole terms for the charge fluctuation and better description of polarization and short-range repulsions.

In Chapter 4, we present several recent developments in free energy methods most relevant to QM/MM simulations and discuss a few topics motivated by these developments using simple but informative models. When a low-level of QM theory, such as a semi-empirical method, is adequate, potential of mean

force/alchemical free energy simulations can be readily carried out with several established techniques, such as umbrella sampling, metadynamics and free energy perturbation. For complex reactions, minimum free energy path methods have become more prevalent, although we caution that the results, especially free energy profiles, can be sensitive to several parameters in these calculations, such as the choice of the CV space used to span the string and force constants used to estimate the mean force with restrained molecular dynamics. Besides, it is also productive to integrate metadynamics with finite temperature string type of calculations to explore the mechanism of complex reactions; metadynamics using well-chosen CVs can be very effective at identifying new reaction channels and therefore provide good initial guess for string calculations, which are local in nature and therefore rely on well-chosen initial pathways. On the other hand, for high-level QM/MM potentials, it likely depends on the problem in hand. When the QM region is fairly rigid and fluctuations orthogonal to the reaction path are largely harmonic in nature, the QM/MM-MFEP framework pioneered by Yang and co-workers is expected to work well. When the QM region is highly dynamical, an appropriate alternative is to determine the minimum free energy pathway at a semi-quantitative level using direct sampling with a low-level (semi-empirical) QM/MM method and then correct the energetics for key regions following perturbative schemes. Provided that the key structures on the free energy surface remain similar at different levels of theory, our tests indicate that the perturbative schemes can lead to substantially improved barriers and in favorable cases, only low-level sampling is needed. In general, to make sure that perturbative schemes give meaningful results, however, it is essential to conduct careful analysis of the energy difference distributions and key structural features using trajectories at different levels of theory.

In Chapter 5, we conduct charging free energie simulationsin both solution and protein systems calculated with GSBP and periodic boundary condition (PBC) simulations. Because many approximations are inherent to the finite sphere boundary framework, we want to have a better understanding whether these approximations make a significant impact on the problem of interest. For charging (oxidation) free energy in proteins, good agreement between GSBP and PBC can be obtained when care is exercised. We observe that the smooth dielectric boundary approximation inherent to GSBP tends to make significant errors when the inner region is featured with a high net charge. However, the error can be corrected with Poisson-Boltzmann calculations using snapshots from GSBP simulations in a straightforward and robust manner. Due to the more complex charge and solvent distributions, the magnitudes of corrections in protein simulations appear

to be different from those derived for solution simulations. Therefore, it remains somewhat challenging to directly compare absolute charging free energies in PBC and GSBP simulations for protein systems; the relative charging/oxidation free energies, however, are more robust and PBC/GSBP results agree much better even for GSBP simulations with a fairly small inner region. For the specific protein system (CueR) studied, the use of a low-dielectric "constant" for the outer region is found to introduce only a small error to the oxidation free energy. These results are likely due to the fact that the site of interest is fairly solvent accessible and therefore well screened from the outer region. For a deeply buried site, a more sophisticated treatment for the dielectrics for the outer region might be needed.

In Chapter 6, we combine QM/MM simulations with metadynamics and the string method, and analyze in detail the mechanism of ATP hydrolysis activity in the myosin domain. Focusing mainly on the hydrolysis-competent pre-powerstroke state, we identify multiple reaction pathways of ATP hydrolysis and computed the associated free energy profiles. The scissile bond in the γ-phosphate of ATP breaks early in the reaction; proton transfer from the lytic water to the γ phosphate through active site residues is an important part of the kinetic bottleneck of the hydrolysis. Different hydrolysis pathways with similar free energy barriers are found, indicating plasticity in the hydrolysis mechanism. To understand the factors that determine the catalytic efficiency, we also analyze hydrolysis pathways in PR (in which hydrolysis is disfavored), and in a hybrid model that has active site features of PPS state, with the overall conformation of the PR state. The results suggest that the ATPase activity depends directly on the positioning of nearby water molecules to conduct efficient in-line attack, but also, indirectly, on residues beyond the active site. From the functional perspective of avoiding futile hydrolysis, the apparent sensitivity of ATPase activity to the positioning of active site motifs is clearly essential, provided that structural transitions in the active site and distal regions are highly cooperative in nature. In the more general context, the results suggest that optimization of residues beyond the active site for electrostatic stabilization and pre-organization is likely important to rational enzyme design.

APPENDICES

Appendix A

List of Publications

1. "Charging Free Energy Calculations Using the Generalized Solvent Boundary Potential (GSBP) and Periodic Boundary Condition: A Comparative Analysis Using Ion Solvation and Oxidation Free Energy in Proteins", X. Lu and Q. Cui, *J. Phys. Chem. B* 117, 2005-2018 (2013).
2. "Parameterization of DFTB3/3OB for Sulfur and Phosphorus for Chemical and Biological Applications", M. Gaus, X. Lu, M. Elstner and Q. Cui, *J. Chem. Theory Comput.* 10, 1518-1537 (2014).
3. "Toward Quantitative Analysis of Metalloenzyme Function Using MM and Hybrid QM/MM Methods: Challenges, Methods and Recent Applications", M. Gaus, P. Goyal, G. Hou, X. Lu, X. Pang, J. Zienau, X. Xu, M. Elstner and Q. Cui, *Quantitative Computations in Biology (edited by Ruhong Zhou)* 33-82 (2014).
4. "Molecular Simulation of Water and Hydration Effects in Different Environments: Challenges and Developments for DFTB Based Models", P. Goyal, H. Qian, S. Irle, X. Lu, D. Roston, T. Mori, M Elstner and Q. Cui, *J. Phys. Chem. B* 118, 11007-11027 (2014).
5. "Parametrization of DFTB3/3OB for Magnesium and Zinc for Chemical and Biological Applications", X. Lu, M. Gaus, M. Elstner and Q. Cui, *J. Phys. Chem. B* 119, 1062-1082 (2015).
6. "QM/MM Free Energy Simulations: Recent Progress and Challenges", X. Lu, D. Fang, S. Ito, Y. Okamoto and Q. Cui, *Mol. Simul.* 42, 1056-1078 (2016)
7. "Regulation and Plasticity of Catalysis in Enzymes: Insights from Analysis of Mechanochemical Coupling in Myosin", X. Lu, V. Ovchinnikov, D. Demapan, D. Roston and Q. Cui, *Biochemistry* 56, 1482-

1497 (2017)