<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>Enzymes in the World of Computational Chemistry - Computational Studies on Structures, Dynamics, Mechanisms, and Specificity of Five Biologically Important Enzymes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Author(s)</strong></td>
<td>Tian, Boxue</td>
</tr>
<tr>
<td><strong>Publication Date</strong></td>
<td>2012-10-30</td>
</tr>
<tr>
<td><strong>Item record</strong></td>
<td><a href="http://hdl.handle.net/10379/3189">http://hdl.handle.net/10379/3189</a></td>
</tr>
</tbody>
</table>

Some rights reserved. For more information, please see the item record link above.
ENZYMES IN THE WORLD OF COMPUTATIONAL CHEMISTRY

Computational Studies on Structures, Dynamics, Mechanisms, and Specificity of Five Biologically Important Enzymes

by Boxue Tian

Research Supervisors:

Prof. Leif A. Eriksson

Prof. Paul V. Murphy

College of Science, School of Chemistry, NUI Galway

September 2012
# Table of Content

Abstract ........................................................................................................................................... 1

Acknowledgements ........................................................................................................................ 4

Abbreviations .................................................................................................................................. 5

Publications and Contributions ....................................................................................................... 6

Chapter 1 Introduction .................................................................................................................... 7
  1.1 Overview .................................................................................................................................... 7
  1.2 Enzyme Structure and Function Prediction ............................................................................. 9
  1.3 Enzyme Dynamics Simulations ............................................................................................... 11
  1.4 Mechanistic Study on Enzymes ............................................................................................. 15
  1.5 Computational Study on Enzyme Specificity ......................................................................... 19
  1.6 In Silico Design of Enzyme Inhibitors ................................................................................... 20

Chapter 2 Methodology .................................................................................................................. 24
  2.1 Overview .................................................................................................................................... 24
  2.2 Molecular Mechanics ............................................................................................................. 25
  2.3 Quantum Mechanics .............................................................................................................. 26
    2.3.1 The Schrödinger equation ............................................................................................... 26
    2.3.2 Hartree-Fock theory ....................................................................................................... 27
    2.3.3 Density Functional Theory ............................................................................................ 28
2.4 Quantum Mechanics/Molecular Mechanics ................................................................. 30
2.5 Molecular dynamics simulations .............................................................................. 32
2.6 Homology Modeling ................................................................................................. 33
2.7 Docking and Virtual Screening ................................................................................ 34

Chapter 3 Result and Discussion .................................................................................. 37

3.1 Computational Study on SrtA (Papers 1, 2) ......................................................... 37
    3.1.1 Introduction ........................................................................................................ 37
    3.1.2 L. monocytogenes SrtA homology model (Paper 1) ........................................ 39
    3.1.3 Arg197 Motion in L. monocytogenes SrtA Homology Model (Paper 1) .......... 41
    3.1.4 Arg197 Motion in S. aureus SrtA (Paper 2) ...................................................... 44
    3.1.5 Relative stability of the inactive and active forms (Paper 2) ......................... 45
    3.1.6 Substrate binding affinity (Paper 2) ................................................................. 47
    3.1.7 Acylation mechanism of SrtA: roles of Arg197 and Thr183 (Paper 2) .......... 48
    3.1.8 Inhibitors against SrtA (unpublished) ........................................................... 50

3.2 Computational Study on MHPCO (Papers 3, 4) .................................................. 52
    3.2.1 Introduction ........................................................................................................ 52
    3.2.2 DFT study on small model (Paper 3) ............................................................... 55
    3.2.3 DFT cluster study on MHPCO active site (Paper 4) ...................................... 58
    3.2.4 QM/MM (ONIOM) study on the whole MHPCO (Paper 4) ..................... 63

3.3 Computational Study on PBGS (Paper 5) ............................................................ 70
References ................................................................................................................................. 104

Appendix ..................................................................................................................................... 114

Appendix 1. Computational details .......................................................................................... 114

A1.1 Paper 1 ................................................................................................................................. 114

A1.2. Paper 2 ............................................................................................................................... 116

A1.3. Paper 3 ............................................................................................................................... 120

A1.4. Paper 4 ............................................................................................................................... 120

A1.5. Paper 5 ............................................................................................................................... 121

A1.6. Paper 6 ............................................................................................................................... 122

A1.7 Paper 7 ............................................................................................................................... 123

Appendix 2. Published papers .................................................................................................... 124
Abstract

The thesis demonstrates computational chemical methodology applied to some biologically important enzymes. Five enzymes have been studied, i.e. sortase A (SrtA), 2-Methyl-3-hydroxypyridine-5-carboxylic acid oxygenase (MHPCO), porphobilinogen synthase (PBGS), D-glucarate dehydratase (GlucD) and oxidosqualene-lanosterol cyclase (OSC). Different levels of theory have been used to address questions related to structure, function, dynamics, mechanism, specificity, and inhibitor design of the above enzymes.

The sortase A (SrtA) enzyme, which catalyzes the peptidoglycan cell wall anchoring reaction of LPXTG surface proteins, has been proposed to be a universal target for therapeutic agents against Gram-positive bacteria. The structure of the *L. monocytogenes* SrtA enzyme-substrate complex was obtained using homology modeling, molecular docking and molecular dynamics simulations. The active site arginine (Arg 197) was found to be able change its hydrogen donor interactions from the LP backbone carbonyl groups of the LPXTG substrate in the inactive form, to the TG backbone carbonyls in the active form. Similar motion of Arg197 was also observed in the *S. aureus* SrtA system. The catalytic mechanism of *S. aureus* SrtA was then systematically studied using MD simulations, ONIOM(DFT:MM) calculations, and QM/MM charge deletion analysis. The catalytic roles of Arg197 and Thr183 were analyzed. Our calculations show that Arg197 has several important roles in the mechanism. It is crucial for substrate binding, and is capable of reversible shift of its hydrogen bonds between the LP and TG carbonyls of the LPXTG substrate motif, depending on the protonation state of the catalytic Cys184-His120 dyad. Arg197 stabilizes the catalytic dyad in the active ion pair form but at the same time raises the barrier to acylation by approximately 8 kcal/mol. Thr183 is also essential for the catalytic reaction in that it correspondingly lowers the barrier by the same amount via electrostatic interactions. The catalytic mechanism proceeds via proton transfer from His120, followed by nucleophilic attack from the thiolate anion of Cys184. The data thus supports the proposed reverse protonation mechanism (RPM), and disproves the hypothesis of the Arg197 generating an oxyanion hole to stabilize the tetrahedral intermediate of the reaction.
MHPCO catalyzes the hydroxylation and subsequent ring-opening of the aromatic substrate MHPC to give the aliphatic product R-(N acetylamino methylene)succinic acid (AAMS), which is the essential ring-opening step in the bacterial degradation of vitamin B6. MHPCO belongs to the flavin-containing aromatic hydroxylases family. However, MHPCO is capable of catalyzing a subsequent aromatic ring-cleavage reaction to give acyclic products rather than hydroxylated aromatic ones. The catalytic mechanism of MHPCO has been systematically studied using DFT (MPWB1K and B3LYP) and ONIOM(DFT:MM) methods. Our DFT calculations show that the rearomatization of the hydroxylated intermediate occurs spontaneously in aqueous solution; this implies that the ring-opening process occurs inside the active site, in which limited water is available. The instability of the hydroxylated intermediate of MHPCO is the main reason why acyclic products are formed. Previously proposed mechanisms for the ring-opening step were studied, and were shown to be less likely to occur. Two new pathways (pathway A and B) with reasonable barrier heights are reported herein. Both DFT and ONIOM calculations show that the ring-opening pathway B, in which an epoxy transition state is formed, is more favored than the direct C2-C3 cleavage pathway A. Our calculations show that the active-site residues Arg211 and Tyr223 have a minor effect on the reaction, while the peptide bond of Pro295-Ala296, the side chain of Tyr82 and several crystal water molecules affect the reaction energy profile considerably. Different QM/MM partitioning schemes have been used to study the enzymatic reaction, and the results show that both the reaction barriers for the hydroxylation and the ring-opening pathways are sensitive to the QM/MM partitioning.

Porphobilinogen synthase (PBGS) catalyzes the asymmetric condensation and cyclization of two 5-aminolevulinic acid (5-ALA) substrate molecules to give porphobilinogen (PBG), and is known as the first common step in the biosynthesis of the tetrapyrroles. The chemical step of PBGS is herein revisited using QM/MM (ONIOM) calculations. Two different protonation states and several different mechanisms are considered. Previous mechanisms based on DFT-only calculations are shown unlikely to occur. According to these new calculations, the deprotonation step rather than ring closure is rate-limiting. Both the C-C bond formation first mechanism and the C-N bond formation first mechanism are possible, depending on how the A-site ALA binds to the enzyme. We furthermore propose that future work should focus on the substrate binding step rather than the enzymatic mechanism.
D-glucarate dehydratase (GlucD) catalyzes the dehydration of D-glucarate or L-idarate to give 5-keto-4-deoxy-D-glucarate (5-KDG). The stereo-specificity of GlucD is explored by combined docking and QM/MM calculations. According to our calculations, both the substrate binding and the chemical steps of GlucD contribute to substrate specificity. The current approach will be used for assisting enzyme function assignment.

Oxidosqualene-lanosterol cyclase (OSC) is a key enzyme in the biosynthesis of cholesterol. The catalytic mechanism and the product specificity of OSC have been studied by using QM/MM calculations. According to our calculations, the protonation of the epoxide ring of oxidosqualene is rate-limiting. The wild type OSC (which generates lanosterol), H232S (which generates parkeol) and H232T (which generates protosta-12,24-dien-3-beta-ol) mutants were modeled, in order to explain the product specificity thereof. We show that the product specificity of OSC at the hydride/methyl-shifting stage is unlikely to be achieved by the stabilization of the cationic intermediates, because the precursor of lanosterol is not the most stable cationic intermediate for the wild type OSC. The energy barriers for the product-determining conversions are related to the product specificity of different OSC mutants. We thus suggest that the product specificity of OSC is likely to be controlled by kinetics, rather than thermodynamics.
Acknowledgements

I would like to express my deepest gratitude to my supervisor Prof. Leif A. Eriksson for his invaluable support and guidance during my PhD study.

I wish to sincerely thank Prof. Paul V. Murphy, Prof. Robert J. Woods and Prof. Matthew P. Jacobson (UCSF) for helpful assistance and advice on my scientific research.

Many thanks to Prof. Åke Strid (Örebro University), Dr. Yao-Quan Tu (KTH), Dr. Emma S. E. Eriksson, Dr. Edvin Erdtman (University of Borås), and Dr. Frank Wallrapp (UCSF) for their contributions to my research, and for their assistance beyond science.

Prof. Tian He (East China University of Science and Technology) is acknowledged for offering me the opportunity to come to Europe.

Special thanks also to Dr. Khalifa. A. K. Musa, Li Gao, Min Wu, Chun-Xia Gao, and other members in Prof. Eriksson’s lab; Dr. Elisa Fadda, Oliver Grant, and other members in Prof. Woods’ lab; Dr. Chakrapani Kalyanaraman, Dr. Lan Hua, and other members in Prof. Jacobson’s lab; as well as Dr. Chong-Wen Zhou, Dr. Qiong Zhang, Dr. Rong-Zhen Liao and other friends who are working on theoretical chemistry and computational modeling.

Dr. Fawaz Aldabbagh (Graduate Research Committee), Louise Coyle (International Affairs Office), Karen Kelly, Judy Buckley and other administrators and faculty in school of chemistry and NUI Galway are acknowledged for assistance on various paper works.

I also wish to express my gratitude to my beloved families; for their understanding and endless love, through the duration of my PhD studies.

Last but not least, thanks a million to NUI Galway for financial support and for provision of computational facilities. The SFI/HEA Irish Centre for High-End Computing (ICHEC) is also acknowledged for the provision of computational facilities and support.
# Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>5HN</td>
<td>5-hydroxynicotinic acid</td>
</tr>
<tr>
<td>ADMET</td>
<td>absorption, distribution, metabolism, excretion and toxicity</td>
</tr>
<tr>
<td>AO</td>
<td>atomic orbital</td>
</tr>
<tr>
<td>BO</td>
<td>Born-Oppenheimer</td>
</tr>
<tr>
<td>DFT</td>
<td>density functional theory</td>
</tr>
<tr>
<td>EE</td>
<td>electronic embedding or electrostatic embedding</td>
</tr>
<tr>
<td>FADHOOH</td>
<td>C-(4a)-flavinhydroperoxide</td>
</tr>
<tr>
<td>GlucD</td>
<td>D-glucarate dehydratase</td>
</tr>
<tr>
<td>HF</td>
<td>Hartree-Fock</td>
</tr>
<tr>
<td>HTVS</td>
<td>high throughput virtual screening</td>
</tr>
<tr>
<td>KS</td>
<td>Kohn-Sham</td>
</tr>
<tr>
<td>L. monocytogenes</td>
<td><em>Listeria monocytogenes</em></td>
</tr>
<tr>
<td>LCAO</td>
<td>linear combination of atomic orbitals</td>
</tr>
<tr>
<td>LDA</td>
<td>local density approximations</td>
</tr>
<tr>
<td>LPXTG</td>
<td>leucine, proline, X, threonine, and glycine, where X is any amino acid</td>
</tr>
<tr>
<td>ME</td>
<td>mechanical embedding</td>
</tr>
<tr>
<td>MHPCO</td>
<td>2-Methyl-3-hydroxypyridine-5-carboxylic acid oxygenase</td>
</tr>
<tr>
<td>MM</td>
<td>molecular mechanical</td>
</tr>
<tr>
<td>MO</td>
<td>molecular orbital</td>
</tr>
<tr>
<td>MUE</td>
<td>mean unsigned error</td>
</tr>
<tr>
<td>NA-Cys</td>
<td>nucleophilic attack by the cysteine thiolate anion</td>
</tr>
<tr>
<td>ONIOM</td>
<td>own N-layer integrated molecular orbital molecular mechanics</td>
</tr>
<tr>
<td>OSC</td>
<td>oxidosqualene-lanosterol cyclase</td>
</tr>
<tr>
<td>PA</td>
<td>proton affinity</td>
</tr>
<tr>
<td>PBGS</td>
<td>porphobilinogen synthase</td>
</tr>
<tr>
<td>PHBH</td>
<td>p-hydroxybenzoate hydroxylase</td>
</tr>
<tr>
<td>PT-His</td>
<td>a proton transfer from histidine</td>
</tr>
<tr>
<td>QM</td>
<td>quantum mechanical</td>
</tr>
<tr>
<td>RMSF</td>
<td>root mean square fluctuation</td>
</tr>
<tr>
<td>RPM</td>
<td>reverse protonation mechanism</td>
</tr>
<tr>
<td>S. aureus</td>
<td><em>Staphylococcus aureus</em></td>
</tr>
<tr>
<td>SBVS</td>
<td>structure based virtual screening</td>
</tr>
<tr>
<td>SCF</td>
<td>Self-Consistent Field</td>
</tr>
<tr>
<td>SD</td>
<td>Slater determinant</td>
</tr>
<tr>
<td>SP</td>
<td>standard precision</td>
</tr>
<tr>
<td>SrtA</td>
<td>sortase A</td>
</tr>
<tr>
<td>T-G</td>
<td>threonine and glycine</td>
</tr>
<tr>
<td>TI</td>
<td>tetrahedral intermediate</td>
</tr>
<tr>
<td>TST</td>
<td>transition state theory</td>
</tr>
<tr>
<td>XP</td>
<td>extra precision</td>
</tr>
</tbody>
</table>
## Publications and Contributions

[D] = designing the project; [R] = running calculations; [P] = preparing the manuscript; [C] = correcting manuscript

### Publications included in this thesis

<table>
<thead>
<tr>
<th>Paper</th>
<th>Authors</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
<th>DOI</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper 1</td>
<td>B. Tian, L. A. Eriksson</td>
<td>Structural changes of Listeria Monocytogenes Sortase A: A key to understanding the catalytic mechanism</td>
<td><em>Proteins</em> 2011, 79, 1564-1572</td>
<td>DOI: 10.1002/prot.22983</td>
<td>[R] [P] [C]</td>
<td></td>
</tr>
</tbody>
</table>

### Publications not included in this thesis

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Journal</th>
<th>Year</th>
<th>DOI</th>
<th>Contributions</th>
</tr>
</thead>
</table>
Chapter 1 Introduction

1.1 Overview

Enzymes are proteins that are capable of catalyzing chemical reactions. Enzymatic reactions are efficient, selective and occur under mild conditions (usually at room temperature and physiological pH). These advantages attract chemists and bio-chemists to investigate enzymatic reactions. Novel chemical reactions are also designed based on the knowledge on enzymatic systems. Hence, enzymatic reactions as well as enzyme-derived reactions widen the roads of organic and medicinal chemistry.

Conventionally, enzymes can be named by their substrates or functions. Given the fact that enzymes are often multi-functional, a nomenclature called EC numbers is introduced by the International Union of Biochemistry and Molecular Biology so that each enzyme gets a single code. The first number in the EC number designates the reaction type, i.e. EC 1 oxidoreductases; EC 2 transferases; EC 3 hydrolases; EC 4 lyases; EC 5 isomerases and EC 6 ligases. The second, third and fourth numbers define more about substrate and product for an enzymatic reaction.

Enzymes are indispensable components to all living organisms, and almost all metabolic pathways that occur in a cell are catalyzed by enzymes. Enzymes have in vitro and in vivo functions. On one hand, an enzyme can catalyze natural or artificial substrates, and such function is in vitro function. On the other, an enzyme together with other enzymes can accomplish more complicated tasks in cell, for example, the metabolism, and such functions are called in vivo function. Understanding the function and collaboration network of enzymes is important for the determination of drug targets.

Enzymatic structures are mainly determined by X-ray crystallography and NMR spectroscopy. Resolved structures are usually stored in the Protein Data Bank (PDB), from which the Cartesian coordinates for each heavy atom in the protein can be obtained. However, X-ray crystallography and NMR spectroscopy are usually time-consuming. Computational methods such as homology
modeling are much more efficient than experimental methods. Applications of computational methods to enzymatic structure and function prediction will be introduced in 1.2.

Enzymatic structures are dynamic, and may undergo large conformational changes during the catalytic process. Static enzyme structures from X-ray crystallography show only conformations (or snapshots) near one local free energy minimum. Although these snapshots record important moments of an enzymatic process, the dynamical information is limited in these snapshots. An increasing number of researchers are interested in the dynamics of enzymes. Computational methods such as MD simulations can be used to investigate the dynamics of enzymes, and will be introduced in 1.3. We will also discuss how enzyme dynamics contribute to the enzymatic reactions.

An enzyme can utilize various catalytic pathways, but only one or two pathways are energetically feasible under physiological conditions. Experimental techniques such as mutagenesis, kinetic isotope effects and X-ray crystallography can provide some hints on the catalytic mechanisms of enzymes. However, reactive intermediates along the catalytic pathway are short-lived and undetectable by current experimental methods. Hence, computational approaches such as QM/MM methods are important for studying catalytic mechanisms of enzymes, which will be discussed in 1.4.

Enzymes are specific, both in selecting substrates and in forming products. Exploring the rationale that how enzymes control substrate and product specificity is useful for further redesigning enzymes to catalyze novel reactions. There are a lot of computational studies on catalytic mechanisms of enzymes, but only a few discuss enzyme specificity. We will give some examples of computational studies on enzyme specificity in 1.5.

Many available drugs are enzyme inhibitors, and computer-aided drug design plays a vital role in drug discovery. Computational tools are powerful for quickly identifying potent inhibitors against an enzyme target, which will be discussed in 1.6.

As we have studied different enzymatic systems in this thesis, we summarize all these systems and their related topics in Table 1 - 1. In this chapter, we will only discuss enzymes in a general
context, and a detailed introduction about the enzymes in Table 1 - 1 is given in Chapter 3 (details are also available in the enclosed manuscripts and published papers).

### Table 1 - 1. Enzymes studied in this thesis

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Structure Prediction</th>
<th>Function Prediction</th>
<th>Dynamics Simulation</th>
<th>Catalytic Mechanism</th>
<th>Specificity</th>
<th>Inhibitor Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>SrtA</td>
<td>Paper 1</td>
<td>Papers 1,2</td>
<td>Paper 2</td>
<td>Papers 3,4</td>
<td>Paper 3</td>
<td>Unpublished</td>
</tr>
<tr>
<td>MHPCO</td>
<td>Papers 3,4</td>
<td>Paper 5</td>
<td>Paper 6</td>
<td>Paper 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSC</td>
<td>Paper 7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 1.2 Enzyme Structure and Function Prediction

The 3D structure of an enzyme provides useful information, such as the active site architecture, interactions between ligand and enzyme, and sometimes even a likely mechanism. More importantly, the 3D structure of a target enzyme is the starting point of structure-based drug design (SBDD). Unfortunately, experimental techniques for 3D structure determination of enzymes are time-consuming, while the number of available gene sequences is steadily increasing. Hence, computational methods for modeling protein structures are widely used, in order to narrow the gap between ‘supply and demand’ (Figure 1 - 1).

There are many different methods for protein structure prediction, for instance, homology modeling (or comparative modeling), fold recognition, knowledge-based *de novo* modeling and *ab initio* modeling from first principles\[^1\]. In this thesis, we will focus on homology modeling as it is so far the most commonly used method for protein structure prediction. Homology modeling relies on two types of input data, i.e. template structure(s) and sequence alignment between query sequence and template sequence(s). The process of homology modeling consists of template searching, sequence alignment, model building, model refinement and evaluation. There are several applications of homology modeling in SBDD, and examples can be found in two recent reviews \[^2, 3\].
Besides SBDD, homology models of enzymes can be used for enzyme function prediction. Enzymes that have high sequence identity (>60%) usually have very similar functions (first three digits or all four digits of an EC number)\textsuperscript{[4]}. However, there are many exceptions, which make functional assignment of enzymes very difficult. An example is the enolase superfamily\textsuperscript{[5]}, where the misannotation of enzymatic function is severe. The key is that the function of an enzyme function is determined by its 3D structure. Hermann et al.\textsuperscript{[6]} presented an approach based on docking a possible metabolite library against the crystal structure of an enzyme and successfully assigned the function of Tm0936. Kalyanaraman et al.\textsuperscript{[7, 8]} further applied both homology modeling and docking to the enolase superfamily, enabling functional prediction of enzymes to be done directly from the gene sequences. These achievements lead to a recent project called enzyme function initiative (EFI)\textsuperscript{[5]}, which calls for a comprehensive collaboration among six scientific cores (Superfamily/Genome, Protein, Structure, Computation, Microbiology, and Data/Dissemination).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Applications of computational chemistry to enzyme structure and function prediction}
\end{figure}
Chapter 1 Introduction

1.3 Enzyme Dynamics Simulations

The combined homology modeling and docking approach can significantly improve the efficiency of functional assignment of enzymes and structure based drug design. However, this method is not perfect and sometimes provides misleading information. This is because 1) the docking and homology modeling methods have limitations; 2) the substrate (or enzyme inhibitor) might not exist in the metabolite (or drug molecule) library; 3) enzymatic structures are dynamic rather than static (see next section).

1.3 Enzyme Dynamics Simulations

Before getting into enzyme dynamics, we need to know the time scales for various motions in an enzymatic reaction (Table 1 - 2\textsuperscript{[9-11]}). Strictly speaking, all the motions listed in Table 1 - 2 can be directly related to ligand binding, enzymatic reaction and product release. Therefore, we have a relatively wide time scale in a single enzymatic reaction. Computational methods such as MD simulations play an important role in studying the dynamics of proteins. So far, MD simulations can achieve millisecond time scale, and simulations of complete protein folding and unfolding processes have been reported\textsuperscript{[12]}. Our knowledge about the dynamic effects in enzyme catalysis is increasing rapidly\textsuperscript{[11, 13-29]}. We will first introduce various models that describe the enzyme catalysis (Figure 1 - 2), and then discuss how the dynamics of an enzyme help catalysis.

Table 1 - 2. Time scales for various motions in an enzymatic reaction\textsuperscript{[9]}

<table>
<thead>
<tr>
<th>Motion</th>
<th>Time Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bond vibrations</td>
<td>10–100 fs</td>
</tr>
<tr>
<td>(TS life time)</td>
<td></td>
</tr>
<tr>
<td>Rotations of side chains</td>
<td>10–100 ps</td>
</tr>
<tr>
<td>Hinge bending at domain interfaces</td>
<td>0.1-10 ns</td>
</tr>
<tr>
<td>Water structure reorganization</td>
<td>10ns</td>
</tr>
<tr>
<td>Rotation of medium-sized side chains</td>
<td>0.001-1ms</td>
</tr>
<tr>
<td>in the protein interior</td>
<td></td>
</tr>
</tbody>
</table>
The lock and key model (Figure 1 - 2a) is probably the earliest model that explains enzyme catalysis, which emphasize that the overall (ensemble averaged) conformation is catalytically active. Two parameters are important for an enzymatic reaction, $K_M$ and $k_{cat}$, which are usually defined as in Equation 1 - 1. They describe how well a ligand binds and how fast the actual chemical reaction proceeds, respectively.
Equation 1 - 1. Definition of $K_M$ and $k_{cat}$

$$K_M = \frac{K_{-1} + K_2}{K_1}$$

$$k_{cat} = K_2$$

The ratio $k_{cat}/K_M$ indicates the enzyme efficiency of a selected ligand. It should be noted that $K_M$ equals the binding affinity only when $K_2 << K_{-1}$ and most of the time $k_{cat}$ and $K_M$ are related (Equation 1 - 1). The induced fit model (Figure 1 - 2b) suggests that the conformation of the enzyme can change when the substrate (or inhibitor) binds. Experimental techniques such as NMR and single molecule fluorescence resonance energy transfer (FRET) suggest that a single enzyme molecule can undergo conformational changes during the reaction and these conformations show very different catalytic efficiency\cite{10,13,30-32}. The existence of more than one catalytically active conformation is no longer debated. We can assume that all the conformations are active, though some are more active than others and the population of various conformations may be changed upon a perturbation such as substrate (or inhibitor) binding (Figure 1 - 2c). This is called the conformational selection with population shift model\cite{15}, i.e. multiple Michaelis complexes $[E_n \cdot S]$ can convert to the product complexes $[E_n \cdot P]$ (Figure 1 - 2c). The lock and key model (Figure 1 - 2a) and the induced fit model (Figure 1 - 2b) can both be treated as special cases of the conformational selection model, in which the enzyme fluctuates around an average conformation with an average catalytic rate and this rate can change when large conformational perturbations occur. An extreme condition in the conformational selection model is that there may exist only a small population (<1%) of very active conformations (states) which have much higher efficiency than the other conformations. One such example is the sortase A enzyme (SrtA), which will be discussed later in the thesis.

The definition of the dynamical effects of an enzymatic reaction is complicated\cite{16}, and is beyond the scope of this thesis. Although some articles in high-level journals\cite{18,19,24,26,28} provided evidence that the dynamical effect might be directly related to the chemical reaction, Warshel et al.\cite{16} pointed out that the evidence is not solid enough to establish a relationship between the dynamical effect and the rate of the reaction step. Our view about enzyme dynamics is that if some dynamic motions can contribute to (or have a connection to) the chemical reaction, they
should occur at similar time scales. Therefore, we suggest that large conformational changes which occur at long time scales (Table 1 - 2) are mainly responsible for substrate binding and product release processes, as illustrated in Figure 1 - 2b and the chemical reaction starts after the substrate binding event. However, if we assume that substrate binding is not a physical process but a chemical one (as in coordination chemistry), and all the three steps, i.e. substrate binding, chemical reaction and product release are chemical steps, then the dynamics of the enzyme can definitely contribute to enzyme catalysis at different stages and time scales, despite that no clear relationship between long and short time scale events has been established. Therefore, we normally consider the enzyme dynamics at different stages separately.

The dynamics of the substrate binding process can be modeled by using steered MD simulations or umbrella sampling simulations with potential of mean force (PMF) calculations on the free energy profile, in which the substrate moves away from the active site according to a predefined pathway (or unbinding pathway)\[^{33}\]. This approach can also be used for studying the product release process. The chemical step can be modeled by defining reaction coordinates, and employing a QM/MM potential. It should be mentioned that when studying enzymatic reactions with computational methods, we have to balance a lot of issues (for instance, QM or MM methods, simulation time and simulation conditions), depending on what information we are after. For example, if the goal is to study conformational changes induced by the substrate binding or product release processes, we can use classical MD simulations (Paper 1, 2). If the goal is to accurately predict the free energy barrier of various steps in an enzymatic reaction whose mechanism is known or relatively simple, we can certainly apply a QM/MM potential or reactive force fields that can describe chemical bond formation or cleavage (such as empirical valence bond theory) to run a free energy MD simulation with a reasonably long simulation time.

However, for most systems, the catalytic mechanism is not at all clear. We usually start with the representative conformation, e.g. the crystal structure, and analyze the catalytic mechanisms with a QM cluster method or single snapshot QM/MM calculations (Papers 2, 3-7), which are computationally more efficient. In fact, in this thesis, most of the results are obtained by using single snapshot QM/MM calculations, though we are aware of its disadvantages.
To date, most of the computational studies start from experimental structures (crystal and NMR structures) or homology models. Hence, we need to know more about our starting point before we go further. The life time of the Michaelis complex is very short and it is thus very difficult to trap. The common experimental method for studying Michaelis complexes is by investigating the structure of the enzyme complexed with substrate mimics (or inhibitors) or the structure of a mutant enzyme complexed with the substrate. However, as shown in Figure 1 - 2c this can lead to an inactive structure when using these pseudo Michaelis complexes, because the population of the active Michaelis conformation might be very small in the native enzyme as well as the pseudo Michaelis complexes. One can get a more reliable structure of the Michaelis complex by using TS analogs (or intermediate analogs) instead of substrate analogs with the wild-type enzyme.

Another issue is how to assign the protonation state of the Michaelis complex. This is very important because most of the simulation studies on enzymes consider only one protonation state. If the protonation state used in the simulation is inactive (see Figure 1 - 2c), the conclusions can be misleading. We suggest that one should at least check the protonation states of the active site residues, i.e. if there are more than one possible protonation state, one should explore the reaction for all of them. We will demonstrate the importance of active site protonation states in Papers 1, 2 and 5.

**1.4 Mechanistic Study on Enzymes**

The previous section discussed enzyme dynamics in the overall enzymatic process. In this section we will focus on the chemical step, in which fascinating reactions with high efficiency occur. Why are enzymatic reactions efficient? Generally speaking, enzymes accelerate chemical reactions by lowering the activation free energies $\Delta G^\neq$ for chemical reactions$^{[34, 35]}$. Two common scenarios are reactant destabilization and transition state (TS) stabilization.

In the first scenario, the reactant could be physically destabilized. Spacial constraints and the electrostatic environment of the enzyme can force the reactant into a high-energy conformation which is closer to the TS, thus reduces $\Delta G^\neq$.

The second scenario, i.e. TS stabilization, is believed to be a common feature of enzymes, although there might be exceptions. Enzymes usually form a pre-optimized environment
(including electrostatics and space constraints) which accommodates and stabilizes the TS by both short-range interactions such as hydrogen-bonding and Van der Waals interactions and long-range electrostatic interactions. It should be noted that we emphasize the ability of an enzyme to reduce the free energy of the TS relative to the reactant rather than the absolute energy of the TS, because in some cases both reactant and TS can be stabilized by the enzyme. We will discuss such a case in the mechanistic study of SrtA (Paper 2). Here we will briefly introduce the generalized transition state theory (TST) and then discuss how to link the computational predictions to experimental data.

TST utilizes two assumptions, the dynamical bottleneck assumption and the equilibrium assumption. The dynamical bottleneck assumption assumes that the reaction rate is controlled by decomposition of the transition-state, and the equilibrium assumption assumes that the transition-state is in equilibrium with the reactant. At temperature $T$, the experimentally determined reaction-rate constant $k(T)$ can be written as in Equation 1 - 2:\[36, 37]:

$$k(T) = \gamma(T)k^{\text{TST}}(T) = \gamma(T) \frac{1}{\beta h} e^{-\beta \Delta G^\ddagger(T)}$$

$$\beta = \frac{1}{k_B T}$$

$$\gamma(T) = \Gamma(T) \kappa(T) g(T)$$

where $k^{\text{TST}}$ is the conventional TST rate constant, $k_B$ is Boltzmann’s constant, $h$ is Planck’s constant, and $\gamma(T)$ is called the transmission coefficient, which has three components: $\Gamma(T)$ (tunneling transmission coefficient), $\kappa(T)$ (recrossing transmission coefficient) and $g(T)$ (nonequilibrium distributions coefficient). Due to the limitation of our computational tools, we did not consider $\gamma(T)$ at all in our studies. If one is interested in these effects, we recommend a recent review article:\[38].

To our knowledge, most of the computational studies on enzyme mechanisms try to predict $\Delta G^\ddagger$ of the rate-limiting step in various pathways, and use the pathway whose rate-limiting $\Delta G^\ddagger$
being lowest as the most likely pathway and compare the $\Delta G^\ddagger$ thereof with experimental data $k_{cat}$ (Figure 1 - 3). It has been shown that conventional TST reasonably well links the computationally predicted $\Delta G^\ddagger$ and the experimental $k(T)$.

![Gibbs Free Energy vs Reaction Coordinates](image)

**Figure 1 - 3. Comparing predicted $\Delta G^\ddagger$ to experimental data**

In this thesis, the QM/MM [ONIOM(B3LYP:Amber)] method is used for most mechanistic studies. This particular combination has been extensively tested and shows good performance\textsuperscript{[39-88]}. Although one may argue that in most cases single point QM/MM methods get good results by error cancellation, we suggest that this method provides a reasonable balance between accuracy and computational cost. More importantly, the results obtained from single point QM/MM calculations can be easily reproduced. We did not use sampling or PMF calculations in this thesis because 1) when MD and sampling are introduced, the results might become more difficult to reproduce; 2) if we think more over Figure 1 - 2, enzymes might not average over all the conformations, i.e. they might use only a small portion (only 1% or less) of the active form to perform the reaction; in this thesis we call it a partially-active model (the traditional enzyme model is defined as the all-active model, Figure 1 - 4).

Assuming that an enzyme has multiple states (different protonation states, or dramatically different conformations) which are connected by physical and chemical equilibrium, and we start
adding substrates; ideally, the substrates can reach all the states of the enzyme, form an ensemble-averaged [E:S] complex and finish the catalysis, as described in Figure 1 - 4a. However, some of the states may be much more efficient than others, and we have the partially-active model (Figure 1 - 4b) in which the concentration of the active form of the enzyme (E_{active}) determines the experimentally observed reaction rate, whereas the inactive form of the enzyme (E_{inactive}) has almost no contribution (we explained this idea in the mechanistic study on sortase A enzyme, Paper 2). It should be noted that the parameter $k_{cat}$ is independent of the concentration of the enzyme. Therefore, as long as we are modeling the active form of the enzyme, the obtained $\Delta G^*$ for the rate-limiting step should be related to the rate constant $k_{cat}$.

![Figure 1 - 4. All-active model versus partially active model for enzyme catalysis](image)

In this chapter, we do not list examples of mechanistic enzyme studies by computational methods, since there are simply too many. If one is interested in such studies, we recommend several review articles\cite{36-38,89-102}. Although the number of computational studies on enzyme catalysis is huge, only a few studies have addressed enzyme specificity, which will be discussed in the next section.
1.5 Computational Study on Enzyme Specificity

Enzymatic reactions have very high specificity, which usually cannot be achieved by common synthetic methodology. Enzyme specificity can be divided into substrate specificity and product specificity. In this thesis, we use the general reaction coordinates for an enzymatic reaction (Figure 1 - 5) to discuss substrate specificity and product specificity.

![Diagram of reaction coordinates for an enzymatic reaction](image)

**Figure 1 - 5. General reaction coordinates for an enzymatic reaction**

The substrate binding and the chemical steps are important for substrate specificity, and $k_{\text{cat}}/k_M$ is used for determination of the best substrate for an enzyme. For most enzymes, $\Delta G_{\text{bind}}^{\ddagger}$ is very low and we can assume that reaction $E+S \rightarrow [E_{\text{active}}\cdot S]$ is in equilibrium. Therefore, the concentration of $[E_{\text{active}}\cdot S]$ is related to the binding affinity $\Delta G_{\text{bind}}$ (note that $\Delta G_{\text{bind}}$ can be positive for some bad binders). However, for some enzymes involving conformational change upon substrate binding, $\Delta G_{\text{bind}}^{\ddagger}$ is large and might be the rate-limiting step for the overall catalytic process. Therefore, to study substrate specificity, one needs to know whether ligand binding is rate-limiting for a specific ligand. Although binding might not be rate-limiting for the original substrate, for other ligands with low binding affinity it might take longer time to bind than to react. Does it mean good binders are good substrates? The answer is no. High binding affinity only guarantees a ligand binds to the enzyme and has the possibility to react. Whether or not a binder really
undergoes the enzymatic reaction depends on the chemical barrier $\Delta G^{\neq}_{\text{chem}}$ (a good binder might be an inhibitor). Therefore, when studying substrate specificity, one has to consider both the substrate-binding and the chemical steps. The prediction of $\Delta G^{\neq}_{\text{chem}}$ has been discussed in 1.4. There are a lot of approaches for the prediction of $\Delta G^{\neq}_{\text{bind}}$, which are summarized in a recent review\cite{103}. $\Delta G^{\neq}_{\text{bind}}$ along one defined pathway or coordinates can be modeled using free energy (PMF) simulations, as mentioned in 1.3.

Product specificity is somehow controlled by substrate specificity, because at least the total number of atoms should be conserved in reactants and products. All three barriers in Figure 1 - 5 may contribute to product specificity. Besides the kinetics, product specificity is also related to the thermodynamic stability of possible products. In this thesis, we discuss product specificity in the following two scenarios: a) Two enzymes with high sequence identity use the same or similar ligands as their substrates, but get rather different products (for example MHPCO versus PHBH, Paper3); b) The same ligand in the same enzyme has the ability to give different products, but only one major product is found (for example OSC, Paper7). To our knowledge, most experimental data does tell us enzymes have product specificity, but does not explain why. Computational approaches can help us to understand product specificity of enzymes, and two good examples are histone lysine methyltransferase $^{[104-107]}$ and squalene cyclase$^{[108]}$. However, one has to consider several factors (all three enzymatic steps might be relevant) in order to reach a convincing conclusion regarding product specificity.

### 1.6 In Silico Design of Enzyme Inhibitors

Computational approaches have become an indispensible tool in today’s drug discovery$^{[109, 110]}$. From target selection to lead generation, and on to lead optimization and ADMET prediction, computational methods have significantly reduced costs in time and money. In this section, lead discovery guided by computational modeling will be discussed. Other computational applications such as lead optimization and ADMET prediction, which are also important for the design of enzyme inhibitors$^{[109-111]}$, will not be covered in this thesis.

Generally speaking, there are two different ways for the discovery of novel lead compounds for a given enzymatic target, i.e. structure based approach$^{[112-116]}$ and pharmacophore based
approach[117-119]. The structure based approach usually starts from the structure of the drug target, and uses docking or scaffold growth method to fill the active site pocket of the target[111]. The pharmacophore based approach usually starts from some known small molecule inhibitors, and uses the 3D molecular features or fingerprints to find molecules that have similar spatial extensions and properties. Note that the pharmacophore based approach does not require the structure of the drug target but if one has the structure of the drug target, the results can be easily refined by using the space constraints from the target structure. The structure based virtual screening (SBVS) approach is the most commonly used method for lead discovery, and will be discussed in this thesis. The pharmacophore based approach will not be discussed here, and one can refer to some recent reviews[117-119] for more information.

SBVS uses two types of input data, i.e. the structure of the target and chemical libraries. SBVS computationally screens chemical libraries by using molecular docking codes, in order to find molecules that have high binding affinities to the target structure. The raw target structure must first be prepared, adding hydrogen atoms, assigning protonation states etc. The chemical libraries used for SBVS usually contain thousands to millions of compounds which are commercially available and satisfy certain criteria such as Lipinski’s rule of five (a rule of thumb for determination of drug likeness). The workflow of SBVS is described in Figure 1 - 6. The output of SBVS is usually a set of molecules that can be potential lead compounds (hits), which will undergo further lead optimization or experimental tests.
A molecular docking code usually consists of a placement method for generating different orientations (also called poses) of a ligand, and a scoring function to rank these. One should bear in mind that all the scoring functions are based on empirical parameters, and that prediction of absolute binding free energies are complicated (that's why in Figure 1 - 6 we emphasize that more ranking steps might be needed). The lead compounds identified by SBVS may undergo experimental tests to confirm their potency. The best ones may be further optimized through structure-activity relationship (SAR) studies with computational or experimental methods.

It should be noted that the use of fragment libraries in SBVS has become more popular in recent years\[112\]. In contrast to the lead-like compound libraries, which usually contain molecules with high molecular weight, the fragment libraries only contain small chemical fragments with low molecular weight (<250 Da). Theoretically, the fragments can be further modified to sample more possibilities in the chemical space. More importantly, fragments are usually easier to synthesize or purchase. Although in the initial SBVS, the hits from fragment libraries usually have lower binding affinity than those from general lead libraries, the fragments have much more
space for improvement. Assuming that the pocket of the target is large, and different hits from fragment libraries occupy different regions of the pocket, one can simply join these hits together with a certain linker to get a better hit.

Before we end this chapter, we want to emphasize that the spirit of the chapter is to demonstrate the power of computational chemistry for enzyme-related systems. There are certainly some other areas not covered by this thesis, for example, artificial enzymes$^{[120-124]}$ (protein design) and allosteric regulation$^{[125-127]}$. With more powerful computers and more user-friendly software interfaces, we believe that computational enzymology will become accessible to all biochemists and enzymologists in the future.
Chapter 2 Methodology

2.1 Overview

In computational chemistry, when we model molecular systems such as enzymes, two of the most important properties are the geometry (\( \mathbf{R} \) is the positions of atoms or nuclei, \( \mathbf{r} \) is the positions of electrons) and the potential energy \( V \). The initial nuclear geometry \( \mathbf{R} \) can be obtained from experimental (e.g. X-ray structure) or computational data, while the potential energy \( V \) is a function of the geometry (\( \mathbf{R}, \mathbf{r} \)).

The Born-Oppenheimer (BO) approximation is used throughout the thesis. The BO approximation allows the total wavefunction \( \Psi \) to be separated into a nuclear wavefunction \( \Psi(\mathbf{R}) \) and an electronic wavefunction \( \Psi(\mathbf{r}) \), with \( \Psi(\mathbf{r}) \) being a function of the nuclear positions \( \mathbf{R} \). Hence, when the nuclear positions \( \mathbf{R} \) change, a potential energy surface (PES) is obtained, and the relationship between \( \mathbf{R} \) and \( V \) is established via the potential energy function \( V(\mathbf{R}) \).

When we introduce molecular mechanics (MM) and quantum mechanics (QM), we are mainly interested in how to obtain \( V \) from \( \mathbf{R} \), i.e. the functional form of \( V(\mathbf{R}) \), because it is the foundation for other simulations such as molecular dynamics and docking. The information regarding the electronic part (\( \mathbf{r} \)) can be obtained from QM calculations but not from MM calculations. \( V(\mathbf{r}) \) is certainly important for studying a system undergoing chemical reaction, such as an enzymatic system. However, complete QM calculations for enzymatic systems are currently impossible. Hence, a hybrid potential QM/MM is usually used for modeling enzymatic reactions.

In 2.2, we discuss the formalism of classical force fields. Other force fields such as polarizable or reactive force field are not included in this thesis. In 2.3, we will briefly introduce basic formalism of Hartree-Fock (HF) theory and density functional theory (DFT). In 2.4, we will introduce the ONIOM formalism for QM/MM calculations, which is used in most of the studies in this thesis. In 2.5, 2.6 and 2.7, we will introduce formalism and algorithms for molecular dynamics (MD) simulation, homology modeling, and docking, respectively.
2.2 Molecular Mechanics

In MM, the smallest particle is an atom, and the interactions of the molecular system are described by a potential function, which is also called a force field. A force field $V_{FF}$ is a function of atomic positions $\mathbf{R}$. $V_{FF}(\mathbf{R})$ consists of two terms, i.e. bonded and non-bonded interactions $[V_{\text{bonded}}(\mathbf{R})$ and $V_{\text{non-bonded}}(\mathbf{R})]$. Bonded interaction has three components, i.e. bond stretching, angle bending and dihedral torsion $[V_{\text{bonds}}(\mathbf{R})$, $V_{\text{angles}}(\mathbf{R})$, and $V_{\text{dihedrals}}(\mathbf{R})$, respectively]. The non-bonded interaction has two components, i.e. electrostatics and van der Waals $[V_{\text{electrostatics}}(\mathbf{R})$ and $V_{\text{VDW}}(\mathbf{R})]$ interactions. Equation 2 - 1 is the most common formula for a classical force field\(^{[128]}\), and the detailed functional form and the parameters vary in different force fields. In Equation 2 - 1, $l$ and $\theta$ are bond lengths and angles ($l_0$ and $\theta_0$ are reference bond lengths and angles), respectively. $\omega$ is the torsion angle; $n$ gives the number of minimum points in the function when the bond rotated through 360º, and $\gamma$ is the phase factor that determines the minimum torsion angle. $q$ refers to atomic charge; $r$ refers to the distance between two atomic centers, and $\varepsilon_0$ is the electric constant. $\varepsilon$ and $\sigma$ are the well depth and the collision diameter.\(^{[128]}\)

![Illustration of various interactions](image)

**Figure 2 - 1. Illustration of various interactions\(^{[128]}\)**
2.3 Quantum Mechanics

Equation 2 - 1.

\[ V_{FF}(\mathbf{R}) = V_{\text{bonded}}(\mathbf{R}) + V_{\text{non-bonded}}(\mathbf{R}) \]

\[ = V_{\text{bonds}}(\mathbf{R}) + V_{\text{angles}}(\mathbf{R}) + V_{\text{dihedrals}}(\mathbf{R}) + V_{\text{electrostatics}}(\mathbf{R}) + V_{\text{VDW}}(\mathbf{R}) \]

\[ = \sum_{\text{bonds}} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2} (1 + \cos(n\omega - \gamma)) \]

\[ + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left( \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} + 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^6 \right] \right) \]

2.3 Quantum Mechanics

2.3.1 The Schrödinger equation

In QM, atoms are represented by nuclei and electrons, and the energy of the system is a function of the positions of nuclei (\(\mathbf{R}\)) and electrons (\(\mathbf{r}\)). The general goal of QM is solving the Schrödinger equation (Equation 2 - 2), where \(\mathbf{H}\) is the energy operator called Hamiltonian, \(\Psi\) is a wavefunction that describes the electronic state of the system, and \(E\) is the total energy of the system. \([129]\)

Equation 2 - 2. Schrödinger equation

\[ \mathbf{H}\Psi(\mathbf{R}, \mathbf{r}) = E\Psi(\mathbf{R}, \mathbf{r}) \]

Applying the BO approximation to the Schrödinger equation, the nuclear positions \(\mathbf{R}\) are deemed to be fixed, and the key is solving the electronic Schrödinger equation (Equation 2 - 3). In Equation 2 - 3, \(T_e\) is the kinetic energy of electrons, \(V_{ne}\) is the nuclear-electron attraction energy, \(V_{ee}\) is the electron-electron repulsion energy, and \(V_{nn}\) is the nuclear-nuclear repulsion energy (a constant).
There are many ways to solve the electronic Schrödinger equation, and only two of the most important approaches, i.e. the Hartree-Fock (HF) theory and the Density Functional Theory (DFT), will be discussed in this thesis.

**Equation 2 - 3. Electronic Schrödinger equation**

\[ H_e \Psi(r) = E_e \Psi(r) \]

\[ H_e = T_e + V_{ne} + V_{ee} + V_{nn} \]

### 2.3.2 Hartree-Fock theory

In the HF theory, the wavefunction \( \Psi(r) \) is written as a Slater determinant (SD), in order to fulfill the antisymmetry requirement. An example SD for an N-electron system is given in Equation 2 - 4. The columns are molecular orbitals (MOs), while the rows are the positions of each electron. An underlying approximation for a single SD is that electron correlation is neglected, which is also called the mean-field approximation. Solving \( \Psi(r) \) is converted to solving the HF equations (Equation 2 - 5).\(^{[129]}\)

**Equation 2 - 4. SD for an N-electron system**

\[
\phi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(1) & \phi_2(1) & \cdots & \phi_N(1) \\ \phi_1(2) & \phi_2(2) & \cdots & \phi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(N) & \phi_2(N) & \cdots & \phi_N(N) \end{vmatrix}
\]

\( F_i \) is called the Fock operator (Equation 2 - 5), which is a one-electron energy operator that describes the kinetic energy and attraction to nuclei (\( h_i \)), as well as interactions with other electrons (\( J \) and \( K \) operators are called Coulomb and exchange operators). \( \lambda_{ij} \) is the Lagrange multipliers.
Chapter 2 Methodology

2.3 Quantum Mechanics

Equation 2 - 5. HF equations

\[ F_i \phi_i = \sum_{j}^{\text{Nelec}} \lambda_{ij} \phi_j \]

\[ F_i = h_i + \sum_{j}^{\text{Nelec}} (J_j - K_j) \]

Each MO can be written as a linear combination of a set of known functions, which are called basis functions or atomic orbitals (AOs). Hence, this approach is also called linear combination of atomic orbitals (LCAO), as written in Equation 2 - 6. According to the variational principle, the best set of MOs should have the lowest energy. Therefore, the HF equations can be solved by minimizing the energy.

Equation 2 - 6. LCAO

\[ \phi_i = \sum_{\alpha}^{M_{\text{basis}}} c_{i\alpha} \chi_{\alpha} \]

For a closed shell system, HF equations are usually represented by the Roothaan-Hall (RH) equations (Equation 2 - 7). \( F \) is the Fock matrix, \( C \) is a matrix of MO coefficients, \( S \) is the overlap matrix of the basis functions, and \( \epsilon \) is a matrix of MO energies. The Self-Consistent Field (SCF) techniques are then used to iteratively solve the RH equations. \(^{129}\)

Equation 2 - 7. RH equations

\[ FC = SC\epsilon \]

2.3.3 Density Functional Theory

DFT is probably the most widely used tool for QM calculations in computational chemistry. Actually, all the QM calculations in this thesis are done by DFT methods such as B3LYP. The
foundation of DFT is the Hohenberg and Kohn theorem, which proves that the ground state electronic energy is determined completely by the electron density. \[^{130}\]

The definition of electron density $\rho$ in an $N$ electron system is described in Equation 2 - 8, where $\mathbf{r}$ and $s$ are spatial and spin variables of electrons, respectively.\[^{130}\] It should be noted that $\rho$ is a function of three spatial variables which integrates to the total number of electrons, and $\rho$ can be measured by experimental techniques such as X-ray diffraction. Generally speaking, $\rho$ is a manifestation of electronic wavefunction. Although we cannot derive the functional form of the wavefunction from $\rho$, $\rho$ itself can provide a lot of useful information such as the ground state electronic energy.

**Equation 2 - 8. Definition of electron density**

$$\rho(\mathbf{r}) = N \sum_{s_1} \cdots \sum_{s_N} \int d\mathbf{r}_2 \cdots \int d\mathbf{r}_N |\Psi(\mathbf{r}, s_1, \mathbf{r}_2, s_2 \cdots \mathbf{r}_N, s_N)|^2$$

$$\int \rho(\mathbf{r}) d\mathbf{r} = N$$

The Kohn-Sham (KS) theory enables DFT to be widely used in practice. In the KS model, a non-interacting reference system, which has exactly the same ground state density as the real system, is introduced. This reference system is also represented by a SD. In order to distinguish the SD in KS and in HF, the symbols $\Theta_{KS}$ and $\varphi$ ($\varphi$ is called a KS orbital) are used in Equation 2 - 9.\[^{130}\]

The $F^{KS}$ operator is a one electron operator in the reference system. $-\frac{1}{2} \nabla^2$ is the kinetic energy operator, and $V_{KS}(\mathbf{r})$ is an effective potential operator (also called KS potential). Once we know $F^{KS}$ as a function of $\rho$, we can solve the KS equations with SCF techniques.

**Equation 2 - 9. An N-electron non-interacting reference system for KS model**

$$\Theta_{KS} = \frac{1}{\sqrt{N!}} \left| \begin{array}{cccc} \varphi_1(1) & \varphi_2(1) & \cdots & \varphi_N(1) \\ \varphi_1(2) & \varphi_2(2) & \cdots & \varphi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(N) & \varphi_2(N) & \cdots & \varphi_N(N) \end{array} \right|$$
Chapter 2 Methodology

2.4 Quantum Mechanics/Molecular Mechanics

Equation 2 - 10. HF equations analog in the KS model

\[ F_i^{KS} \varphi_i = \sum_{j=1}^{N_{elec}} \lambda_{ij} \varphi_j \]

\[ F_i^{KS} = -\frac{1}{2} \nabla^2 + V_{KS}(r) \]

The general energy functional for the KS theory can be written in Equation 2 - 11, where \( T[\rho(r)] \) is the real kinetic energy, \( T_{KS}[\rho(r)] \) is the KS kinetic energy, \( E_{ee}[\rho(r)] \) is the real electron-electron interaction energy, \( J[\rho(r)] \) is the Coulomb interaction energy, \( E_{Ne}[\rho(r)] \) is the nuclear-electron attraction energy, and \( E_{XC}[\rho(r)] \) is the exchange-correlation energy. The trick of the KS theory is that a large unknown term, i.e. the real kinetic energy \( T[\rho(r)] \) and the real electron-electron interaction energy \( E_{ee}[\rho(r)] \), is converted into a small unknown term, i.e. the exchange-correlation energy \( E_{XC}[\rho(r)] \). The \( E_{XC}[\rho(r)] \) term can be constructed by using different models, such as local density approximations (LDA), generalized gradient approximations (GGA) and hybrid functionals.\(^{[130]}\)

Equation 2 - 11. KS functional form for DFT

\[ E[\rho(r)] = T_{KS}[\rho(r)] + J[\rho(r)] + E_{XC}[\rho(r)] + E_{Ne}[\rho(r)] \]

\[ E_{XC}[\rho(r)] = (T[\rho(r)] - T_{KS}[\rho(r)]) + (E_{ee}[\rho(r)] - J[\rho(r)]) \]

2.4 Quantum Mechanics/Molecular Mechanics

The QM/MM approach employs a combined QM/MM potential to treat systems which are too large to be modeled with QM-only potentials. Different QM/MM approaches use different energy expressions (e.g. subtractive or additive schemes), as well as different schemes to treat the covalent QM/MM interface (e.g. link-atom, boundary-atom or localized-orbital schemes). In this thesis, we only introduce the ONIOM approach\(^{[131]}\), and comprehensive introduction of the other approaches such as localized-orbital approach can be found elsewhere\(^{[89, 91, 97, 132]}\).

Equation 2 - 12. ONIOM energy expression
Chapter 2 Methodology

2.4 Quantum Mechanics/Molecular Mechanics

\[ E_{ONIOM} = E^{MM}[\text{Real}] + E^{QM}[\text{Model}] - E^{MM}[\text{Model}] \]

The ONIOM approach introduces a hydrogen link-atom to treat the QM/MM interface (Figure 2-2), and the two-layer energy expression of ONIOM is given in Equation 2-12. We consider an example ONIOM calculation of an alanine molecule, whose side-chain is modeled by QM. In the ONIOM calculations, the QM region (called model system, Figure 2-2) becomes a methane, and the QM/MM energy of the alanine is \( E^{MM[\text{alanine}]} + E^{QM[\text{methane}]} - E^{MM[\text{methane}]} \).

![Figure 2-2. An example ONIOM model system](image.png)

Another issue is how a QM/MM method describes the electrostatic interaction between the MM and the QM region. Two common schemes are mechanical embedding (ME) and electrostatic embedding (also called electronic embedding, EE). In the ME scheme, MM-QM electrostatics is the same as MM-MM electrostatics; while in the EE scheme, MM atomic charges are incorporated into the QM Hamiltonian (Equation 2-13). In Equation 2-13, \( N \), \( J \), and \( i \) are the atoms from the MM region, atoms from the QM region, and electrons; \( s \) is a scaling factor, which equals zero for charges less than three bonds away from the QM region; \( Z \) and \( q \) are charges in the MM and QM regions, respectively.

\[ H_{EE}^{QM} = H^{QM} - \sum_i \sum_N s_N q_N r_{iN} + \sum_j \sum_N \frac{Z_J s_N q_N}{r_{jN}} \]
2.5 Molecular dynamics simulations

An MD simulation generates a series of time-correlated snapshots of a system, in order to obtain an ensemble with statistical meaning. MD simulation techniques can be applied to many areas such as biochemistry and material science.

Classical MD simulations use Newton’s second equation to generate trajectories (Figure 2 - 3). To perform a classical MD simulation, initial positions $\mathbf{R}_0$ and velocities $\mathbf{v}_0$ as well as a time step $\Delta t$ (normally 1-2 fs) must be provided. $\mathbf{R}_0$ can be obtained from X-ray or NMR structures, and $\mathbf{v}_0$ values are usually generated according to atom types and simulation temperature. A potential function $V(\mathbf{R})$ is further used to evaluate the force on each atom, in order to get accelerations $\mathbf{a}$. New positions $\mathbf{R}_{n+1}$ can then be obtained from $\mathbf{R}_n$ repeatedly (Figure 2 - 3). There are integration algorithms to solve the Newton’s equations of motion, and the expressions of two widely-used ones, i.e. the leapfrog algorithm and the velocity Verlet algorithm, are given in Equation 2 - 14 and Equation 2 - 15, respectively.

![Figure 2 - 3. MD workflow](image)
Equation 2 - 14. Leapfrog algorithm

\[ R(t + \Delta t) = R(t) + \Delta t \cdot v(t + \frac{1}{2} \Delta t) \]

\[ v(t + \frac{1}{2} \Delta t) = v(t - \frac{1}{2} \Delta t) + \Delta t \cdot a(t) \]

Equation 2 - 15. Velocity Verlet algorithm

\[ R(t + \Delta t) = R(t) + \Delta t \cdot v(t) + \frac{1}{2} \Delta t^2 \cdot a(t) \]

\[ v(t + \Delta t) = v(t) + \frac{1}{2} \Delta t \cdot [a(t) + a(t + \Delta t)] \]

When a sufficiently long trajectory is generated, various properties of the system can be analyzed. However, for simulations of large systems (e.g. enzymes), the desired time-scale is difficult to reach. The root-mean-square deviation (RMSD) values of the C-\(\alpha\) carbon of the protein are sometimes used to determine whether a protein system is fully equilibrated, but it is not rigorous. To our knowledge, there is no rigorous way to decide whether a simulation samples sufficient points in the phase space, although there are other techniques (e.g. replica exchange MD) that enable us to enhance sampling.

2.6 Homology Modeling

Homology modeling method is a way to construct the 3D structure of a target protein using the gene sequence of the target protein and the known structures of one or more related homologous proteins (called templates). Generally speaking, the accuracy of the homology model depends on the sequence identity between target protein and templates, as well as sequence alignment. Multiple sequence alignment is usually used, as aligning the target protein to its family members can preserve the common features (conserved residues) of the family.

Given templates and alignments, a homology model can be constructed and refined (Figure 2 - 4). Model building involves several steps, such as initial geometry specification, insertions and
deletions of amino acids, loop selections, sidechain packing, final model selection and refinement.\textsuperscript{[133]} The aligned region of the model structure is usually conserved from the template structure during the initial model construction, and different homology model codes use different algorithms to predict unaligned region and refine the final structure.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{homology_modeling_example.png}
\caption{An example of homology modeling}
\end{figure}

\section*{2.7 Docking and Virtual Screening}

Molecular docking is used for predicting and ranking the orientations of a ligand, in the pocket of a receptor (e.g. enzyme active site). A molecular docking code usually consists of a placement method for generating different orientations (also called poses) of a ligand, and a scoring function to rank them.\textsuperscript{[128]} Shape-complementarity methods are the most widely used for pose generation. In shape-complementarity methods, the shapes of the ligands and receptors are described by a
number of descriptors such as solvent-accessible surface area, hydrogen bond donor or acceptor etc.\textsuperscript{[128]} Poses are then generated according to complementarity between the descriptors. Docking poses are usually ranked by scoring functions, which estimate the binding free energies of the ligands. Additional binding affinity calculations with the other methods are also frequently used to obtain more accurate results, e.g. Generalized Born/Surface Area (GB/SA), Poisson-Boltzmann/Surface Area (PB/SA) and Generalized Born/Volume Integral (GB/VI) models.

In the GB/SA model, the solvation free energy ($G_{sol}$) is a sum of a solvent-solvent cavity term ($G_{cav}$), a solute-solvent van der Waals term ($G_{vdw}$), and a solute-solvent electrostatic polarization term ($G_{pol}$). $G_{cav} + G_{vdw}$ together is approximately linearly related to their solvent accessible surface areas (SA), and $G_{pol}$ is obtained by solving the generalized Born (GB) equation (or GB with modifications, Equation 2 - 16).\textsuperscript{[134]} In Equation 2 - 16, $SA_k$ ($\AA^2$) is the total solvent-accessible surface area of all atoms of type $k$, $\sigma$ (kcal/(mol $\AA^2$)) is an empirically determined atomic solvation parameter, $\alpha_i$ ($\AA$) is the Born radius of atom $i$, and $D_{ij}$ is the squared ratio of the $i,j$th atom pair separation to their mean Born diameters.\textsuperscript{[134]} PB/SA\textsuperscript{[135]} and GB/VI\textsuperscript{[136]} models evaluate $G_{sol}$ with different functional forms, but the essential ideas for these three models are very similar, where the key is to get a good model for taking the solvent effect into account.

**Equation 2 - 16. The GB/SA model**\textsuperscript{[134]}

\[
G_{sol} = G_{cav} + G_{vdw} + G_{pol}
\]

\[
G_{cav} + G_{vdw} = \sum_{k=1}^{N} \sigma_k SA_k
\]

\[
G_{pol} = -166.0(1 - \frac{1}{\epsilon}) \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{q_i q_j}{(r_{ij}^2 + \alpha_i^2 e^{-D_{ij}})^{0.5}}
\]

Virtual screening is a multiple-ligand docking process, and the aim is to select good binders from chemical libraries which contain thousands to millions of compounds. An example library is the ZINC database\textsuperscript{[137]}. In virtual screening calculations, multiple docking and re-scoring might be used to select the best binders step by step. For example, in the Schrödinger software, there are
three different levels of accuracy for virtual screening, i.e. high throughput virtual screening (HTVS), standard precision (SP) and extra precision (XP).
Chapter 3 Result and Discussion

3.1 Computational Study on SrtA (Papers 1, 2)

3.1.1 Introduction

In Gram-positive pathogenic bacteria, the cell wall anchored surface proteins, for example, the LPXTG (leucine, proline, X, threonine, and glycine, where X is any amino acid) protein, play vital roles in pathogenicity\[^{138}\]. The LPXTG protein has an LPXTG motif near the carboxyl terminal followed by a hydrophobic region and a positively charged tail\[^{139}\]. The cell wall anchored LPXTG proteins are capable of adhering to and invading the host cells\[^{140, 141}\]. The sortase A (SrtA) enzyme catalyzes the cell wall anchoring reaction of LPXTG proteins\[^{142-144}\]. SrtA cleaves the peptide bond between the threonine and glycine (T-G) residues in the LPXTG motif and links the threonine to the amino group of the pentaglycine cell wall crossbridge, thus forming the cell wall attached protein (Figure 3 - 1)\[^{144}\].

Figure 3 - 1. Cell wall anchoring reaction of LPXTG surface proteins catalyzed by SrtA (Paper 1)
It has been shown that inactivation of SrtA significantly reduces the virulence of the pathogens\cite{145, 146}. In addition, SrtA is not required for the cell growth, and SrtA inhibitors will unlikely lead to development of drug resistance\cite{146}. Therefore, SrtA has been proposed to be a universal target for therapeutic agents against Gram-positive bacteria\cite{143, 144, 147}.

In this thesis, SrtA enzymes in two Gram-positive pathogens, i.e. \textit{Listeria monocytogenes} (\textit{L. monocytogenes}) and \textit{Staphylococcus aureus} (\textit{S. aureus}), are studied by using computational methods. \textit{L. monocytogenes} is a virulent foodborne pathogen, and infection by which causes listeriosis, with clinical manifestations including septicemia, meningitis, etc.\cite{148, 149} \textit{S. aureus}, especially Methicillin-resistant \textit{S. aureus}, is a major health problem, which causes tens of thousands of deaths each year.\cite{143} Our primary goal is to obtain structural and mechanistic information of SrtA, and the long-term goal is to design SrtA inhibitors, which could be potential drugs against Gram-positive pathogens such as \textit{L. monocytogenes} and \textit{S. aureus}.

The knowledge about the structure of the SrtA enzyme-substrate complex is limited, and two experimental structures are of particular interest, i.e. a crystal structure of \textit{S. aureus} SrtA (C184A mutant) with a bound substrate\cite{150}, and a recent NMR structure of \textit{S. aureus} SrtA\cite{151}, in which the modified tetrapeptide LPAT* was attached to the active site cysteine through a disulfide bond, mimicking the short lived LPXT-enzyme intermediate.\cite{151} In the X-ray structure His120 is far away from Cys184, whereas in the NMR structure, the Cys184 and His120 residues are close to each other. Mutagenesis\cite{151-153} demonstrated that H120A mutation results in a 10000-fold decrease in enzyme activity, suggesting that His120 is even more important than Cys184 (1000-fold decrease when in the C184A mutant). Hence, the X-ray structure is unable to explain the role of His120. In addition, the X-ray structure of the \textit{S. aureus} SrtA-LPXTG complex is significantly different from the structures of other SrtA homologues.\cite{151}

SrtA uses a ping-pong mechanism\cite{154, 155}. SrtA first cleaves the T-G peptide bond, forming an LPXT-enzyme intermediate. The second substrate pentaglycine (from the outer cell wall cross bridge) is then attached to the LPXT moiety (Figure 3 - 2, \textit{S. aureus} numbering). According to mutagenesis studies\cite{152, 153}, Cys184, His120, Arg197, and Thr183 are indispensable, as alanine substitution of any of these four residues gives inactive SrtA. The carbonyl carbon of the T-G peptide bond in the substrate is attacked by the Cys184 thiolate anion, and the leaving NH group
is protonated by the His120 imidazolium cation. Arg197 is considered to be important for substrate binding via hydrogen bonds. The role of Thr183 is currently unknown.

Figure 3-2. The ping-pong mechanism of SrtA (Paper 2)

Although several catalytic mechanisms were proposed, only the reverse protonation mechanism (RPM) is consistent with all the experimental observations (e.g. pKa measurements, pH dependence experiments, and solvent isotope effect measurements). In RPM, a small fraction of the active SrtA is competent for catalysis; Cys184 and His120 exist in the thiolate anion and imidazolium cation forms (Figure 3-2) before the substrate is bound, and the ionization state of Cys184 is independent of His120 protonation. Unfortunately, RPM is difficult to be proved due to the sparse population (less than 1%) of the active form of SrtA. Moreover, it is difficult to obtain the structure of the active enzyme-substrate complex by using either X-ray diffraction or NMR techniques. To this end, computational methods are key tools to characterize the structure of the active SrtA-LPXTG complex and study the catalytic mechanism thereof.

3.1.2 L. monocytogenes SrtA homology model (Paper 1)

As structure of L.monocytogenes SrtA is not available, homology modeling is used to model the catalytic domain thereof (From Tyr 75 to Lys 222). When choosing a template, the sequence identity is the most important criterion (a higher sequence identity template gives a more reliable
model structure). *Streptococcus pyogenes* (*S. pyogenes*) SrtA\(^{[157]}\) was chosen as the template to model *L. monocytogenes* SrtA, based on the multiple sequence alignment (Figure 3 - 3). The catalytic domain sequence identity between *L. monocytogenes* and *S. pyogenes* (47.6%) is much higher than that between *L. monocytogenes* and *S. aureus* (26.9%). Therefore, the *S. aureus* SrtA structure is not used for homology modeling, though it is experimentally the most studied SrtA enzyme. The obtained *L. monocytogenes* SrtA homology model superposes well with the template (Ca-RMSD 1.20 Å), and the most significant difference is the β4/H2 loop, which is consistent with the fact that the sequence identity is quite low in this region (Figure 3 - 3). The substrate pentapeptide (LPTTG in *L. monocytogenes*) is then built, optimized and docked into the active site of the *L. monocytogenes* SrtA homology model. The obtained SrtA-LPXTG complex is then equilibrated by MD simulations, where interesting results are observed.
3.1.3 Arg197 Motion in *L. monocytogenes* SrtA Homology Model (Paper 1)

30ns MD simulations are performed on the inactive and active forms (active and inactive forms refer to the zwitterionic and neutral forms of the Cys184-His120 catalytic dyad) of the *L. monocytogenes* SrtA-LPTTG complexes. Computational details can be found in Appendix. Initial structure of the active form is obtained by manually modifying the protonation state of the last snapshot from 30ns trajectory of the inactive form. The root mean square fluctuation (RMSF) of the MD simulation trajectory indicates the flexibility of the residues (Figure 3 - 4). For the inactive form, during the 30ns MD simulation, the LP residues in LPTTG motif maintain their positions, whereas the TTG part is very flexible, as no specific active site residue holds the TTG part in position (Figure 3 - 4). For the active form, after 30ns simulation, Arg197 switches its two hydrogen bonds from the LP to the TG residues. As a result, the flexibility of the TTG part is reduced (Figure 3 - 4), since Arg197 interacts with this part in most of the simulation.

![Residue RMSF Graph](image)

**Figure 3 - 4.** Comparison of the residue RMSF of the LPTTG substrates in the inactive and active SrtA-LPTTG complexes (Paper 1)
To monitor the Arg197 motion, the distance between Arg197 and His127, is plotted as a function of time (Figure 3 - 5). The Arg197 motion is divided into three stages (Figure 3 - 5): Stage 1, the guanidinium cation donates two hydrogen bonds to the LP backbone carbonyl oxygen atoms, which is equivalent to the conformational orientation in the inactive form; Stage 2, the guanidinium cation shifts its hydrogen bonding to the carbonyls of the TT part; Stage 3, the hydrogen bonds have switched to the TG residue carbonyls. In the Arg197 motion, a rotation of the sidechain guanidinium cation is observed. The most likely reason Arg197 motion is charge–charge interaction, i.e. with Cys188 being deprotonated, the thiolate anion attracts the guanidinium cation, and pulls it towards the Cys-His dyad.

![Diagram showing the movement of Arg197 and His127](image)

**Figure 3 - 5.** Distance between Arg197-His127 (center of mass) changes as function of simulation time for the active SrtA-LPTTG complex (Paper 1)
Figure 3 - 6. Superposed structures of the inactive (in red) and active (in blue) SrtA-LPTTG complexes after minimization (Paper 1)

Given the Arg197 motion, the following hypotheses about the role of Arg197 have been made: (1) Arg197 alters the relative stabilities of the inactive and active forms of the Cys184-His120 catalytic dyad via electrostatic interactions; (2) Arg197 affects substrate binding affinity; (3) Arg197 stabilizes the transition state or the tetrahedral intermediate. In the MD simulations of *L. monocytogenes* SrtA, the initial structure is a computational model rather than experimental data. Hence, to further prove our hypotheses, systematic simulations as well as higher level QM/MM calculations have been performed using the NMR structure of *S. aureus* SrtA\(^{151}\).
3.1.4 Arg197 Motion in *S. aureus* SrtA (Paper 2)

20 ns MD simulations are performed using the active form of the *S. aureus* SrtA-LPATG complex, followed by another 20 ns MD simulations with the Cys-His dyad switched to the inactive form (computational details see Appendix). The obtained results are consistent with that for the simulations of *L. monocytogenes* SrtA homology model. In the current simulation, Arg197 forms hydrogen bonds to the carbonyl oxygen of the TG peptide bond in the active form, and swings over to form hydrogen bonds to the LP part in the inactive form. It should be noted that Position1 and Position2 refer to the preferred positions of Arg197 in the active and inactive SrtA-substrate complexes.

Figure 3 - 7. Hydrogen bonds between Arg197 and substrate in the 20 ns MD simulation of the inactive SrtA-LPXTG complex, starting from the simulated active complex. Position1 and Position2 refer to the preferred positions of Arg197 in the active and inactive SrtA-substrate complexes (Paper 2)
The hydrogen-bonding interactions between Arg197 and the substrate in Position1 and Position2 are shown in Figure 3 - 7, and a superposed view is shown in Figure 3 - 8. In Position1, the side chain of Arg197 forms one direct hydrogen bond with the carbonyl group of the TG peptide bond (H11···OThr, Figure 3 - 7). In the 20ns simulation of the inactive form of the SrtA-LPXTG complex, the H11···OThr hydrogen bond is broken spontaneously, while two new hydrogen bonds are formed (H21···OLeu and HE···OLeu, respectively, Figure 3 - 7). The repositioning of Arg197 may be essential for the SrtA catalysis, as mentioned in the hypotheses above. To elucidate the role of Arg197, different systems were therefore investigated by means of QM/MM (ONIOM) calculations.

---

**Figure 3 - 8.** Superposed structures with Arg197 in Position1 (in red) and Position2 (in blue) (Paper 2)

### 3.1.5 Relative stability of the inactive and active forms (Paper 2)

The relative stability of the inactive and active forms is important for the SrtA enzyme, as pKa measurements\textsuperscript{[154, 156]} indicate that the majority of the enzyme exists in the inactive form. In this
respect, SrtA differs from the cysteine proteases such as papain\[^{158}\]. However, the bell-shaped pH dependencies of \(k_{\text{cat}}/K_M\) imply that one residue in the Cys-His dyad must be protonated and the other deprotonated for activity\[^{154}\]. Mladenovic et al.\[^{159}\] have shown that cysteine proteases are capable of altering the relative stabilities of the neutral versus zwitterionic forms of the Cys-His dyad, which explained the abnormal pKa values thereof.\[^{160,161}\] In order to validate the hypothesis that the Arg197 motion is capable of altering the relative stabilities of the inactive form versus the active form in SrtA, systematic QM/MM (ONIOM) calculations (Figure 3 - 9; see Appendix for computational details) are performed on the transformations between the inactive and active forms in Position1, Position1-R197A (Position1 with Arg197 mutated to alanine), and Position2 (Arg197 located far away from the Cys-His dyad).

![Figure 3 - 9. QM/MM partitioning scheme: Model-1 in blue, Model-2 blue + red and Model-3 blue + red + green (Paper 2)](image)

We first studied the relative stabilities of the inactive and active forms with and without the substrate bound. According to five MD simulations in Position1, the active form is more stable than the inactive form (Table 3 - 1, entry 1 and 4) by 24-30 kcal/mol. Given the small difference between the five simulation snapshots (Table 3 - 1, entry 1), only one starting structure was used in the Position2 and Position1-R197A systems. According to our results, the inactive form becomes more stable than the active form in either Position1-R197A or Position 2. In addition,
the active form becomes less stable with the substrate bound (Table 3-1). The rationale is that the hydrophobic substrate prevents solvent molecules from entering the enzyme pocket and that the Cys-His ion pair is destabilized in the resulting nonpolar environment. With the LPXTG bound to the inactive enzyme, the proton transfer from Cys184 to His120 becomes more endothermic (Table 3-1, entries 3 and 6), unless the Arg197 side chain moves to the TG region. Therefore, we suggest that the general base mechanism, where the substrate binds to the inactive form, is unlikely to occur. Our calculations therefore favor the RPM\textsuperscript{[154]}; i.e., the LPXTG substrate directly binds to the active form of the enzyme and undergoes acylation, whilst the majority of the enzyme is in the inactive form.

### Table 3-1. Proton States of the Cys-His Dyad under Different Conditions (Paper 2)

<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>Relative Energies$^a$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Active</td>
<td>Inactive</td>
</tr>
<tr>
<td>1</td>
<td>Position1 without LPXTG</td>
<td>0.0</td>
<td>27.4 ± 3.0$^b$</td>
</tr>
<tr>
<td>2</td>
<td>Position1-R197A without LPXTG</td>
<td>0.0</td>
<td>-10.8</td>
</tr>
<tr>
<td>3</td>
<td>Position2 without LPXTG</td>
<td>0.0</td>
<td>-8.2</td>
</tr>
<tr>
<td>4</td>
<td>Position1 with LPXTG</td>
<td>0.0</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>Position1-R197A with LPXTG</td>
<td>0.0</td>
<td>-15.2</td>
</tr>
<tr>
<td>6</td>
<td>Position2 with LPXTG</td>
<td>0.0</td>
<td>-14.1</td>
</tr>
</tbody>
</table>

$^a$ Energy of inactive form relative to active form was used for each case. Model-3 was used.

$^b$ Average energy of five snapshots, with error bar.

### 3.1.6 Substrate binding affinity (Paper 2)

The binding affinity of the substrate was evaluated using the MM/GBVI implicit solvent approach\textsuperscript{[162]} in the MOE software\textsuperscript{[133]}. The average value of binding affinities from 10 snapshots (with all solvent molecules removed) in the last 1 ns MD trajectory was evaluated (Table 3-2). The hydrogen bonding between Arg197 and LPXTG plays an important role in substrate binding,
as the R197A mutation results in a 4-6 kcal/mol decrease in binding affinity (Table 3 - 2 entry 1 vs entry 3, and entry 2 vs entry 4). Therefore, the hypothesis that Arg197 affects substrate binding affinity has been validated. However, the binding affinity is not significantly affected by the Arg197 motion (entry 1 vs entry 5, and entry 2 vs entry 6), suggesting that the substrate does not selectively bind to the active form of the enzyme.

**Table 3 - 2. Substrate Binding Affinity under different conditions (Paper 2)**

<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>MM/GBVI Binding Affinity$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Position1 active</td>
<td>-55.8 ± 0.8</td>
</tr>
<tr>
<td>2</td>
<td>Position1 inactive</td>
<td>-54.9 ± 0.6</td>
</tr>
<tr>
<td>3</td>
<td>Position1-R197A active</td>
<td>-49.6 ± 0.9</td>
</tr>
<tr>
<td>4</td>
<td>Position1-R197A inactive</td>
<td>-50.9 ± 1.2</td>
</tr>
<tr>
<td>5</td>
<td>Position2 active</td>
<td>-55.0 ± 0.5</td>
</tr>
<tr>
<td>6</td>
<td>Position2 inactive</td>
<td>-54.0 ± 0.8</td>
</tr>
</tbody>
</table>

$^a$ in kcal/mol

**3.1.7 Acylation mechanism of SrtA: roles of Arg197 and Thr183 (Paper 2)**

The acylation of cysteine proteases have previously been modeled in a number of theoretical studies\(^{163-165}\). Arad et al.\(^{165}\) suggested that a proton transfer from histidine (PT-His) occurs prior to, or concerted with, the nucleophilic attack by the thiolate anion (NA-Cys). The papain system was further studied by Harrison et al.\(^{164}\) with the AM1/MM and B3LYP/MM methods, and the concerted pathway without the involvement of a tetrahedral intermediate (TI), was suggested. Recently, Ma et al.\(^{163}\) determined the two-dimensional free energy surfaces of the acylation of human cathepsin by using AM1/MM molecular dynamics simulations. The PT-His and NA-Cys reactions were found highly coupled, and a TI was formed along the NA-Cys pathway.\(^{163}\)
To explore the third hypothesis, i.e. whether the oxyanion hole (Position1) generated by Arg197 stabilizes the TI in the acylation step, QM/MM calculations were used to model the acylation step in the active form of SrtA in Position1. As we have demonstrated that the Cys-His ion-pair, which is the reactant of the acylation step, is significantly stabilized by Arg197, the barrier height of acylation may in fact increase due to the stabilization of the reactant. The deacylation step was not modeled, as the deacylation is highly similar to the reverse process of the current acylation.

Initial scans of two reaction coordinates, i.e. NA-Cys (SCys184···CThr distance) and PT-His (HEHis120···NGly distance), suggest that PT-His process is the major route to reach the TS, and NA-Cys may occur concertedly with or immediately after PT-His. Results for the TS are summarized in Table 3 - 3. The obtained barrier heights are in reasonable accordance with the experimental value ~20 kcal/mol[154].

Table 3 - 3. Properties of the TS in Position1 for different ONIOM schemes (Paper 2)

<table>
<thead>
<tr>
<th>ONIOM Models</th>
<th>Barrier Height(^a)</th>
<th>Imaginary(^b) Frequency</th>
<th>SCys184···CThr distance in TS(^c)</th>
<th>HEHis120···NGly distance in TS(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-1</td>
<td>19.4</td>
<td>-730.4</td>
<td>2.880</td>
<td>1.412</td>
</tr>
<tr>
<td>Model-2</td>
<td>14.8</td>
<td>-700.2</td>
<td>2.907</td>
<td>1.573</td>
</tr>
<tr>
<td>Model-3</td>
<td>17.5</td>
<td>-536.8</td>
<td>3.350</td>
<td>1.190</td>
</tr>
</tbody>
</table>

\(^a\) in kcal/mol  
\(^b\) in i cm\(^{-1}\)  
\(^c\) in Angstrom

To further investigate the role of Arg197 and Thr183 on the acylation step, we modeled this for Position1-R197A, Position1-T183A and Position2 (Table 3 - 4). According to our calculations, R197A mutation significantly lowers the acylation barrier (from 19.4 to 11.7 kcal/mol) whereas the T183A mutation works in the opposite manner (from 19.4 to 28.8 kcal/mol). Experimentally, both R197A and T183A mutations results in ~1000 fold decrease in enzyme activity, and the roles of these two residues are still somewhat controversial[151]. The acylation barrier in Position2 is lower than that of Position1, implying that the Arg197 motion activates the Cys-His dyad but
at the same time raises the acylation barrier. Hence, Arg197 itself gives rise to an increased barrier for the acylation step. To compensate for this, the presence of Thr183 affects the catalysis by lowering the acylation barrier by roughly the same amount via electrostatic interactions.

**Table 3 - 4. Comparison of Acylation Barriers at Different States (Paper 2)**

<table>
<thead>
<tr>
<th>Systems</th>
<th>Acylation Barrier$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position1</td>
<td>19.4</td>
</tr>
<tr>
<td>Position2</td>
<td>16.8</td>
</tr>
<tr>
<td>Position1-R197A</td>
<td>11.7</td>
</tr>
<tr>
<td>Position1-T183A</td>
<td>28.2</td>
</tr>
</tbody>
</table>

$^a$ Model-1 was used, in kcal/mol

### 3.1.8 Inhibitors against SrtA (unpublished)

To date, several inhibitors against the *S. aureus* SrtA have been reported, including non-specific sulfhydryl modifiers, natural inhibitors, peptide analogues, mechanism-based inhibitors, as well as synthetic small molecules identified by high-throughput screening$^{[166-170]}$. Unfortunately, none of these inhibitors present sufficient efficiency to be therapeutically useful. Several representative inhibitors with their IC$_{50}$ values were summarized in Figure 3 - 10.
3.1 Computational Study on SrtA (Papers 1, 2)

Figure 3 - 10. Representative inhibitors against *S. aureus* SrtA\(^{166-170}\)

SBVS calculations are performed using the NMR structure of *S. aureus* SrtA\(^{151}\) and the ZINC database\(^{134}\). Top 100 hit molecules for each subset were identified (unpublished results). Further experimental validation of these hit molecules is needed in order to select the best candidates for structural optimization (lead optimization). The results are not appropriate to be shown due to patent reasons.
3.2 Computational Study on MHPCO (Papers 3, 4)

3.2.1 Introduction

Flavoprotein monoxygenases (FMOs) catalyze various oxidative reactions.\[171\] Within the FMO family, the aromatic hydroxylases have been extensively studied. The aromatic hydroxylase enzymes hydroxylate aromatic rings at the ortho position of phenolic substrates via electrophilic aromatic substitution reactions.\[172, 173\] MHPCO is a unique member of this aromatic hydroxylase family because the reactions catalyzed by the enzyme give aliphatic products rather than aromatic ones.\[174-179\] Three compounds have been shown to be substrates of MHPCO (Figure 3 - 11), namely, 2-methyl-3-hydroxypyridine-5-carboxylic acid (MHPC), 5-hydroxynicotinic acid (5HN) and N-methyl-5-hydroxynicotinic acid (NMHN).\[180, 181\] The conversion of MHPC to α-(N-acetylaminomethylene)succinic acid (AAMS) is the key step in the bacterial degradation of vitamin B\(_6\).\[182\]

Figure 3 - 11. Reactions catalyzed by MHPCO (Paper 3)

The monoxygenation process catalyzed by aromatic hydroxylases can be divided into two stages: reduction and oxidation (Figure 3 - 12).\[173\] In the reduction stage, the FAD cofactor is reduced by NADPH, yielding FADH\(^{-}\). In the oxidation stage, FADH\(^{-}\) reacts with molecular oxygen to form C-(4a)-flavinhydroperoxide (FADHOOH), which is a short-lived intermediate.\[183\] FADHOOH then acts as an electrophile in the oxygenation, and the terminal hydroxyl group is transferred from FADHOOH to the substrate through electrophilic aromatic substitution.\[173\]
Catalytic mechanisms of MHPCO were previously proposed by Chaiyen et al.\textsuperscript{180,181} and McCulloch et al.\textsuperscript{172}. These mechanisms share the similarity that hydroxylation and ring-opening are the two key catalytic steps. As with other aromatic hydroxylases, e.g. p-hydroxybenzoate hydroxylase (PHBH), the substrate is first hydroxylated by FADHOOH through an electrophilic aromatic substitution, which is the rate-limiting step (Figure 3 - 13). In the ring-opening step, a water molecule is added to the hydroxylated intermediate either before or after the cleavage of the C2-C3 bond (Figure 3 - 13).
The hydroxylation step of PHBH has been modeled previously with both QM\cite{184,185} and QM/MM\cite{186-190} methods. Bach et al.\cite{184} studied the mechanism of FADHOOH oxidation of a series of heteroatom nucleophiles with several QM methods which provided insight on how to distinguish between oxygen atom transfer and hydroxyl transfer mechanisms. The hydroxylation mechanism of PHBH was then modeled with inclusion of two active site residues, Arg214 and Tyr201 (in MHPCO, the corresponding residues are Arg211 and Tyr223), which were shown to play a minor role in lowering the reaction barrier.\cite{185} Ridder et al. studied PHBH\cite{186,188} and phenol hydroxylase\cite{187} using QM/MM calculations, suggesting that the amide oxygen of the active site proline (Pro293 in PHBH, Pro364 in phenol hydroxylase, and Pro295 in MHPCO) can stabilize the TS by 2-3 kcal/mol\cite{186}. Senn et al.\cite{190} investigated the hydroxylation of PHBH by using QM/MM molecular dynamics simulations and analyzed the hydrogen bonding network in the active site along the reaction coordinate of OH transfer. They found that the hydrogen bond between the Pro293 amide oxygen and the proton in the transferred hydroxyl varied substantially, making the role of the Pro293 backbone carbonyl inconclusive. Mata et al.\cite{189} performed more accurate QM/MM calculations on the PHBH system using higher level QM methods such as LMP2 and LCCSD. A barrier height of ~14 kcal/mol was obtained, which is consistent with the experimental value.

The catalytic mechanism of MHPCO has been systematically studied in Paper 3 and Paper 4. In Paper 3, DFT calculations are performed on a small model, containing only FADHOOH and a
substrate 5HN (Paper 3). Previously proposed pathways by Chaiyen\textsuperscript{[180]} and McCulloch\textsuperscript{[172]} are shown unlikely to occur. Two new pathways for ring-opening are presented based on theoretical considerations, and the unusual product specificity of MHPCO is explained. Based on the data from Paper 3, extensive DFT and ONIOM(DFT/MM) calculations are further used in Paper 4, in order to address the questions that are not answered by the small model study, e.g. catalytic roles of active site residues. Computational methods for Paper 3 and Paper 4 are listed in Table 3 - 5, and more details can be found in Appendix.

\textbf{Table 3 - 5. Summary of computational methods for Paper 3 and Paper 4}

<table>
<thead>
<tr>
<th>Model System</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper 3</td>
<td>5HN + FADHOOH</td>
</tr>
<tr>
<td>Paper 4</td>
<td>SI: MHPC + FADHOOH + active site atoms</td>
</tr>
<tr>
<td></td>
<td>SII: MHPC + FADHOOH + whole enzyme</td>
</tr>
</tbody>
</table>

\textbf{3.2.2 DFT study on small model (Paper 3)}

According to previous titration experiments, the substrates exist in their tripolar forms when binding to the MHPCO.\textsuperscript{[181]} In PHBH, although the substrate p-hydroxybenzoate should not have the phenolic hydroxyl deprotonated at neutral pH, it attains a phenolate form in the enzyme.\textsuperscript{[173]} It was argued that, with the phenolic hydroxyl deprotonated, the substrate is in a higher energy state, thus the energy barrier for hydroxylation is lowered.\textsuperscript{[185]} These enzymes hence activate the substrates by controlling their protonation states.

In Paper 3, we used an ethanolyl group (-CH\textsubscript{2}CH\textsubscript{2}OH) to replace the ribityl chain in order to simplify the large system. We obtain an energy barrier of 33.8 kcal/mol for the hydroxylation reaction, which is higher than previous DFT studies using the B3LYP method.\textsuperscript{[185]} The too high barrier found herein can be ascribed to the fact that no explicit active site residues were included, that might stabilize the TS through hydrogen bonding or electrostatic interactions. When
optimizing the product complex, this immediately rearranges to complex III instead of II (Figure 3 - 14). The final product complex of the hydroxylation step is determined by the proton affinity (PA) of FADHO$^-$ versus the hydroxylated intermediate, which are 345 and 333 kcal/mol in MPWB1K/6-311+G(2d,2p) calculations, respectively. Hence, complex III is more stable than II, in accordance with the spontaneous rearrangement found. The energy surface for proton transfer from FADHO$^-$ to 5HN (i.e. conversion from complex III to II) was scanned, showing that the reaction is strictly endothermic along the reaction coordinate, even as the O-H distance in complex II reached 0.98 Å. However the total energy difference between II and III is only 1.5 kcal/mol, which is much smaller than the PA difference (12 kcal/mol). This implies that the product complex is not only determined by the PA values, and more factors must be considered. Hence we suggest that both complex II and III can be the product of the hydroxylation step, with complex III being slightly favored (Figure 3 - 14).

MHPCO is an unusual member in the aromatic hydroxylase family because after the hydroxylation, there is a further ring-cleavage process to yield an acyclic product instead of an aromatic one. According to Chaiyen et al.[180, 181] the ring-opening process is faster than the rearomatization of I1 (I1 and I2 are defined in Figure 3 - 14) so as to prevent the formation of the aromatic product. However, our calculations show that the C2-C3 bond strength in I1 is not that weak (>40 kcal/mol). In addition, when we optimized the geometry of I1 in water (optimization...
with PCM model), the rearomatization of \( \text{I}1 \) occurs spontaneously. We suggest that once the hydroxylated intermediate is released into aqueous solution, the rearomatization of \( \text{I}1 \) occurs (Figure 3 - 13). Hence, in the active site of MHPCO there should be limited amounts of water present around the substrate, and the hydroxylated intermediate is not released into water until the aromatic ring is cleaved. The enzyme structure study by McCulloch et al. shows that the active site pocket of MHPCO is significantly different from the other aromatic hydroxylases.\(^{[172]}\) Therefore, the unusual product specificity of MHPCO is determined by both the unstable hydroxylated intermediate and the particular protein structure.

In previous studies by Chaiyen et al, it is suggested that after the hydroxylation, with the FADHO\(^-\) moiety leaving, the hydroxylated substrate (similar to \( \text{I}1 \)) undergoes the ring-cleavage with the help of water.\(^{[180, 181]}\) Recently, McCulloch et al. proposed a new mechanism involving that the hydroxylated intermediate cleaves at the C2-C3 bond to yield a ketene intermediate which is further hydrated to the product.\(^{[172]}\) Both mechanisms have been evaluated in vacuo (Figure 3 - 15). According to our calculations, the conversions \( \text{I}1 \rightarrow \text{I}3 \) and \( \text{I}4 \rightarrow \text{I}5 \) are difficult (\( \text{TS1-3} \), \( \Delta \Delta G^\neq = 41.9 \text{ kcal/mol} \); \( \text{TS4-5} \), \( \Delta \Delta G^\neq = 38.5 \text{ kcal/mol} \)), implying that the pathways proposed by McCulloch et al. and Chaiyen et al. are unlikely to occur.

![Figure 3 - 15. Previous mechanisms proposed by Chaiyen et al.\(^{[180, 181]}\) and McCulloch et al.\(^{[172]}\) (Paper 3)](image)

When using the intermediate \( \text{I}2 \), two different pathways leading to the acyclic product are found (Figure 3 - 16). Compared with Figure 3 - 15, these two pathways have much lower barrier.
heights. In pathway A, the C2-C3 bond cleaves to yield the ketene intermediate $I_{6a}$. The energy barrier of the reaction is 14.1 kcal/mol ($TS_{2-6a}$), which is much lower than the direct ring-opening reaction in Figure 3 - 15 ($TS_{1-3}$, $\Delta \Delta G^\neq = 41.9$ kcal/mol). In pathway B, $I_{2}$ first forms an epoxy transition state ($TS_{2-7}$, $\Delta \Delta G^\neq = 11.1$ kcal/mol), yielding $I_{7}$ With the weak C3-O bond broken ($TS_{7-6b}$, $\Delta \Delta G^\neq = 0.5$ kcal/mol), intermediate $I_{6b}$ is obtained. Ketene intermediates $I_{6a}$ and $I_{6b}$ are stoichiometrically identical, and react with water to yield two different configurations of the product (A and B) which are observable by $^1$H-NMR spectroscopy.$^{[181]}$ Although the barriers for the two ring-opening pathways become higher ($TS_{2-6a}$, 20.7 kcal/mol; $TS_{2-7}$, 16.0 kcal/mol) with the presence of FADHOOH, these two pathways are much more favorable than previously proposed mechanisms.

**Pathway A**

![Pathway A diagram]

**Pathway B**

![Pathway B diagram]

**Figure 3 - 16. Current ring-opening pathways (Paper 3)**

**3.2.3 DFT cluster study on MHPCO active site (Paper 4)**

As mentioned above, the too high hydroxylation barrier from MPWB1K calculations can be ascribed to the fact that no explicit active site residues were included. Hence, in Paper 4, we considered two model systems, i.e. $S_1$ and $S_II$. $S_1$ model system is used for determining the QM/MM partitioning of the ONIOM calculations and identifying important residues for the reaction. The $S_1$ model contains atoms that attain hydrogen bonds to the substrate MHPC and to the peroxide of FADHOOH, and thus includes FADHOOH, MHPC, Pro295, and Ala296 and the
side chains of Tyr82, Tyr223, and Arg211, as well as eight crystal water molecules (149 atoms, Figure 3 - 17). The initial coordinates of these residues were obtained from the prepared crystal structure. The hydrophobic residues are excluded to our model, as they are not expected to substantially affect the reaction pathways. Moreover, inclusion of the hydrophobic residues will significantly increase the computational demand. All the intermediates and TSs in Figure 3 - 18 are optimized using the B3LYP functional. Two basis sets are used, i.e. 6-31G(d) (BS1) and 6-311+G(2d,p) (BS2), for geometry optimizations and single point energy calculations, respectively.

Figure 3 - 17. Active site of MHPCO\(^a\) (Paper 4)

\(^a\)All 149 atoms in the figure were included in the S\(_1\) model, DFT-only calculations, and the red-labeled atoms were included as the QM part in the largest ONIOM model calculation (96 QM atoms).
Chapter 3 Result and Discussion

3.2 Computational Study on MHPCO (Papers 3, 4)

Figure 3 - 18. Intermediates and TSs studied in Paper 4

Through the QM calculations on the $S_1$ system, the aim is to identify residues that have a major effect on the potential energy profile along the reaction paths and treat those residues quantum mechanically in the ONIOM calculations. The mean unsigned error (MUE) between the relative energies of $S_1$ and $S_1^{\text{Res}}$ was used to assess the importance of the residue to the reaction energy profile. For example, in the $S_1$ system, we located all intermediates and TSs in Figure 3 - 18, and the relative energies are $E_1$ ($E_1^{} = 0$), $E_{\text{TS}-2}$, $E_3$ … $E_{\text{TS}-8}$, $E_9$. When we remove the one residue, for example Arg211, we obtain the system $S_1^{\text{Arg211}}$, with the relative energies $E_1'$ ($E_1'^{} = 0$), $E_{\text{TS}-2}'$, $E_3'$, …, $E_{\text{TS}-8}'$, $E_9'$. The MUE of residue Arg211 is then evaluated according to Equation 3 - 1. The MUE of the active-site residues are summarized in Table 3 - 6. According to Table 3 - 6, we found that the magnitudes of the MUE values for the BS2 basis set were smaller than the results with the BS1 basis set for all the cases investigated, implying that the use of different basis sets can influence the relative energies considerably. The larger basis set is more suitable for the purpose of this study, since errors on relative energies are generally reduced (becomes more stable) when a higher basis set is employed (Table 3 - 6).
Chapter 3 Result and Discussion

3.2 Computational Study on MHPCO (Papers 3, 4)

Equation 3 - 1. Definition of MUE (using Arg211 as an example; Paper 4)

\[ MUE_{Arg211} = \frac{1}{8} \left[ |E_{TS-2} - E_{TS-2}'| + |E_3 - E_3'| + \cdots + |E_{TS-8} - E_{TS-8}'| + |E_9 - E_9'| \right] \]

Table 3 - 6. MUE\textsuperscript{a} of active site residues and waters (Paper 4)

<table>
<thead>
<tr>
<th>Residue/Water</th>
<th>MUE (B3LYP/BS1)</th>
<th>MUE (B3LYP/BS2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro295-Ala296</td>
<td>10.4</td>
<td>8.8</td>
</tr>
<tr>
<td>Tyr82</td>
<td>12.0</td>
<td>8.1</td>
</tr>
<tr>
<td>Arg211</td>
<td>2.2</td>
<td>2.1</td>
</tr>
<tr>
<td>Tyr223</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td>W654</td>
<td>7.2</td>
<td>5.5</td>
</tr>
<tr>
<td>W410</td>
<td>2.6</td>
<td>1.7</td>
</tr>
<tr>
<td>W435</td>
<td>4.9</td>
<td>3.9</td>
</tr>
<tr>
<td>W493</td>
<td>4.9</td>
<td>3.7</td>
</tr>
<tr>
<td>W546</td>
<td>1.8</td>
<td>1.5</td>
</tr>
<tr>
<td>W547</td>
<td>3.3</td>
<td>2.7</td>
</tr>
<tr>
<td>W633</td>
<td>5.4</td>
<td>4.0</td>
</tr>
<tr>
<td>W642</td>
<td>2.7</td>
<td>2.2</td>
</tr>
</tbody>
</table>

\textsuperscript{a}As defined in Equation 3 - 1, in kcal/mol

Comparing the relative energy (BS2) profiles of \( S_1 \) and \( S_{1\text{Res}} \), the peptide bond of Pro295-Ala296 and the side chain of Tyr82 significantly affect the energy profile (MUE values 8.8 and 8.1 kcal/mol, respectively). All intermediates and TSs are stabilized (lower relative energy) with Pro295-Ala296 removed, while Tyr82 plays an opposite role (Figure 3 - 19). According to our calculations, Arg211 and Tyr223 have minor effects on the reaction (MUE values are 2.1 kcal/mol and 0.6 kcal/mol, respectively). Therefore, Arg211 and Tyr223 as well as two water
molecules (W410 and W546, MUE values are 1.7 and 1.5 kcal/mol) are not treated as part of the QM region in the ONIOM calculations.

```Figure 3 - 19. Relative energy (BS2) profiles of $S_I$ model calculations: $S_I$ in black, $S_I$-Pro295-Ala296 in red and $S_I$-Tyr82 in Blue (Paper 4)```

Pro295-Ala296 and Tyr82 affect not only the relative energies of various points on the PES (MUE), but also the barrier heights. The hydroxylation barrier is lowered from 31.5 kcal/mol for $S_I$ to 26.4 kcal/mol when Pro295-Ala296 is removed, suggesting that the hydrogen bond between the amide oxygen and the peroxide (OOH) proton stabilizes the reactant more than the TS. This does not support the QM/MM study by Ridder et al.\cite{187, 188}, who suggested that the conserved proline peptide bond can stabilize the TS of the hydroxylation step. A likely reason is that the MHPC system is different from the PHBH one in terms of substrate, active-site interactions, and residue orientations. In the active site of MHPC, there are more water molecules than in PHBH, and the substrate interacts with these water molecules instead of a direct interaction with residues such as Tyr82, Tyr223, and Arg211. Besides the hydroxylation step, Pro295-Ala296 affects the
barrier height of TS-8. Tyr82 has minor effect (less than) on the hydroxylation step, but can affect the barrier heights of the ring-cleavage pathways (3 and 5 kcal/mol for TS-4 and TS-6 respectively).

### 3.2.4 QM/MM (ONIOM) study on the whole MHPCO (Paper 4)

The $S_{II}$ system contains the whole MHPCO. A two-layer ONIOM scheme is used for the QM/MM calculations on the $S_{II}$ system. The ONIOM-ME scheme was used for geometry optimizations and frequency calculations, and the ONIOM-EE scheme was used for single-point energy calculations on ONIOM-ME optimized geometries. Eight different QM/MM partitioning schemes are used, and the largest QM part contains 96 QM atoms, as described in Figure 3 - 17. The partitioning schemes are summarized in Table 3 - 7. For the ONIOM-1 calculations, both B3LYP and M062X methods were used for the QM layer, while for the other partitioning schemes, only the B3LYP functional was used to treat the QM part. Details see Appendix.

#### Table 3 - 7. Different ONIOM partitioning schemes employed (Paper 4)

<table>
<thead>
<tr>
<th>ONIOM calculations</th>
<th>QM region</th>
<th>QM atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONIOM-1</td>
<td>FAD, MHPC, Pro295-Ala296, Tyr82, six water molecules</td>
<td>96</td>
</tr>
<tr>
<td>ONIOM-2</td>
<td>FAD, MHPC, Pro295-Ala296, six water molecules</td>
<td>81</td>
</tr>
<tr>
<td>ONIOM-3</td>
<td>FAD, MHPC, Tyr82, six water molecules</td>
<td>88</td>
</tr>
<tr>
<td>ONIOM-4</td>
<td>FAD, MHPC, six water molecules</td>
<td>73</td>
</tr>
<tr>
<td>ONIOM-5</td>
<td>FAD, MHPC, Pro295-Ala296, Tyr82</td>
<td>78</td>
</tr>
<tr>
<td>ONIOM-6</td>
<td>FAD, MHPC, Pro295-Ala296</td>
<td>63</td>
</tr>
<tr>
<td>ONIOM-7</td>
<td>FAD, MHPC, Tyr82</td>
<td>70</td>
</tr>
<tr>
<td>ONIOM-8</td>
<td>FAD, MHPC</td>
<td>55</td>
</tr>
</tbody>
</table>
We will focus on the ONIOM(B3LYP: Amber) geometries in the ONIOM-1 calculations (96 QM atoms). In the hydroxylation, the peroxide proton forms a hydrogen bond in 1 with the amide oxygen of Pro295-Ala296 (H⋅⋅⋅O distance 1.750 Å, Figure 3 - 20a); a hydrogen bond that remains in TS-2 (H⋅⋅⋅O distance 1.780 Å, Figure 3 - 20b). W654 links FAD and MHPC through hydrogen bonds. Tyr82 forms an indirect hydrogen bond with the N1 proton in the substrate. In TS-2, the hydroxyl lies in the middle between FAD and MHPC with the O⋅⋅⋅C and O⋅⋅⋅O distances being 1.940 and 1.958 Å, respectively. In a very accurate model for the PHBH enzyme, the average O⋅⋅⋅C and O⋅⋅⋅O distances in the transition state are 2.10 and 1.83 Å (B3LYP/GROMOS geometry), respectively. A likely reason for the geometry differences (O⋅⋅⋅C and O⋅⋅⋅O distances) between TS-2 in our study and the transition state reported by Mata et al. could be that the MHPCO system is not entirely the same as the PHBH system in terms of both substrate and protein environment. A back-transfer process of the hydroxyl proton was observed in this study, which was not reported in PHBH. When the hydroxyl attaches to C2 of MHPC, the hydroxyl proton transfers back to the FAD moiety and forms a hydrogen bond with the transferred hydroxyl oxygen in intermediate 3 (H⋅⋅⋅O distance 1.769 Å, Figure 3 - 20c), consistent with the gasphase calculations. The hydrogen bond between W654 and the phenolic oxygen of MHPC becomes weaker in 3 compared with in 1 (1.652 Å vs 2.016 Å), implying that this oxygen is less negatively charged in the oxygenated intermediate.
3.2 Computational Study on MHPCO (Papers 3, 4)

In the ring-opening pathway A, the C2···C3 bond in MHPC cleaves directly to give intermediate 5. In TS-4, the C2···C3 distance increases to 2.058 Å, and the hydrogen bond between W654 and phenolic oxygen becomes much weaker than in 3 (H···O distance 2.920 Å vs 2.016 Å, Figure 3-21a). The breaking of this hydrogen bond may increase the barrier height for this path, and will be discussed in more detail later. In the ketene intermediate 5, W654 forms a strong hydrogen bond (H···O distance 1.663 Å) with the newly formed carbonyl group (Figure 3-21b), which in turn stabilizes this intermediate. The hydrogen bond of Tyr82 is unchanged between 3 and TS-4.

In ring-opening pathway B, the oxygen on C2 moves toward C3 and forms an epoxy-like transition state TS-6 (O···C3 distance 1.717 Å, Figure 3-21c), followed by cleavage of the C2-C3 bond to give intermediate 9 (Figure 3-21d). We were unable to locate intermediate 7, probably because the conversion from 7 to 9 is barrierless in the protein environment. In TS-6, W654 forms a strong hydrogen bond with the phenolic oxygen in MHPC (H···O distance 1.698 Å,

---

**Figure 3-20. Geometries of ONIOM-1 calculations: a) 1, b) TS-2 and c) 3 (Paper 4)**
Figure 3 - 21c), this being much weaker in both 3 and 9 (H⋯O distances are 2.016 and 2.047 Å, respectively) which implies the oxygen is more negative in the TS than in either reactant or intermediate. The hydrogen bond between Tyr82 and the substrate in 3 and TS-6 again does not change significantly, implying that the Tyr82 does not affect the ring-opening energy barriers via forming and breaking hydrogen bonds. Therefore, it is likely that Tyr82 affects the reactions through electrostatic interactions.

Figure 3 - 21. Geometries from ONIOM-1 calculations of the ring-opening pathways: a) TS-4, b) 5, c) TS-6 and d) 9 (Paper 4)

The ONIOM-EE (BS2) energies are used to characterize the potential energy profile along the reaction coordinates (Figure 3 - 22). From Figure 3 - 22, we note that the ONIOM-1 calculations with B3LYP:Amber and M062X:Amber do not show any significant differences, except for TS-2. The M062X:Amber obtains the energy barrier 47.1 kcal/mol for the hydroxylation, while for B3LYP:Amber, the value is much lower (32.2 kcal/mol). As the experimental value is only 16.7
kcal/mol, there must be additional factors that are not considered in our system, such as the dynamics of the enzyme, or that the single configuration (crystal structure) of the protein available to date may not be sufficiently accurate to describe the MHPCO reaction.

Figure 3 - 22. Relative Energy profiles of $S_I$ (QM, B3LYP, in black) and $S_{II}$ (ONIOM-EE, B3LYP:Amber in red and M062X:Amber in blue) calculations (Paper 4)

Comparing the ONIOM(B3LYP:Amber) calculations ($S_{II}$) with the QM (B3LYP) calculations ($S_I$), the activation energies for the hydroxylation are similar (32.2 kcal/mol vs 31.5 kcal/mol), suggesting that the MM layer has minor effect on the hydroxylation step. For the ring-opening pathways, the results show several differences. First, according to the QM calculations, the intermediate 5 (product of pathway A) is more stable than 9 (product of pathway B); however, in the ONIOM calculations, 9 becomes more stable. Second, the energy barriers obtained from QM (B3LYP) calculations (22.0 kcal/mol for pathway A and 16.6 kcal/mol for pathway B) are considerably higher than those from ONIOM-1 (B3LYP-Amber) calculations (14.3 kcal/mol for pathway A and 5.6 kcal/mol for pathway B), suggesting that the ring-opening process becomes easier in the presence of the protein environment. Moreover, the barrier difference between the two pathways (5.4 kcal/mol in QM vs 8.7 kcal/mol in ONIOM) becomes larger in the ONIOM
calculations, and pathway B ($\Delta E^z = 5.6$ kcal/mol) becomes even more favored than pathway A ($\Delta E^z = 14.3$ kcal/mol). Hence, according to the ONIOM calculations, pathway B is more favored over pathway A, both thermodynamically and kinetically. Active site waters may play an important role in controlling the ring-opening paths. In intermediate 3, W654 forms hydrogen bond with the phenolic oxygen of MHPC (Figure 3 - 20c). To reach TS-4 of pathway A, this hydrogen bond must be broken (Figure 3 - 21a), whereas for TS-6 of pathway B, this hydrogen bond becomes even stronger (Figure 3 - 21c).

**Table 3 - 8. Comparison of activation energy in different ONIOM partitioning schemes (Paper 4)**

<table>
<thead>
<tr>
<th>ONIOM scheme$^a$</th>
<th>Activation Energy$^b$</th>
<th>Hydroxylation</th>
<th>Ring-opening Pathway A</th>
<th>Ring-opening Pathway B</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONIOM-1</td>
<td>32.2</td>
<td>14.3</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>ONIOM-2</td>
<td>32.5</td>
<td>14.8</td>
<td>5.5</td>
<td></td>
</tr>
<tr>
<td>ONIOM-3</td>
<td>29.1</td>
<td>16.9</td>
<td>2.2</td>
<td></td>
</tr>
<tr>
<td>ONIOM-4</td>
<td>25.3</td>
<td>17.2</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td>ONIOM-5</td>
<td>29.8</td>
<td>18.2</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>ONIOM-6</td>
<td>29.9</td>
<td>18.2</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>ONIOM-7</td>
<td>23.0</td>
<td>22.7</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>ONIOM-8</td>
<td>26.2</td>
<td>20.6</td>
<td>2.1</td>
<td></td>
</tr>
</tbody>
</table>

$^a$ONIOM partitionings described in Table 3 - 7

$^b$ONIOM-EE/BS2 energies, in kcal/mol

Besides the main ONIOM model calculations (ONIOM-1, 96 QM atoms), we also performed other ONIOM calculations using several different QM/MM partitioning schemes (Table 3 - 7), in order to explore if the results are affected by alternative QM/MM partitioning schemes. The activation energies for TS-2, TS-4, and TS-6 in the different QM/MM partitioning schemes are
summarized in Table 3 - 8. The energy barriers are very sensitive to the QM/MM partitioning scheme, especially for the treatment of the peptide bond in Pro295-Ala296. When this peptide bond is added to the QM region, the barrier heights of the hydroxylation and the ring-opening pathway B increase, while the barrier for the ring-opening pathway A decreases.
3.3 Computational Study on PBGS (Paper 5)

3.3.1 Introduction

Tetrapyrrole derivatives, such as porphyrin, chlorine and corrin, are essential to most life forms\cite{191}. Porphobilinogen synthase (PBGS) catalyzes the asymmetric condensation and cyclization of two 5-aminolevulinic acid (5-ALA) substrate molecules to give porphobilinogen (PBG), and is known as the first common step in the biosynthesis of the tetrapyrroles (Figure 3 - 23)\cite{192-194}.

![Reaction catalyzed by PBGS](image)

**Figure 3 - 23. Reaction catalyzed by PBGS**

PBGSs from most organisms utilize metal ions (Zn$^{2+}$, Mg$^{2+}$, K$^{+}$ and Na$^{+}$) as cofactors. The metal ions of PBGS reside at two sites, i.e., the active site and the allosteric site\cite{195}. Most PBGSs have Zn$^{2+}$ or Mg$^{2+}$ in the active site\cite{195-199}, albeit some do not use any active site metal ion at all\cite{200}. PBGSs that have active site Zn$^{2+}$ are commonly found in metazoan (including human), archea, and yeast organisms. We herein focus on the yeast PBGS to illustrate the catalytic process of the enzyme (Figure 3 - 24).
Figure 3 - 24. Catalytic process of PBGS

The two substrate molecules bind to the enzyme separately. To distinguish the two ALA molecules, the labels A (acetyl-) and P (propionyl-) are used throughout this study (Figure 3 - 23) referring to the length of the resulting carboxylic acid tail in the final PBG molecule. When the active site lid is open, the P-site substrate binds to the enzyme first, and forms a Schiff base with the P-site lysine (Lys263, yeast numbering; Figure 3 - 24)\textsuperscript{196, 197, 201}. A water molecule is released and the second substrate enters the active site. With the A-site ALA bound, the active site lid becomes closed, to prevent solvent molecules from entering the active site\textsuperscript{202}. Binding of the A-site ALA is controversial, as the intermediate with both substrates bound is short-lived and
has never been observed\textsuperscript{[202]}. Crystal structures of PBGS-inhibitor complexes\textsuperscript{[203-205]} have shown that the A-site ALA might also form a Schiff base with the A-site lysine (Lys210). Mutagenesis studies suggest that the P-site lysine is more important than the A-site lysine in terms of decrease in enzymatic activity\textsuperscript{[206]} . It is argued that the A-site ALA should bind in a geometry that will support the highly exothermic reaction and block access to the bulk solvent, irrespective of whether an A-site Schiff base is formed or not\textsuperscript{[202]}.

In the current QM/MM study, we have considered several starting structures and two different protonation states for the Michaelis complex, aiming to fill the “blank” in Figure 3 - 24. Given the short distance between the nitrogen atom of Lys210 and that of P-site bound ALA (\textasciitilde4Å in crystal structure 1H7O\textsuperscript{[196]}; Figure 3 - 24), we suggest that the A-site ALA may bind directly to the P-site ALA via Schiff base formation to get PS1-S or PS2-S (Figure 3 - 25); such Michaelis complexes were not considered previously (PS2-S is an intermediate in an earlier DFT study\textsuperscript{[207]}). The Schiff base formation as well as Schiff base exchange reactions are not considered in this study since multiple groups (water molecules, metal ions as well as active site residues) are probably involved\textsuperscript{[208, 209]}, along with conformational changes of the enzyme\textsuperscript{[202]}.

![Figure 3 - 25. Michaelis complexes considered in the current study](image)

The role of the active site zinc ion has been discussed in a number of studies\textsuperscript{[195, 197, 202, 209-217]}, but no consensus has been reached. In the crystal structure of the yeast PBGS-PBG* intermediate\textsuperscript{[197]},
Zn$^{2+}$ binds to three cysteine residues (Cys133, Cys135 and Cys143) as well as the amino nitrogen of the A-site ALA. It is suggested that Zn$^{2+}$ is important for facilitating the binding and reactivity of the A-site ALA, as well as maintaining the stability of the enzyme structure\textsuperscript{[197, 202]}. It has also been hypothesized that the active site Zn$^{2+}$ is important for removal of the hydroxyl group during the Schiff base formation and for controlling various proton transfer processes via electrostatic interactions\textsuperscript{[209]}. Recently, the crystal structure of the yeast PBGS-PBG\textsuperscript{*} intermediate\textsuperscript{[197]} (which has an active site Zn$^{2+}$) and the crystal structure of \textit{Toxoplasma gondii} PBGS-PBG complex\textsuperscript{[200]} (Tg-PBGS, which has no active site metal ion) were resolved, which reveal that the orientation of the product, as well as the positioning of most of the conserved residues are almost identical (Figure 3 - 26)\textsuperscript{[200]}. We therefore base our models on the assumption that active site metal ion is used for binding the A-site ALA, but not for the chemical step.

Figure 3 - 26. QM region of the QM/MM calculations: a) yeast PBGS; b) Tg-PBGS

Another issue is how the pyrrole ring of PBG is formed\textsuperscript{[218, 219]}. Generally speaking, there are two routes to reach the pyrrole ring, either C3-C4 bond formation first or C5-N1 bond formation first (numbering cf. Figure 3 - 23). Recent crystal structures\textsuperscript{[196, 197, 203, 204]} of PBGS-inhibitor complexes as well as isotope experiments\textsuperscript{[209, 220]} support the C3-C4 bond formation first pathway\textsuperscript{[209]}, hereafter defined as Mechanism 1 (Figure 3 - 27). In contrast, small model DFT
calculations favor the C5-N1 bond formation first pathway, which is defined as Mechanism 2 (Figure 3 - 30). It should be noted that the protonation state of Mechanism 2 is different from that of Mechanism 1. The favored mechanisms might hence be different for different protonation states.

In this study, various mechanisms for the chemical step are modeled using the yeast PBGS. Mechanism 1 and 4, which we show are the most likely ones, are also modeled using the Tg-PBGS structure, in order to validate the assumption that the active site metal ion is not involved in the chemical step. We wish to address the following questions: 1) whether Schiff base formation of A-site ALA with A-site Lys210 is necessary for catalysis; 2) which protonation state or starting structure is preferred; 3) which bond (C3-C4 or C5-N1) is formed first in the reaction. A summary of computational methods is listed in Table 3 - 9, and more details can be found in Appendix.

Table 3 - 9. Summary of computational methods

<table>
<thead>
<tr>
<th></th>
<th>Yeast PBGS</th>
<th>Tg-PBGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protonation State</td>
<td>PS1 and PS2</td>
<td>PS1</td>
</tr>
<tr>
<td>Mechanism studied</td>
<td>All mechanisms</td>
<td>Mechanism 1 and 4</td>
</tr>
<tr>
<td>Geometry optimization</td>
<td>B3LYP/BS1:AMBER&lt;sup&gt;a&lt;/sup&gt;</td>
<td>B3LYP/BS3:AMBER&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>B3LYP/BS2:AMBER&lt;sup&gt;b&lt;/sup&gt;</td>
<td>B3LYP/BS4:AMBER&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup>BS1 : SDD for Zn<sup>2+</sup>, 6-31G(d) for other elements

<sup>b</sup>BS2 : SDD for Zn<sup>2+</sup>, 6-311+G(2d,p) for other elements

<sup>c</sup>BS3 : 6-31G(d)

<sup>d</sup>BS4 : 6-311+G(2d,p)
3.3.2 Mechanism 1

To date, Mechanism 1 is the most favored one for PBGS\cite{196, 197, 203, 205, 209}. According to this mechanism, both ALA molecules bind to the enzyme in the protonated Schiff base form (PS1-D-1 in Figure 3 - 27). $H_R$ at the A-site ALA is abstracted by its own carboxyl group to give the enamine intermediate (PS1-D-2), followed by C3-C4 bond formation. Isotope effect experiments\cite{209} show that deuterium substitution of $H_R$ has a larger isotope effect on $V_{\text{max}}$ than what $H_S$ has, but the effect on $V_{\text{max}}/K_M$ is identical for deuterium substitution at both $H_R$ and $H_S$ which suggests that deprotonation might not be the rate-limiting step\cite{209}. PS1-D-3 is unstable and converts to the ring-closed intermediate PS1-D-4\cite{205}. The steps from PS1-D-4 to the final product are relatively easy. According to our QM/MM calculations, a minor change to the previous mechanism is observed, in that Asp131/Ser179 can abstract a proton from the N1 atom of the P-site ALA and stabilize PS1-D-4, which was not considered previously.

The energies for the three stages of Mechanism 1 are summarized in Figure 3 - 28. Our results indicate that the overall rate-limiting step is the deprotonation stage, which has a barrier of 22.8 kcal/mol. This is somewhat higher than the phenomenological barrier height of $\sim$18 kcal/mol, derived from the experimental turnover rate\cite{220} using the Eyring equation. The difference between theoretical results and experimental data might be due to the neglect of entropic contributions and quantum tunneling effects. In addition, deprotonation might occur in the substrate binding process and directly give the deprotonated intermediates\cite{202, 209}. As the geometries are only optimized locally, the transition from the open form to the closed form or other large conformational changes is also not considered in this work.
Figure 3 - 27. Mechanism 1
The ring closure stage can start from two different conformations, PS1-D-2a and PS1-D-2b. The difference between these is whether the Schiff base proton of Lys263 forms a hydrogen bond with the N1 atom (dashed line, Figure 3 - 27) or not. This hydrogen bond stabilizes PS1-D-2a and increases the barrier height for the C3-C4 bond formation ($\Delta E^\neq = 25.5$ kcal/mol in PS1-D-2a vs $\Delta E^\neq = 19.0$ kcal/mol in PS1-D-2b). Interestingly, the conversion PS1-D-2a $\rightarrow$ PS1-D-4 is concerted (Figure 3 - 29). Although intermediate PS1-D-3 can be found in the stepwise pathway from PS1-D-2b (Figure 3 - 27), the conversion PS1-D-3 $\rightarrow$ PS1-D-4 is barrierless (Figure 3 - 28), implying that PS1-D-3 is unstable. As the stepwise ring closure from PS1-D-2b ($\Delta E^\neq = 19.0$ kcal/mol) is more favorable than the concerted ring closure from PS1-D-2a ($\Delta E^\neq = 25.5$ kcal/mol), we suggest that before the C3-C4 bond formation, PS1-D-2a must convert to PS1-D-2b by breaking the hydrogen bond between N1 and the iminium proton of Lys263 through a rotation of the N-C3-C2-N1 dihedral angle. The PBG formation stage is fast and we did not observe any step that has a barrier height of more than 10 kcal/mol (Figure 3 - 28). The key conversions PS1-D-1 $\rightarrow$ PS1-D-2a and PS1-D-2b $\rightarrow$ PS1-D-3 are also modeled in the Tg-PBGS system, and the results are consistent with those in the yeast PBGS system. This suggests that the active site metal ion is presumably not involved in the chemical step.

![Figure 3 - 28. Relative energies of key intermediates and transition states in Mechanism 1](image-url)
3.3.3 Mechanisms 2 and 3

Mechanism 2 and Mechanism 3 were previously studied using DFT-only (B3LYP/6-31G(d)//IEFPCM) calculations, and Mechanism 2 was shown to be more favorable than Mechanism 3\textsuperscript{[207]}\textsuperscript{[207]}. However, in our QM/MM calculations on the PS2 system (Figure 3 - 30), the barrier heights for deprotonation are dramatically different from the DFT-only calculations. It should be noted that protonation state PS2 is used in both Mechanism 2 and 3, albeit there is strong experimental evidence that the P-site Schiff base is protonated when the first ALA (P-site ALA) binds (Figure 3 - 24)\textsuperscript{[201]}.
In the previous B3LYP/6-31G(d)//IEFPCM calculations, the barrier heights for conversions PS2-S-1 $\rightarrow$ PS2-S-3 and PS2-D-1 $\rightarrow$ PS2-D-2 were 12.1 and 19.4 kcal/mol, respectively. However, in the current ONIOM calculations, the barrier heights for these two steps are considerably higher, 37.5 and 42.0 kcal/mol, suggesting that such conversions are unlikely to occur in PBGS. Such large differences between QM and QM/MM calculations may be due to the easier rotation of the N=C double bond in the Lys263 imine group (the nitrogen then abstracts the C4 proton H$^5$) in the QM-only calculations, which have more degrees of freedom. In addition, the overall electrostatic effects may also influence the barriers of these two conversions. An alternative pathway is also considered, in which Lys210 acts as a proton shuttle (Figure 3 - 30), but the barrier height is still
very high (PS2-S-TS1-2, $\Delta E^z = 30.0$ kcal/mol). We therefore suggest that the mechanisms in the DFT-only calculations\cite{207} might be problematic and that neglecting the enzymatic environment is not suitable for modeling these enzymatic reactions. Interestingly, the barriers for the subsequent ring closure stage are actually not high (15.5 and 16.9 kcal/mol, respectively), and if there is some other pathway that gives PS2-S-3 or PS2a-D-2, the ring closure should be straightforward (Table 3 - 10 and Table 3 - 11).

### Table 3 - 10. Relative energies of key intermediates and transition states in Mechanism 2

<table>
<thead>
<tr>
<th>Stage</th>
<th>Intermediate or TS</th>
<th>Relative Energy$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deprotonation</td>
<td>PS2-S-1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS1-2</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>PS2-S-2</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS2-3</td>
<td>16.3</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS1-3</td>
<td>37.5</td>
</tr>
<tr>
<td>Ring closure</td>
<td>PS2-S-3</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS3-4</td>
<td>23.0</td>
</tr>
<tr>
<td>Post PBG formation</td>
<td>PS2-S-4</td>
<td>-8.1</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS4-5</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>PS2-S-5</td>
<td>-22.3</td>
</tr>
</tbody>
</table>

$^a$ONIOM(B3LYP/BS2:AMBER)//ONIOM(B3LYP/BS1:AMBER) energies, yeast PBGS system, in kcal/mol

### Table 3 - 11. Relative energies of key intermediates and transition states in Mechanism 3

<table>
<thead>
<tr>
<th>Stage</th>
<th>Intermediate or TS</th>
<th>Relative Energy$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deprotonation</td>
<td>PS2a-D-1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>PS2a-D-TS1-2</td>
<td>42.0</td>
</tr>
<tr>
<td>Ring closure</td>
<td>PS2a-D-2$^b$</td>
<td>-13.3</td>
</tr>
<tr>
<td></td>
<td>PS2a-D-TS2-3</td>
<td>3.6</td>
</tr>
<tr>
<td>Post PBG formation</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

$^a$ONIOM(B3LYP/BS2:AMBER)//ONIOM(B3LYP/BS1:AMBER) energies, yeast PBGS system, in kcal/mol

$^b$conformation similar to PS1-D-2b is used
3.3.4 Mechanism 4

Mechanism 4 is similar to Mechanism 2, with the main variation that the P-site Schiff base is now protonated (Figure 3 - 31). Although deprotonation in Mechanism 4 is still as difficult as in Mechanism 1 (22.4 kcal/mol vs 22.8 kcal/mol), the ring closure becomes much easier (13.0 kcal/mol vs 19.0 kcal/mol; Figure 3 - 32). Given only 0.4 kcal/mol difference for the rate-limiting step, we suggest that both Mechanism 1 (C3-C4 bond formation first) and Mechanism 4 (C5-N1 bond formation first) are possible. It should be noted that the K210A mutation causes 1000 fold decrease in PBGS activity; however, this mutation also significantly affects the Schiff base formation of P-site ALA\(^{[206]}\). Therefore, mutagenesis is unlikely to tell whether Mechanism 1 or Mechanism 4 is actually the preferred one. We propose that whether Mechanism 1 or Mechanism 4 occurs in PBGS is determined by the binding mode of the A-site ALA, which might be different in different PBGS enzymes.

![Figure 3 - 31. Mechanism 4](image)
3.3 Computational Study on PBGS (Paper 5)

### 3.3.5 Other Pathways and General Consideration of the PBGS Mechanism

Besides the four mechanisms discussed above, we also modeled other possible pathways, for example, using PS2b-D as starting structure (Figure 3 - 25 and Figure 3 - 33). The common feature of all the mechanisms is that the deprotonation of A-site ALA is rate-limiting. Once a deprotonated intermediate is generated, the ring closure is straightforward ($\Delta E^\ddagger = 15$-19 kcal/mol in the various mechanisms). However, in all our attempts to find alternative mechanisms for PBGS, the rate-limiting (deprotonation stage) barriers are always higher than those of Mechanism 1 and Mechanism 4 (Table 3 - 12). It seems that none of the base groups (Figure 3 - 33) considered in the current study is strong enough to abstract the C4 protons on A-site ALA (Table 3 - 12). Even in Mechanism 1 and Mechanism 4, the activation energy ~22 kcal/mol is relatively high for an enzymatic reaction. Therefore, we suggest that future work should focus on the binding mode of A-site ALA and the deprotonation of the C4 protons, rather than which bond (C3-C4 or C5-N1) is formed first in the chemical step.

**Figure 3 - 32.** Relative energies of key intermediates and transition states in Mechanism 4

![Relative energies of key intermediates and transition states in Mechanism 4](image-url)
3.3 Computational Study on PBGS (Paper 5)

Figure 3-33. Base groups that can abstract the C4 protons

Table 3-12. Summary of energy barriers for deprotonation

<table>
<thead>
<tr>
<th>Base group&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Deprotonation barrier&lt;sup&gt;b&lt;/sup&gt;</th>
<th>H&lt;sup&gt;R&lt;/sup&gt;</th>
<th>H&lt;sup&gt;S&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>PS1-D-B1</td>
<td>22.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PS1-D-B2</td>
<td>-</td>
<td>29.0</td>
</tr>
<tr>
<td>B</td>
<td>PS2a-D-B1</td>
<td>31.1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PS2a-D-B2</td>
<td>-</td>
<td>36.8</td>
</tr>
<tr>
<td></td>
<td>PS2a-D-B3</td>
<td>-</td>
<td>42.0</td>
</tr>
<tr>
<td>C</td>
<td>PS2b-D-B1</td>
<td>45.7</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PS2b-D-B2</td>
<td>-</td>
<td>48.0</td>
</tr>
<tr>
<td>D</td>
<td>PS1-S-B1</td>
<td>22.4</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PS1-S-B2</td>
<td>25.6</td>
<td>28.8</td>
</tr>
<tr>
<td>E</td>
<td>PS2-S-B1</td>
<td>37.5</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PS2-S-B2</td>
<td>26.1</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>PS2-S-B3</td>
<td>-</td>
<td>46.5</td>
</tr>
</tbody>
</table>

<sup>a</sup>base groups are defined in Figure 3-33

<sup>b</sup>ONIOM(B3LYP/BS2:AMBER)//ONIOM(B3LYP/BS1:AMBER) energies, in kcal/mol
3.4 Computational Study on the Stereo-specificity of GlucD (Paper 6)

3.4.1 Introduction

Enzymatic reactions with high stereo-specificity are difficult to achieve by synthetic approaches. Understanding enzyme specificity at the atomic level is important for enzyme function prediction and redesign of enzymes. An enzymatic reaction consists of three common steps, i.e. substrate binding, chemical reaction and product release. The first two steps are important for substrate specificity, and \( \frac{k_{cat}}{k_{M}} \) is often used to experimentally determine the best substrate for an enzyme. Nowadays, modeling enzyme-ligand systems with computer simulations has evolved to a routine methodology\[34, 109\]. To evaluate \( \frac{k_{cat}}{k_{M}} \) with computational approaches, one needs to model both substrate binding and chemical steps. However, most computational studies only focus on one of these aspects.

In most cases, enzyme (in vitro) function prediction problem is equivalent to studying enzyme substrate specificity, where the key is to find a ligand with high \( \frac{k_{cat}}{k_{M}} \) values. Therefore, combining properties for the substrate binding step (e.g. binding affinity) and for the chemical step (e.g. activation energy) should give more reasonable results in determining whether or not a ligand is a substrate. Docking-based approaches have been successfully used in enzyme function prediction\[5-8, 222, 223\]. However, an obvious limitation of docking-based approaches is that no information about the chemical step can be obtained\[8\].

We herein report a benchmarking study on the stereo-specificity of D-glucarate dehydratase (GlucD) based on combined docking and quantum mechanical/molecular mechanical (QM/MM) methods. GlucD catalyzes the dehydration of D-glucarate or L-idarate to give 5-keto-4-deoxy-D-glucarate (5-KDG), as well as the interconversion between D-glucarate and L-idarate (Figure 3 - 34)\[224-229\]. In the dehydration of D-glucarate, an active site histidine (H339) abstracts a proton from C5 to give the enediolate intermediate I1. The protonated H339 then transfers the proton to the hydroxyl on C4 to give the enol intermediate I2, which subsequently converts to the final product 5-KDG (Figure 3 - 34). Dehydration of L-idarate is initialized by K207, and then undergoes the same intermediates I1 and I2.
3.4.2 Specificity at the substrate binding step

GlucD is highly stereo-specific in the non-reactive regions, and as a result, none of m-allarate, D-mannarate or D-altrarate is a substrate of GlucD (Figure 3 - 35). In the current work, docking-based methods were used for studying the GlucD specificity at the binding step. Computational details can be found in Appendix. Protonation states of H339 and K207 were assigned according to the Michaelis complex for D-glucarate (Figure 3 - 34), as the reactive regions of the four ligands in Figure 3 - 35 are identical. Glide-SP, Glide-XP and induced fit docking (IFD) methods were used to get the binding poses of the four ligands in Figure 3 - 35, and the MM/GBSA method to evaluate binding affinities (Table 3 - 13). All three docking methods are able to reproduce the binding pose of D-glucarate as seen in the crystal structure. D-glucarate is the best binder in the IFD//MM/GBSA calculations and is the second best binder in the Glide-SP//MM/GBSA and the Glide-XP//MM/GBSA calculations (Table 3 - 13). As the natural substrate D-glucarate is expected to be the best binder, this suggests that the binding affinity at
the IFD/MM/GBSA level is appropriate for qualitatively distinguishing substrate and non-substrate for GlucD. It should be mentioned that in Table 3-13, MM/GBSA calculations are performed on the reactive poses of each ligand, i.e. the pose whose reactive region is similar to the ligand in the crystal structure of GlucD. We choose reactive poses because only these poses can undergo the chemical steps, which will be discussed later in the text. The reactive poses are obtained from the top 10 docking poses, and in most cases, the reactive pose actually has the highest MM/GBSA affinity, details see Table 3–14. Component energies for the IFD pose of D-glucarate are listed in Table 3–15, which shows that the binding free energies is dominated by electrostatics.

![Figure 3-35. Structures of D-glucarate and three non-substrates](image)

**Table 3-13. Relative binding affinities of the reactive poses of D-glucarate and the three non-substrates**

<table>
<thead>
<tr>
<th>Ligands</th>
<th>Glide-SP$^a$</th>
<th>Glide-XP$^a$</th>
<th>IFD$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-glucarate</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>m-allarate</td>
<td>-9.3</td>
<td>11.4</td>
<td>47.8</td>
</tr>
<tr>
<td>D-mannarate</td>
<td>9.5</td>
<td>34.0</td>
<td>24.2</td>
</tr>
<tr>
<td>D-altrarate</td>
<td>8.0</td>
<td>-21.0</td>
<td>18.0</td>
</tr>
</tbody>
</table>

$^a$binding affinities are calculated at the MM/GBSA level, in kcal/mol
### 3.4 Computational Study on the Stereo-specificity of GlucD (Paper 6)

#### Table 3 – 14a. The highest MM/GBSA binding affinities for the top 10 poses

<table>
<thead>
<tr>
<th>Ligands</th>
<th>Glide-SP</th>
<th>Glide-XP</th>
<th>IFD-XP</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-glucarate</td>
<td>-81.31</td>
<td>-74.87</td>
<td>-161.01</td>
</tr>
<tr>
<td>m-allarate</td>
<td>-90.57</td>
<td>-68.56</td>
<td>-117.75</td>
</tr>
<tr>
<td>D-mannarate</td>
<td>-71.77</td>
<td>-45.99</td>
<td>-136.77</td>
</tr>
<tr>
<td>D-altrarate</td>
<td>-76.10</td>
<td>-95.85</td>
<td>-148.42</td>
</tr>
</tbody>
</table>

*a* binding affinities are calculated at the MM/GBSA level, in kcal/mol

#### Table 3 – 14b. MM/GBSA binding affinities for the reactive poses (these poses are similar to the initial structure used for QM/MM calculations)

<table>
<thead>
<tr>
<th>Ligands</th>
<th>Glide-SP</th>
<th>Glide-XP</th>
<th>IFD-XP</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-glucarate</td>
<td>-81.31</td>
<td>-74.87</td>
<td>-161.01</td>
</tr>
<tr>
<td>m-allarate</td>
<td>-90.57</td>
<td>-63.52</td>
<td>-113.20</td>
</tr>
<tr>
<td>D-mannarate</td>
<td>-71.77</td>
<td>-40.85</td>
<td>-136.77</td>
</tr>
<tr>
<td>D-altrarate</td>
<td>-73.27</td>
<td>-95.85</td>
<td>-143.03</td>
</tr>
</tbody>
</table>

*a* binding affinities are calculated at the MM/GBSA level, in kcal/mol

#### Table 3 - 15. Component Energies of D-glucarate in the MM/GBSA//IFD calculations

<table>
<thead>
<tr>
<th>Component</th>
<th>Ligand</th>
<th>Receptor</th>
<th>Complex</th>
<th>Complex – (Ligand + Receptor)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bond + Angle + Torsion</td>
<td>22.5</td>
<td>2157.8</td>
<td>2170.4</td>
<td>-9.9</td>
</tr>
<tr>
<td>Electrostatic</td>
<td>20.8</td>
<td>-14748.5</td>
<td>-15341.5</td>
<td>-613.8</td>
</tr>
<tr>
<td>van der Waals</td>
<td>10.6</td>
<td>-1979.9</td>
<td>-1975.8</td>
<td>-6.4</td>
</tr>
<tr>
<td>SGB total</td>
<td>-177.6</td>
<td>-2257.1</td>
<td>-1938.5</td>
<td>496.2</td>
</tr>
</tbody>
</table>
3.4.3 Specificity at the chemical step

The QM/MM [ONIOM (B3LYP:AMBER)] method was used for modeling the chemical step of GlucD, with the QM layer as described in Figure 3 - 36. All the QM/MM energies reported here are relative energies at the [B3LYP/6-311+G(2d,p):AMBER] level. Although there is experimental evidence to support the catalytic mechanism of GlucD (as described in Figure 3 - 34)\textsuperscript{[225-227]}, no computational study on this mechanism has yet been reported. According to QM/MM calculations on two related enzymes, i.e. mandelate racemase (MR)\textsuperscript{[231-233]} and enolase (in the following text, EL means the enzyme enolase, and EN stands for the enolase superfamily)\textsuperscript{[234, 235]}, the catalytic mechanisms of the EN superfamily are not always the same, depending on binding modes of the substrates and the protonation states of the active site residues. MR, EL and GlucD all belong to the EN superfamily\textsuperscript{[224]}, and GlucD shares more similarities with EL in terms of substrate binding mode, in which the carboxylate group of the substrate is a bidentate ligand to the Mg\textsuperscript{2+} (Figure 3 - 36).

According to our calculations on D-glucarate substrate, TS2 is the rate-determining state (Figure 3 – 37), and the obtained energy barrier 15.1 kcal/mol is consistent with the phenomenological activation energy ~15 kcal/mol (At 295.15K, k\textsubscript{cat} value is 35 s\textsuperscript{-1} for D-glucarate\textsuperscript{[226]}). The barrier height for the proton abstraction (PA) step is also significant (TS1, 12.9 kcal/mol, Figure 3 – 37). Although no experimental results are available to verify what the rate-limiting step is, it has been suggested that the PA step is at least partially rate-limiting\textsuperscript{[226]}, because all the enzymes in the EN superfamily share this common step.
3.4 Computational Study on the Stereo-specificity of GlucD (Paper 6)

Figure 3-36. QM region (~100 atoms) in the QM/MM calculations

Figure 3-37. Energy profile (in kcal/mol) along the reaction coordinate for D-glucarate

We thus focus on both transition states (TS1 and TS2) in order to study the specificity at the chemical step. D-glucarate is the best substrate in terms of energies for both TSs (Table 3-16). In TS1 and TS2, the energy difference between D-glucarate and the three non-substrates is in the range of 2-7 kcal/mol (Table 3-16), which is large enough for us to conclude that specificity is encoded in the chemical step. Comparison of the energy profiles of the four ligands can be found in Figure 3-38. Besides the energies of the TSs, we notice that the atomic charges derived from
3.4 Computational Study on the Stereo-specificity of GlucD (Paper 6)

electrostatic potentials (ESP) relate to the specificity at the PA step. In the Michaelis complexes of the four ligands, the charge difference between the proton on C5 and the nitrogen acceptor on H339 ($q_H-q_N$, c.f. Figure 3 - 36) is in qualitatively accordance with the energies of $\text{TS1}$ (Table 3 - 16). As the optimized geometries at the reactive region are very similar for the four ligands, we suggest that the energies of $\text{TS1}$ are mainly affected by the variations of the overall electrostatic environment. As optimizations of the Michaelis complexes are computationally more efficient than optimizations of both Michaelis complex and the transition state, the ESP atomic charges of the Michaelis complex may thus be used for excluding poor substrate candidates when the computational resources are limited.
Table 3 - 16. Relative Energies of TS1 and TS2 for D-glucarate and three non-substrates

<table>
<thead>
<tr>
<th>Ligands</th>
<th>$\Delta E_{TS1}$</th>
<th>$\Delta \Delta E_{TS1}$</th>
<th>$\Delta E_{TS2}$</th>
<th>$\Delta \Delta E_{TS2}$</th>
<th>$q_H - q_N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-glucarate</td>
<td>12.9</td>
<td>0.0</td>
<td>15.1</td>
<td>0.0</td>
<td>1.10</td>
</tr>
<tr>
<td>m-allarate</td>
<td>16.1</td>
<td>3.2</td>
<td>17.3</td>
<td>2.2</td>
<td>0.61</td>
</tr>
<tr>
<td>D-mannarate</td>
<td>15.5</td>
<td>2.6</td>
<td>18.3</td>
<td>3.2</td>
<td>0.98</td>
</tr>
<tr>
<td>D-altrarate</td>
<td>19.3</td>
<td>6.4</td>
<td>20.2</td>
<td>5.1</td>
<td>0.33</td>
</tr>
</tbody>
</table>

$^a$B3LYP/6-311+G(2d,p):AMBER energies, in kcal/mol

$^b$ $\Delta \Delta E_{TS} = \Delta E_{TS[ligand]} - \Delta E_{TS[D-glucarate]}$, in kcal/mol

$^c$ESP charges at B3LYP/6-311+G(2d,p):AMBER level, H and N are described in Figure 3 - 36.

Figure 3 – 38. Comparison of the energy profiles of the four ligands
3.5 Computational Study on the product specificity of OSC (Paper 7)

3.5.1 Introduction

Oxidosqualene-lanosterol cyclase (OSC), which catalyzes the conversion from oxidosqualene to lanosterol, is a key enzyme in the biosynthesis of cholesterol\[^{[236]}\]. OSC inhibitors are drug candidates against hypercholesterolemia\[^{[237-239]}\]. The proposed catalytic mechanism of OSC is shown in Figure 3 - 39\[^{[238]}\]. Protonation of the epoxide ring of oxidosqualene by Asp455 triggers ring-forming reactions, yielding the protosterol cation 4. A series of 1,2-hydride and 1,2-methyl group shifts then occurs, followed by a final deprotonation step leads to the product lanosterol (Figure 3 - 39)\[^{[238]}\]. In the crystal structure of the OSC-lanosterol complex, His232 and Tyr503 (human OSC numbering) are in favorable positions for the final deprotonation step (Figure 3 - 40)\[^{[238]}\]. Mutation of the active site histidine His232 to different amino acids leads to diverse products, three of which are shown in Figure 3 - 40\[^{[240]}\].

Although multiple cationic/olefin intermediates are formed during the catalytic process of OSC, the sole product lanosterol is obtained. We divide the OSC specificity into two stages, i.e. before or after the formation of protosterol cation 4. In the first stage, the 4-ring chair-boat-chair scaffold protosterol cation is the dominant intermediate formed in the wild type OSC, whilst various byproducts from 1-ring to 3-ring are observed in OSC mutants\[^{[240-249]}\]. In the second stage, although the hydride and methyl shifts may results in many different 4-ring products (e.g. P1, P2 and P3), only lanosterol is formed in the wild type OSC.
Figure 3 - 39. Proposed mechanism of OSC$^{[238]}$  

To date, no computational study on the overall mechanism of OSC has been reported, except that the first step (conversion $R \rightarrow 1$) was studied by QM calculations on small models$^{[250]}$. The mechanism of squalene-hopene cyclase (SHC), a closely related enzyme, has been investigated by a number of computational studies$^{[108, 251-257]}$, where the ring formation steps have been discussed in great detail. In the current work, we are interested in the product specificity at the hydride/methyl-shifting stage, which was never studied previously. In particular, two representative mutants were considered, i.e. H232S and H232T, which generate sole product Parkeol ($P_2$) and protosta-12,24-dien-3-beta-ol ($P_3$), respectively (Figure 3 - 40)$^{[240]}$. 

3.5.2 DFT calculations on key intermediates and products

The two DFT methods B3LYP and M06-2X were used to analyze the gas phase energies of the cation intermediates in Figure 3 - 39 and Figure 3 - 40, and the results are summarized in Table 3 - 17. Recently, Kürti et al.\textsuperscript{[251]} studied the energies of various cations involved in the OSC catalysis using quantum mechanics calculations. They suggest that OSC utilizes a nonstop process, and that product specificity of OSC is probably achieved by the stabilization of the cationic intermediates\textsuperscript{[251]}. According to their hypothesis, to selectively produce lanosterol,
3.5 Computational Study on the product specificity of OSC (Paper 7)

intermediate 8 must have lower energy than 6 and 9, (Figure 3 - 40). However, our DFT calculations disapprove their hypothesis. Intermediate 8 was found to be less stable than 6 or 9 in terms of gas phase energy (Table 3 - 17). The gas phase energies of the three products in Figure 3 - 40 were also compared (Table 3 - 18). According to our calculations, the energies of P1 and P3 are close, while P2 seems to be much less stable. Taken together, the gas phase calculations on the key intermediates and products do not favor P1 as the sole product. To investigate the product specificity of OSC, it thus appears that one must take the enzymatic environment into account, rather than simply look at the gas phase energies of the isolated intermediates.

Table 3 - 17. Gas phase energies of various cation intermediates

<table>
<thead>
<tr>
<th>Cationic intermediate$^a$</th>
<th>$E_{B3LYP}^b$</th>
<th>$E_{M06-2X}^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>-9.1</td>
<td>-22.3</td>
</tr>
<tr>
<td>4</td>
<td>-8.5</td>
<td>-36.3</td>
</tr>
<tr>
<td>5</td>
<td>-14.0</td>
<td>-41.3</td>
</tr>
<tr>
<td>6</td>
<td>-16.8</td>
<td>-44.3</td>
</tr>
<tr>
<td>7</td>
<td>-12.8</td>
<td>-40.6</td>
</tr>
<tr>
<td>8</td>
<td>-15.3</td>
<td>-43.5</td>
</tr>
<tr>
<td>9</td>
<td>-17.2</td>
<td>-46.0</td>
</tr>
</tbody>
</table>

$^a$The structures of intermediates 1-9 are described in Figure 3 - 39 and Figure 3 - 40

$^b$relative energies at the B3LYP/BS2//B3LYP/BS1 level

$^c$relative energies at the M062X/BS2//B3LYP/BS1 level

Table 3 - 18. Gas phase energies of the three major products

<table>
<thead>
<tr>
<th>Product$^a$</th>
<th>$E_{B3LYP}^b$</th>
<th>$E_{M06-2X}^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>P2</td>
<td>11.8</td>
<td>11.6</td>
</tr>
<tr>
<td>P3</td>
<td>0.5</td>
<td>-1.0</td>
</tr>
</tbody>
</table>

$^a$product structures are described in Figure 3 - 40

$^b$relative energies at the B3LYP/BS2//B3LYP/BS1 level

$^c$relative energies at the M062X/BS2//B3LYP/BS1 level
3.5.3 **ONIOM calculations on wild type OSC**

The proposed mechanism of OSC (Figure 3 - 39) was subsequently studied by using the QM/MM (ONIOM) approach (Figure 3 - 41). In a recent QM/MM study by Rajamani et al.\textsuperscript{[108]}, the catalytic mechanism of SHC, a related enzyme, was systematically analyzed. The two major findings were 1) the five- to six-membered ring expansion process is unlikely to occur for either C- or D-ring formation in the cyclization reaction; 2) the distribution of two key products, the 6,6,6,5-tetracyclic carbon skeleton and the 6,6,6,6,5-pentacyclic hopanoids, is controlled by kinetics.\textsuperscript{[108]} In the current study on OSC, we are interested in the product specificity at the hydride/methyl-shifting stage after the formation of the 6,6,6,5-tetracyclic intermediate 4. Intermediates 1-8 and the TSs linking these were thus optimized by using ONIOM calculations (Figure 3 - 39). As mentioned above, His232 was set to the HID protonation isomer.

![Energy profile of the proposed mechanism of the wild type OSC (H232 is in the HID form)](image)

**Figure 3 - 41. Energy profile of the proposed mechanism of the wild type OSC (H232 is in the HID form)**

The first step is the protonation of \textbf{R}, which is in concerted with A ring formation, to give 1. As this step has the highest barrier in the catalytic process (23.0 kcal/mol), we suggest that it is the rate-limiting step of OSC. It should be noted that the obtained barrier height 23.0 kcal/mol is
Chapter 3 Result and Discussion

3.5 Computational Study on the product specificity of OSC (Paper 7)

much lower than a previous small model QM study\textsuperscript{[250]} (40 kcal/mol with B3LYP/BS1). The optimized geometry of TSR-1 is shown in Figure 3 - 42a. The B ring formation is relatively easy, whereby intermediate 2 is obtained. The 6,6,6-tricyclic intermediate 3 was found to be unstable, which is consistent with the previous QM/MM study on SHC\textsuperscript{[108]} as well as the fact that 6,6,6-tricyclic products are never observed. The optimized geometry of TS2-4 is shown in Figure 3 - 42b. It should be mentioned that in SHC, the C-ring formation may lead to 6,6,5-tricyclic, 6,6,6,5-tetracyclic and 6,6,6,6,5-pentacyclic intermediates. In the current study on OSC, 6,6,5-tricyclic and 6,6,6,6,5-pentacyclic intermediates were not considered, as our focus is on the product specificity at the hydride/methyl-shifting stage. The conversion of 4 to 8 is fast and exothermic (Figure 3 - 41), whereafter 8 converts to lanosterol.

![Figure 3 - 42. Optimized geometries of a) TSR-1; b) TS2-4](image)

3.5.4 ONIOM calculations on OSC mutants

Intermediate 8 is the most stable one of those shown in Figure 3 - 39 (Figure 3 - 41), which seems to support the hypothesis that product specificity of OSC is achieved by the stabilization of the cationic intermediates\textsuperscript{[240, 246, 247, 251]} However, our QM and QM/MM calculations show that intermediate 9, which is the precursor of Parkeol (P2), is more stable than 8 (the precursor of P1; Table 3 - 19, column 1, 4 and 5). The relative energies of key intermediates 6, 8 and 9 in the wild type OSC and those in the H232S and H232T mutants were thus compared. The trends of the relative energies of these intermediates are the same. The relative energies of the three products
in different mutants were also studied, and the results show that $P_2$ is more stable than $P_1$ (Table 3 - 19). Hence, according to our calculations, $P_2$ is thermodynamically more favorable than $P_1$ for all the systems investigated in this study.

**Table 3 - 19. Relative energies of the key intermediates and products in the QM/MM calculations**

<table>
<thead>
<tr>
<th>Intermediates or Products</th>
<th>Relative Energy$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wild type HID</td>
</tr>
<tr>
<td>6</td>
<td>24.5</td>
</tr>
<tr>
<td>8</td>
<td>16.0</td>
</tr>
<tr>
<td>9</td>
<td>12.1</td>
</tr>
<tr>
<td>$P_1$</td>
<td>0.0</td>
</tr>
<tr>
<td>$P_2$</td>
<td>-6.6</td>
</tr>
<tr>
<td>$P_3$</td>
<td>1.5</td>
</tr>
</tbody>
</table>

$^a$ ONIOM-EE(B3LYP/BS2:AMBER) energies, relative to $P_1$, in kcal/mol

In the previous QM/MM study by Rajamani et al.$^{[108]}$, the product specificity of SHC was found to be controlled by kinetics. We thus investigate whether the product specificity of OSC is determined by kinetics. In the current work, only the barrier heights for the conversions in the last few steps were studied, in order to explain why $P_1$, $P_2$ and $P_3$ are selectively generated in wild type OSC, H232S and H232T mutants, respectively (Table 3 - 20 and Figure 3 - 43). In the wild type OSC and the H232S mutant, the formation of 7 is faster than $P_3$ (Table 3 - 20 and Figure 3 - 43), and thus $P_3$ is not generated. In the H232T mutant, however, the formation of $P_3$ is faster than the formation of 7 (Table 3 - 20 and Figure 3 - 43). This hence explains why $P_3$ is the product in the H232T mutant. Similarly, in the wild type OSC, as the formation of $P_1$ is faster than the formation of 9 (Table 3 - 20 and Figure 3 - 43), $P_1$ is the preferred product, although $P_2$ is thermodynamically more stable.
Chapter 3 Result and Discussion

3.5 Computational Study on the product specificity of OSC (Paper 7)

Figure 3 - 43. Energy profiles of wild type OSC (in black), H232S (in red) and H232T (in green) mutants

Table 3 - 20. Barrier heights for the key conversions that determine product specificity

<table>
<thead>
<tr>
<th>Conversions</th>
<th>Barrier heights$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wild type</td>
</tr>
<tr>
<td></td>
<td>HID</td>
</tr>
<tr>
<td>$6 \rightarrow 7$</td>
<td>6.4</td>
</tr>
<tr>
<td>$6 \rightarrow P3$</td>
<td>9.0</td>
</tr>
<tr>
<td>$8 \rightarrow 9$</td>
<td>5.8</td>
</tr>
<tr>
<td>$8 \rightarrow P1$</td>
<td>0.8</td>
</tr>
<tr>
<td>$9 \rightarrow P2$</td>
<td>2.0</td>
</tr>
</tbody>
</table>

$^a$ONIOM-EE(B3LYP/BS2:AMBER) energies, in kcal/mol
Chapter 4 Conclusions

4.1 SrtA

The structures of the *L. monocytogenes* SrtA-substrate complexes in both inactive and active forms were studied using homology modeling, docking and MD simulation techniques. We found that when the SrtA enzyme converts to the active form, Arg 197 will shift its hydrogen bonds from the LP residues to the TG residues of the LPXTG substrate. Similar motion was also observed in the *S. aureus* sortase A within 20 ns MD simulations.

The catalytic mechanism of SrtA has been systematically studied using the *S. aureus* SrtA. We suggest that the RPM is the most likely mechanism for SrtA. QM/MM calculations and charge deletion analysis show that Arg197 stabilizes the thiolate anion in the Cys184-His120 catalytic dyad (the precursor of the acylation step), which is essential for SrtA catalysis. Arg197 is important for substrate binding, but the Arg197 motion does not affect the binding affinity. On the basis of the MD simulations and QM/MM calculations of the relative stabilities of the inactive and active Cys-His dyad, we suggest that the general base mechanism is unlikely to occur. IRC-like optimization of TS yields the final product, due to the barrier heights for NA-Cys and the cleavage of the peptide bond being very small (or lacking) when the PT-His step is completed. Interestingly, with the presence of Arg197, the barrier height of acylation increases, which disproves the previous hypothesis that Arg197 may stabilize the TS by forming an oxyanion hole. To compensate for this, Thr183 significantly reduces the acylation barrier via electrostatic interactions, which explains the unknown role of Thr183 in SrtA catalysis. All results are consistent with available experimental data, and provide a detailed explanation for the observed features of the SrtA enzymatic mechanism.

4.2 MHPCO

DFT (MPWB1K) calculations were performed on a small model (5HN + FADHOOH) to study the hydroxylation and ring-opening mechanism of MHPCO. In the hydroxylation, a hydroxyl group is transferred from FADHOOH to the substrate through aromatic electrophilic substitution,
which is similar to the other aromatic hydroxylases. A barrier height of 33.8 kcal/mol is obtained for this step. The hydroxyl proton is spontaneously transferred back between the hydroxylated substrate and the remaining FADHO$^-$ moiety. MPWB1K calculations show that complex III is more stable than II. The ring opening occurs inside the enzyme active site rather than in water because in water the rearomatization is predicted to take place immediately, without any barrier, to give an aromatic product. This has, however, not been detected as the product of MHPCO. The instability of the hydroxylated intermediate of MHPCO is the main reason that the acyclic products are formed. Two possible ring-opening pathways are found. In pathway A, the ring opens directly to give the trans configuration of the ketene intermediate. In pathway B, the cis configuration is obtained by way of an epoxy transition state.

DFT (B3LYP) calculations were then performed on the MHPCO active-site atoms (S$_I$) alongside ONIOM calculations on the full enzyme (S$_II$), in order to understand the catalytic mechanism and the effect of the protein on the reaction. The S$_I$ model calculations show that the active-site residues Arg211 and Tyr223 have minor effect on the reaction. The peptide bond of Pro295-Ala296, the side chain of Tyr82, and several crystal water molecules can affect the reaction energy profile, and the atoms of these residues were added to the QM layer in the ONIOM calculations.

According to both S$_I$ and S$_II$ calculations, the peptide bond of Pro295-Ala296, which acts as a hydrogen bond acceptor, stabilizes the reactant more than TS in the hydroxylation step. The hydrogen bond between Tyr82 and the substrate changed insignificantly for all pathways investigated, suggesting that Tyr82 is likely to affect the reaction barriers via electrostatic interactions instead of hydrogen bond. Comparing the S$_I$ model calculations with the S$_II$ model, we found that the MM layer has minor effect on the hydroxylation step, but can affect the ring-opening reactions. The activation energies obtained for the hydroxylation step are 32.2 kcal/mol for ONIOM-1(B3LYP/BS2:Amber) model and 31.5 kcal/mol for the QM model (B3LYP/BS2), which are much higher than the experimental value (16.7 kcal/mol). Therefore, additional factors must be present that were not considered in our system, such as the dynamics of the enzyme, and that a single conformation (crystal structure) of the protein may not be sufficient to model the MHPCO reaction. In addition, our calculations show that the ring-opening pathway B, in which
an epoxy like transition state is formed, is much more favored than the direct C2-C3 cleavage pathway A, in the presence of the protein environment.

Eight different QM/MM partitioning schemes have been used to study the corresponding reactions, and we conclude that the energy barriers are very sensitive to the QM/MM partitioning scheme, especially for the treatment of the peptide bond in Pro295-Ala296. When this peptide bond is added to the QM region, the barrier heights of the hydroxylation and the ring-opening pathway B increase, while the barrier for the ring-opening pathway A decreases.

### 4.3 PBGS

The chemical step of PBGS is studied by using extensive and systematic QM/MM calculations. We propose that A-site ALA can form a Schiff base with either the A-site lysine or the P-site ALA, and that the Schiff base between A-site ALA and A-site lysine might not be necessary for the ring closure. Our calculations show that Mechanisms 1 and 4 are the most likely ones among all tested. Mechanism 2, which is favored in DFT-only calculations, is shown unlikely to occur. According to our calculations, the preferred protonation state is protonation state 1 in which both Schiff base nitrogens are protonated. The question whether C3-C4 or C5-N1 forms first in the reaction might be system-dependent, given that two competing mechanisms (Mechanism 1 and 4) exist and that the ring closure stage is presumably not rate-limiting. The rate-limiting barrier of ~22 kcal/mol for Mechanisms 1 and 4 is higher than that obtained from experimental data ~18 kcal/mol, and possible explanations are that neither entropic nor quantum tunneling effects are considered in this work, and that deprotonation of the A-site ALA might occur already in the substrate binding process where large conformational changes are involved. Unfortunately, the current methodology is not able to evaluate such changes. MD simulations may be employed to study the conformational changes induced by the binding of A-site ALA, but are obviously non-trivial. To better understand the PBGS mechanism, we propose that future focus should be on the binding process of A-site ALA and the deprotonation of the C4 protons.
4.4 GlucD

Docking and QM/MM calculations are performed to study the GlucD stereo-specificity at the non-reactive regions. Our results show that GlucD specificity is encoded in both the substrate binding and the chemical steps. We emphasize that enzyme substrate specificity is a two parameter \( \frac{k_{\text{cat}}}{k_{M}} \) problem, and computational studies should investigate properties for both steps (possibly using mathematics to integrate these properties in the future). The method presented here may be used to assist in enzyme function prediction.

4.5 OSC

The catalytic mechanism and the product specificity of OSC were investigated by using QM/MM calculations. The first step, i.e. protonation of the epoxide ring of oxidosqualene, was found to be rate-limiting. According to our DFT and ONIOM calculations, intermediate 9, which is the precursor of parkeol, is more stable than 8 (the precursor of lanosterol). Therefore, we suggest that the relative stability of the cationic intermediates may to some extent contribute to the OSC product specificity, but is not the main factor that governing this. The barrier heights at the hydride/methyl-shifting stage were found to be consistent with the product profiles of different OSC mutants. Hence, we suggest that the product specificity of OSC is likely to be controlled by kinetics. Mutagenesis results in changes of the electrostatic environment of the enzyme, and thus affects the reaction barriers at the hydride/methyl-shifting stage. When the base group deprotonates the carbon cation intermediate, product is generated irreversibly. Our calculations are consistent with all the experimental data available.
References


References


Appendix

Appendix 1. Computational details

A1.1 Paper 1

Forcefield and programs

The Amber force fields (Amber 99 and Amber 03) were used for all the MM energy calculations, geometry optimizations, and MD simulations, as well as homology modeling. Homology modeling, docking simulations and binding energy calculations with the molecular mechanics generalized born/volume integral (MM/GBVI) implicit solvent method were performed using the MOE software (Version 2009.10; Amber 99 force field). For further MD simulations, the GROMACS software (Version 4.0.4; Amber 03 force field) was used. The Gaussian09 software (Rev. A.02; Amber 03 force field) was used for geometry optimizations and energy calculations on selected snapshots. All figures were generated using MOE.

Homology modeling and docking

The catalytic domain (From Tyr 75 to Lys 222) of L.monocytogenes SrtA was modeled by homology modeling. The protein sequence of L. monocytogenes SrtA was obtained from the NCBI database (GeneID: 7704527). The crystal structure of Streptococcus pyogenes (PDB code 3FN5, 1.5 Å resolution) was chosen as the template for the homology modeling calculations, based on the multiple sequence alignment. 1250 intermediate homology models (25 backbone models × 50 side chain models) were generated, and the best-scoring one, determined by the GB/VI scoring method which compares the electrostatic salvation energy, was taken to be the final model. The protonation of the final protein model was determined by Protonate 3D in MOE (pH = 7.0). All Arg and Lys residues were assigned to their protonated forms and Asp and Glu to their deprotonated forms. All His residues were set to be neutral with ND1 protonated (the active site histidine, His 127, was modified to the protonated form, and Cys 188 changed from neutral to deprotonated, respectively, in the MD simulations on the active enzyme-substrate complex).
Appendix 1. Computational details

To mimic the LPXTG motif in the *L. monocytogenes* substrate protein Internalin A (InlA), an LPTTG (in InlA, X = threonine) pentapeptide, which was acetylated and amidated at the N-terminal and C-terminal, respectively, was built, optimized and docked into the active site of the *L. monocytogenes* SrtA homology model. The triangle matcher placement methodology in which ligand triplets of atoms are aligned on triplets of alpha spheres was used in the docking simulations. The poses were ranked by a London dG score, which estimates the binding free energy of the ligand. The best pose was used for the subsequent MD simulations.

**MD simulations**

30ns MD simulations were performed on the inactive and active forms (the Cys-His catalytic dyad in their neutral and zwitterionic forms, respectively) of the SrtA-LPTTG complexes. The SrtA-LPTTG complex was solvated in a periodic box with a buffer distance of 7.0 Å, containing 4103 TIP3P water molecules. Four Na\(^+\) ions were added at random positions using the genion command in GROMACS, to satisfy the electroneutrality condition. The whole system (~15,000 atoms) was energy minimized by steepest descent (200 steps) to remove close contacts. Before the actual MD simulation, a position restrained simulation (500 ps duration, 1.0 fs time step, NPT ensemble, T = 298 K, P = 1 bar) was performed, to enable the water molecules to reach more favorable positions. The particle-mesh Ewald (PME) summation was used for long-range electrostatics. A 10 Å cutoff was used for both Coulomb and Lennard-Jones interactions. The temperature and pressure was controlled through the Berendsen coupling algorithm, with the time constants 0.1 ps for temperature and 1.0 ps for pressure coupling. All bond lengths were constrained using the LINCS algorithm. In the actual MD simulations (30ns duration, 1.0 fs time step, NVT ensemble, T = 298 K), the pressure coupling was removed, and other parameters were the same as those in the position restrained simulation. For the MD simulation on the active SrtA-LPTTG complex, the last snapshot of the 30ns MD simulation of the inactive complex was used as the starting structure, with Cys 188 and His 127 being modified to their zwitterionic pair form. The Ca-RMSD of the inactive and active SrtA-LPTTG complexes versus simulation time were plotted to ensure the systems were sufficiently equilibrated. In the root mean square fluctuation calculations, trajectories were collected after 3ns equilibration.
Interaction energy calculations

Comparison of the enzyme-substrate interactions in the inactive and active Michaelis complexes is of interest. The interaction energy (IE) of the enzyme-substrate complex was investigated using MM (Amber 03 Force field) and semiempirical methods (AM1, PM3, and PM6), as well as the MM/GBVI implicit solvent method. The IE is defined as the energy difference between the enzyme-substrate complex (E-S), and individual enzyme (E) and substrate (S), Equation A - 1.

Equation A - 1

\[ \text{IE} = E_{E-S} - (E_E + E_S) \]

To eliminate the kinetic energy, geometry optimizations were performed with the Amber 03 force field, on snapshots from the MD trajectories (with solvent molecules and Na\(^+\) ions). Single point energy calculations using the Amber force field, the different semiempirical methods, and the MM/GBVI solvent model were then performed on the optimized geometries of the full enzyme-substrate complexes with all water molecules and Na\(^+\) ions removed, and the average values were calculated.

A1.2. Paper 2

Preparation of the Michaelis Complex

The first snapshot of the NMR structure of the S. aureus LPXT*-SrtA intermediate (residues 62-206, PDB code: 2KID) was used as the initial model in the current study. In order to recover the Michaelis complex, an Nme-LPATG-Ace peptide was built, optimized and superposed on the LPAT* moiety of this NMR structure, using the Flexible Alignment module in MOE.

MD Simulations

MD simulations were performed to investigate if the Arg197 motion discovered in Paper 1 also applies to the S. aureus case. The Amber 03 force field was used for all the MD simulations. The Cys184-His120 dyad was initially set to the active (ion pair) form, while the protonation of the remaining residues was kept the same as in the NMR structure. The obtained structure was
Appendix

Appendix 1. Computational details

solvated by TIP3P water molecules in a periodic box with a buffer distance of 8.0 Å. A Cl⁻ ion was randomly added to satisfy the electroneutrality condition, using the genion module in GROMACS (version 4.0.4) software. The whole system (~20000 atoms) was energy minimized by steepest descent (200 steps) to remove close contacts. A position restrained simulation (2 ns duration, 1.0 fs time step, NPT ensemble, T = 298 K, P = 1 bar) was performed, to enable the water molecules to reach more favorable positions. The leapfrog algorithm was used for the integration of the classical Newton equations of motion. PME summation was used for long-range electrostatics. A 10 Å cutoff was used for both Coulomb and Lennard-Jones interactions. The temperature and pressure were controlled through the Berendsen coupling algorithm, with the time constants 0.1 ps for temperature and 1.0 ps for pressure coupling. All bond lengths were constrained using the LINCS algorithm. In the production run (20 ns duration, 1.0 fs time step, NPT ensemble, T = 298 K, P = 1 bar), the temperature was controlled using the Nose-Hoover thermostat with a time constant of 0.1 ps and the pressure was controlled using the Parrinello-Rahman barostat with a time constant 1.0 ps. Other parameters were similar to the position restrained simulations.

Prior to the production run, we performed 1 ns MD simulations using the production run parameters and recorded snapshots at 200 ps, 400 ps, 600 ps, 800 ps, and 1 ns as initial structures for further production runs. After the production run of the active form SrtA-LPXTG complex, we modified the protonation states of the Cys-His catalytic dyad to the inactive form and performed another 20 ns MD simulation in order to investigate if the Arg197 side chain can switch to the LP moiety (corresponding to the reverse simulation of Paper 1). 20 ns reversibility tests were also performed to confirm the Arg197 motion. All MD simulations were performed using GROMACS.

**QM/MM (ONIOM) Calculations**

Representative MD snapshots were further energy minimized with the Amber 03 force field. In the QM/MM calculations, MM parameters for the active Cys-His catalytic dyad were used. Four main systems were investigated, including the last snapshot of the active form from the 20 ns MD trajectory (called Position1), the last snapshot of the inactive form from the 20 ns MD trajectory (called Position2), the R197A mutation of Position1 (called Position1-R197A), and the T183A
mutation of Position1 (called Position1-T183A). Mutations were readily obtained using the MOE program. It should be noted that Position1 and Position2 refer to the preferred positions of Arg197 in the active and inactive SrtA-substrate complexes, whereas active and inactive forms refer to the zwitterionic and neutral forms of the Cys184-His120 catalytic dyad. For the R197A and T183A systems, position restrained simulations were performed prior to the MM optimizations, to equilibrate the surrounding water molecules.

The substrate-free systems for Position1, Position2, and Position1-R197A were also studied for comparison of the relative stabilities of the inactive and active forms of the enzyme. Position restrained simulations were performed after the removal of the substrates for the three substrate-free systems.

A two-layer ONIOM (QM/MM) scheme was used for the calculations of the SrtA catalytic mechanism. The standard DFT method B3LYP and the two basis sets 6-31G(d) (BS1) and 6-311+G(2d,p) (BS2) were used for treating the QM region. Geometries of intermediates and TSs were optimized using the ONIOM(B3LYP/BS1:Amber) method with the ONIOM-ME (mechanical embedding) scheme, followed by frequency calculations at the same level of theory. Atoms far away from the substrate carbonyl group (> 15 Å) were kept fixed during the optimizations, and the normal modes of these atoms were excluded from the frequency calculations with the Gaussian 09 keyword `notatoms`. Single-point energy calculations with the ONIOM(B3LYP/BS2:Amber) method were performed on the optimized geometries with the ONIOM-EE (electronic embedding) scheme. Geometries were not optimized with the ONIOM/EE scheme due to the high computational cost and poor convergence. All ONIOM calculations were performed using the Gaussian 09 software (Rev. A.02).

The QM/MM partitioning is described in Figure 3 - 9. Three different partitioning schemes were used for studying the acylation step at Position1. Model-1 (blue) contains the ATG moiety of the substrate as well as the side chains of Cys184 and His120.

Model-2 (blue + red) added Arg197 and two solvent molecules (SOL-1 and SOL-2) to the QM region, and Model-3 further added Thr183, the Thr183-Cys184 backbone, and another seven
solvent molecules to the QM part. Model-1 was also used for the calculations on the Position2, Position1-R197A, and Position1-T183A systems.

Charge Deletion Analysis and Binding Affinity Calculations

Charge deletion analysis was used to investigate the influence of individual residues as well as solvent molecules on the enzymatic catalysis. Deletion of atomic charges of a certain residue in the MM region can result in energy changes in the QM region ($\Delta E_{QM}$). If the QM energy changes of a reactant [$\Delta E_{QM}(R)$] and a transition state [$\Delta E_{QM}(TS)$] are different, the reaction barrier is affected. The electrostatic effect of a certain residue on the reaction barrier [$\Delta \Delta E_{QM}(R \rightarrow TS)$] can be defined as in Equation A - 2. A positive sign of $\Delta \Delta E_{QM}(R \rightarrow TS)$ indicates that the effect of this residue is to lower the barrier height, whereas a negative sign implies that it plays the opposite role. Similarly, the relative stabilities of a reactant and a product (for instance, the relative stabilities of the inactive and active forms of the enzyme) can also be evaluated; see Equation A - 3. All the $\Delta \Delta E_{QM}$ values reported in this study are ONIOM-EE(B3LYP/BS2:Amber) single point energies. To reduce computational cost, every three adjacent residues were combined into a group, and charge deletion analysis was initially used for identifying important groups ($\Delta \Delta E_{QM} > 1 \text{ kcal/mol}$). On the basis of this initial screening, the individual residues in the important groups, as well as all active site residues (not frozen in the ONIOM calculations), were subsequently analyzed.

Equation A - 2

$$\Delta \Delta E_{QM}(R \rightarrow TS) = \Delta E_{QM}(TS) - \Delta E_{QM}(R)$$

Equation A - 3

$$\Delta \Delta E_{QM}(R \rightarrow P) = \Delta E_{QM}(P) - \Delta E_{QM}(R)$$

The binding affinity of the substrate was evaluated using the MM/GBVI approach in the MOE software, which includes implicit solvent effects. The average value of binding affinities from 10 snapshots (with all solvent molecules removed) in the last 1 ns MD trajectory was evaluated (for the Position1-R197A system, only the last snapshot of the position restrained simulation was used).
A1.3. Paper 3

The DFT method MPWB1K was used to study the reaction pathways. Intermediates and transition states were first optimized in vacuo at the B3LYP/6-31G(d) level. Further geometry optimizations by using MPWB1K/6-31+G(d,p) were performed on the geometries of the previous step, followed by frequency calculations at the same level of theory. Transition states were tested by using IRC calculations to make sure that they connected the correct reactants and products in each step. The IEFPCM with UAHF radii was used to optimize geometries and calculate frequencies for all conversions that occur in aqueous solution (ε = 78.3). Single-point energy calculations were performed on the optimized structures by using the higher basis set 6-311+G(2d,2p) in the proton affinity calculation of the hydroxylated intermediates. All calculations were performed by using the Gaussian 03 program.

A1.4. Paper 4

Preparation of the Model System

Two model systems were investigated, i.e. S_I and S_II. For S_I, all the intermediates and TSs in were optimized using the B3LYP functional. Through the QM calculations on S_I, we evaluated the effect of individual active site residues on the enzyme reaction and determined which residues to include in the QM part of the ONIOM calculations.

The second model system (S_II) is the major system studied and contains the full enzyme, substrate, and crystal water molecules (6900 atoms). The S_II system was initially optimized with the Amber force field and then studied with the ONIOM-(DFT:Amber) methodology using different QM/MM partitioning schemes. While the B3LYP functional was the major QM method in the ONIOM calculations, the M062X functional was also used for comparison.

Computational Details for the S_I System: Active-Site Residue Decomposition Analysis

Prior to the ONIOM calculations, we performed QM calculations on the active-site atoms. The Cα atoms of the active-site amino acid side chains were kept frozen during the optimizations. Frequency calculations were performed on the optimized geometries at the same level of theory.
Appendix

Appendix 1. Computational details

Transition states were tested through IRC calculations to make sure that they connected the correct reactants and products in each step. To evaluate the energetic contribution of a certain active-site residue to the reaction, we removed that residue from $S_I$ (or set those atoms as dummy atoms) and recalculated the single-point energy without geometry optimizations, so as to compare the difference in energy profile between the $S_I$ system and the system without that residue ($S_I\text{-Res}$). The mean unsigned error (MUE) between the relative energies of $S_I$ and $S_I\text{-Res}$ was used to assess the importance of the residue to the reaction energy profile. It should be noted that this strategy has not been previously reported.

**Computational Details for the $S_{II}$ System: ONIOM Calculations on the Full Enzyme**

The $S_{II}$ system was initially optimized with the Amber 03 force field. Missing parameters of the substrate MHPC and the cofactor FADHOOH were derived by Antechamber, and the partial atomic charges were assigned using the AM1-BCC charge method. The ONIOM-ME scheme was used for geometry optimizations and frequency calculations, and the ONIOM-EE scheme was used for single-point energy calculations on ONIOM-ME optimized geometries. Atoms far away from the substrate (MHPC) atoms (>15 Å) were kept fixed during the optimization. Eight different QM/MM partitioning schemes were used, and the largest QM part contains 96 QM atoms (Table 3 – 7). For the ONIOM-1 calculations, both B3LYP and M062X methods were used for the QM layer, while for the other partitioning schemes, only the B3LYP functional was used to treat the QM part.

**A1.5. Paper 5**

The crystal structures of the yeast PBGS-PBG* intermediate (PDB code 1OHL) and the *Toxoplasma gondii* PBGS-PBG complex (PDB code 3OBK) were obtained from the RCSB Protein Data Bank. Monomers of the two structures were used. Hydrogen atoms were added using the MOE software (Version 2010. 10). Protonation states of the active site atoms are summarized in Figure 3 - 26. The complex with charged Lys263 is called Protonation State 1 and that with Lys263 neutral is called Protonation State 2. In yeast PBGS, the three active site cysteines are set to the deprotonated form, as the Zn-S distances in the crystal structure suggest this pattern (the Zn-S distances in 1OHL are 2.33 Å, 2.33 Å and 2.30 Å, respectively). The Zn-S
distances (2.25–2.35 Å) obtained from our QM/MM optimizations are consistent with those found in the crystal structure.

The QM region in the QM/MM calculations of the two systems is described in Figure 3 - 26. Starting from the crystal structures of the PBGS-PBG* and PBGS-PBG complexes, the different Michaelis complexes were traced backwards. The results for the forward and reverse reactions are consistent. The systems were initially optimized using the AMBER 03 force field. Missing parameters of PBG were derived using Antechamber and the partial atomic charges were assigned using the RESP charge method. Geometries were optimized with the ONIOM-EE (B3LYP:AMBER) method. The SDD basis set was used for Zn$^{2+}$, and 6-31G(d) for the other atoms. Atoms far away from the PBG atoms (>$10$ Å) were held fixed during the QM/MM optimizations. Frequency calculations were performed on only the QM atoms (using the ‘geom.= readfreeze’ option in Gaussian 09) to confirm the correct normal modes of the imaginary frequency in those TSs. Full frequency calculations were not performed due to shortage of memory. Single-point energy calculations were performed on the optimized geometries with a larger basis set (BS3 or BS4; Table 3 - 9). All the ONIOM calculations were performed using the Gaussian 09 software (Rev. B.01).

A1.6. Paper 6

Docking

The Schrödinger software and the OPLS-AA force field are used for all the docking and binding affinity calculations. The crystal structure of GlucD (1ECQ) is obtained from the PDB data bank, and is prepared by the Protein Preparation Wizard. Ligands are prepared by the LigPrep module. Docking calculations are performed using Glide module as well as the induced fit docking protocol. In Glide docking, both the Standard Precision (SP) and the Extra Precision (XP) modes are used, and 10 poses for each mode are retained for further binding affinity calculations. In the IFD calculations, the XP mode and default parameters are used. MM/GBSA calculations are performed on the top 10 docking poses using the Prime module. The binding affinities for the reactive poses, i.e. the poses whose reactive regions are similar to the ligand in the crystal structure of GlucD, are shown in Table 3 - 15.
Appendix 1. Computational details

**QM/MM calculations**

The Gaussian 09 software (Rev. B.01) is used for all the QM/MM calculations. Michaelis complexes are obtained by superimposing the prepared ligands to the substrate analog 4-deoxy-D-glucarate in the crystal structure (1ECQ). The systems were initially optimized using the AMBER 03 force field. Ligand parameters were derived by Antechamber and the partial atomic charges were assigned using the RESP charge method. A two-layer ONIOM (QM/MM) scheme with the electronic embedding (EE) is used. Geometries were optimized at ONIOM-EE (B3LYP/6-31G*::AMBER) level of theory. Atoms far away from ligand atoms (>10 Å) were kept fixed during the QM/MM optimizations. Frequencies calculations were performed to ensure that each transition state has only one imaginary frequency. Single-point energy calculations were performed on the optimized geometries with a larger basis set ONIOM-EE (B3LYP/6-311+G(2d,p)::AMBER). ESP charges are calculated by using the (pop = MK) keyword in Gaussian 09.

**A1.7 Paper 7**

**DFT calculations**

The DFT functionals B3LYP and M06-2X were used to study the energies of the key intermediates and products. Geometries were optimized in vacuo using B3LYP with the 6-31G(d) basis set (BS1), followed by frequency calculations at the same level of theory. Single-point energy calculations were performed on the optimized structures using the higher basis set 6-311+G(2d,p) (BS2). All DFT calculations were performed using the Gaussian 09.

**ONIOM calculations**

The crystal structure of the OSC-lanosterol complex (PDB code 1W6K) was obtained from the RCSB Protein Data Bank. Hydrogen atoms were added using the Molecular Operating Environment (MOE) software (Version 2010.10). Lanosterol, Asp455, Tyr503 and His232 were included in the QM region in the QM/MM calculations. In the OSC-lanosterol complex, Asp455 was set to the deprotonated form, and Tyr503 was set to the protonated form (in the corresponding reactant complex, Asp255 is protonated and Tyr503 is deprotonated). Three
different protonation states for His232 were considered in the wild type OSC, i.e. HID, HIE and HIP (the HID model is favored in the MOE protonation 3D prediction). It should be noted that HID232 might be the base for the deprotonation step for the wild type OSC, but in the two mutants in Figure 3-39, Tyr503 is the only base suitable for deprotonation. As the deprotonation by Tyr503 is easier than that by HID232 (data not shown), we only discuss the Tyr503 residue as the base for deprotonation. Intermediates and TSs in the proposed mechanism from the product complex (OSC-lanosterol) to the reactant complex (OSC-oxidosqualene) were optimized using the wild type OSC (HID form). Key intermediates 6-9 (Figure 3-38) were then optimized using the H232S and H232T mutants, as well as HIE and HIP forms of the wild type OSC. The mutants were built using a similar approach as previously reported (Paper 4).

The ONIOM method was used for all QM/MM optimizations. The systems were initially optimized using the AMBER 03 force field. Missing parameters of lanosterol were derived using Antechamber with the RESP charge method. A two-layer ONIOM (QM/MM) scheme with electronic embedding (EE) was used. Geometries were optimized at the ONIOM-EE (B3LYP/BS1:AMBER) level. Atoms far away from the lanosterol atoms (>10 Å) were held fixed during the QM/MM optimizations. Frequency calculations were performed on the QM atoms only (using the ‘geom.= readfreeze’ option in G09) to confirm the correct normal modes of the imaginary frequency in those TS’s. Full frequency calculations were not performed due to shortage of memory. Single-point energy calculations were performed on the optimized geometries with ONIOM-EE (B3LYP/BS2:AMBER). All ONIOM calculations were performed using the Gaussian 09 software.
Paper 1
**Structural Changes of *Listeria monocytogenes* sortase A: A key to understanding the catalytic mechanism**

Boxue Tian and Leif A. Eriksson*

School of Chemistry, National University of Ireland, Galway, Ireland

**ABSTRACT**

*Listeria monocytogenes* is one of the most virulent foodborne pathogens. *L. monocytogenes* Sortase A (SrtA) enzyme, which catalyzes the cell wall anchoring reaction of the leucine, proline, X, threonine, and glycine proteins (LPXTG, where X is any amino acid), is a target for the development of antilisteriosis drugs. In this study, the structure of the *L. monocytogenes* SrtA enzyme-substrate complex was obtained using homology modeling, molecular docking and molecular dynamics simulations. Explicit enzyme-substrate interactions in the inactive and active forms of the enzyme were compared, based on 30ns simulations on each system. The active site arginine (Arg 197) was found to be able change its hydrogen donor interactions from the LP backbone carbonyl groups of the LPXTG substrate in the inactive form, to the TG backbone carbonyls in the active form, which could be of importance for holding the substrate in position for the catalytic process.


Key words: *Listeria monocytogenes*; sortase A; arginine; homology model; inhibitor design.

**INTRODUCTION**

*Listeria monocytogenes* is a Gram-positive bacterium and one of the most virulent foodborne pathogens. Infection by *L. monocytogenes* causes listeriosis, with clinical manifestations including septicemia, meningitis, and many other lethal diseases. In *L. monocytogenes* cells, the leucine, proline, X, threonine, and glycine (LPXTG, where X denotes any amino acid) surface proteins play vital roles in pathogenicity. This type of proteins has an LPXTG motif near the carboxyl terminal followed by a hydrophobic region and a positively charged tail. The LPXTG proteins are covalently anchored to the bacterial cell wall and are capable of adhering to and invading the host cells. The cell wall anchoring reaction of LPXTG proteins is a transpeptidation that is catalyzed by the enzyme sortase A (SrtA). SrtA cleaves the peptide bond between the threonine and glycine (TG) residues in the LPXTG motif and links the threonine to the amino group of the pentaglycine cell wall crossbridge, thus forming the cell wall attached protein (Fig. 1).

It has been shown that inactivation of *L. monocytogenes* SrtA will significantly reduce the virulence of the pathogens. Moreover, as SrtA is not required for the cell growth, SrtA inhibitors will unlikely lead to development of drug resistance. Inhibitors against *L. monocytogenes* SrtA are hence very promising candidates of antilisteriosis drugs. In fact, SrtA has been proposed to be the universal target for therapeutic agents against Gram-positive bacteria. The *Staphylococcus aureus* SrtA is so far the most studied SrtA enzyme, and inhibitors against *S. aureus* SrtA have been reported. Kinetic studies on *S. aureus* have shown that the SrtA enzyme utilizes a ping-pong mechanism and that the transpeptidation involves two steps, acylation and deacylation. Three active site residues are found to be indispensable to the transpeptidation reaction, namely the cysteine-histidine catalytic dyad (Cys 184 and His 120; *S. aureus* numbering) and the active site arginine (Arg 197; *S. aureus* numbering).

In the acylation step, the thiolate sulfur of the active site cysteine attacks the substrate threonine carbonyl carbon and cleaves the peptide bond, forming the LPXT-enzyme intermediate. In the deacylation step, the second substrate pentaglycine is attached to the LPXT moiety, forming the product. The active site histidine is proposed to act as a general acid and base, which protonates and deprotonates the nitrogen in the peptide bond in the acylation and deacylation reactions, respectively. Mutation studies on the active site arginine showed that this residue is as important for the catalytic reaction as the Cys-His catalytic dyad. It was suggested that this arginine residue plays dual roles through (i) binding of the substrates via...
hydrogen bond donation, and (ii) stabilizing the tetrahedral intermediate formed during the catalytic reaction.\textsuperscript{25}

Limited knowledge of the structure of the SrtA enzyme-substrate complex is currently available. Zong \textit{et al.}\textsuperscript{26} reported a crystal structure of \textit{S. aureus} SrtA with a bound substrate, in which the active site cysteine was mutated. In a recent NMR structure of \textit{S. aureus} SrtA,\textsuperscript{25} the modified tetrapeptide LPAT* was attached to the active site cysteine through a disulfide bond, mimicking the short lived LPXT-enzyme intermediate. Obviously, this structure is not the correct enzyme-substrate complex. In addition, it was shown that the enzyme is active only when the Cys-His catalytic dyad exists in the zwitterionic form.\textsuperscript{17} As less than 1% of the enzyme exists in this form,\textsuperscript{17} it is difficult to obtain the structure of the active enzyme-substrate complex by using either X-ray diffraction or NMR techniques. To this end, theoretical methods are key tools to characterize the structure and dynamics of the active SrtA-LPXTG complex.

**COMPUTATIONAL DETAILS**

**Forcefield and programs**

The Amber force fields (Amber 99 and Amber 03)\textsuperscript{27,28} were used for all the molecular mechanics (MM) energy calculations, geometry optimizations, and molecular dynamics (MD) simulations, as well as homology modeling. Homology modeling, docking simulations and binding energy calculations with the molecular mechanics generalized born/volume integral (MM/GBVI) implicit solvent method\textsuperscript{29} were performed using the molecular operating environment (MOE) software\textsuperscript{30} (Version 2009.10; Amber 99 force field). For further MD simulations, the GROMACS software\textsuperscript{31} (Version 4.0.4; Amber 03 force field) was used. The Gaussian09 software\textsuperscript{32} (Rev. A.02; Amber 03 force field) was used for geometry optimizations and energy calculations on selected snapshots. All figures were generated using MOE.

**Homology modeling and docking**

The catalytic domain (From Tyr 75 to Lys 222) of \textit{L. monocytogenes} SrtA was modeled by homology modeling. The protein sequence of \textit{L. monocytogenes} SrtA was obtained from the NCBI database (GeneID: 7704527). The crystal structure of \textit{Streptococcus pyogenes}\textsuperscript{33} (PDB code 3FN5, 1.5 Å resolution) was chosen as the template for the homology modeling calculations, based on the multiple sequence alignment (Supporting Information S1). 1250 intermediate homology models (25 backbone models × 50 side chain models) were generated, and the best-scoring one, determined by the GB/VI scoring method\textsuperscript{29} which compares the electrostatic solvation energy, was taken to be the final model. The protonation of the final protein model was determined by Protonate
3D in MOE (pH = 7.0). All Arg and Lys residues were assigned to their protonated forms and Asp and Glu to their deprotonated forms. All His residues were set to be neutral with ND1 protonated (the active site histidine, His 127, was modified to the protonated form, and Cys 188 changed from neutral to deprotonated, respectively, in the MD simulations on the active enzyme-substrate complex; see below).

To mimic the LPXTG motif in the L. monocytogenes substrate protein Internalin A (InlA), an LPTTG (in InlA, X = threonine) pentapeptide, which was acetylated and amidated at the N-terminal and C-terminal, respectively, was built, optimized and docked into the active site of the L. monocytogenes SrtA homology model. The triangle matcher placement methodology in which ligand triplets of atoms are aligned on triplets of alpha spheres was used in the docking simulations. The poses were ranked by a London dG score, which estimates the binding free energy of the ligand. The best pose was used for the subsequent MD simulations.

**MD simulations**

30ns MD simulations were performed on the inactive and active forms (the Cys-His catalytic dyad in their neutral and zwitterionic forms, respectively) of the SrtA-LPTTG complexes. The SrtA-LPTTG complex was solvated in a periodic box with a buffer distance of 7.0 Å, containing 4103 TIP3P water molecules. Four Na\(^+\) ions were added at random positions using the genion command in GROMACS, to satisfy the electroneutrality condition. The whole system (15,000 atoms) was energy minimized by steepest descent (200 steps) to remove close contacts. Before the actual MD simulation, a position restrained simulation (500 ps duration, 1.0 fs time step, NPT ensemble, \(T = 298\) K, \(P = 1\) bar) was performed, to enable the water molecules to reach more favorable positions. The particle-mesh Ewald sum-mation was used for long-range electrostatics. A 10 Å cutoff was used for both Coulomb and Lennard-Jones interactions. The temperature and pressure was controlled through the Berendsen coupling algorithm, with the time constants 0.1 ps for temperature and 1.0 ps for pressure coupling. All bond lengths were constrained using the LINCS algorithm. In the actual MD simulations (30ns duration, 1.0 fs time step, NVT ensemble, \(T = 298\) K), the pressure coupling was removed, and other parameters were the same as those in the position restrained simulation. For the MD simulation on the active SrtA-LPTTG complex, the last snapshot of the 30ns MD simulation of the inactive complex was used as the starting structure, with Cys 188 and His 127 being modified to their zwitterionic pair form. The Co-RMSD of the inactive and active SrtA-LPTTG complexes versus simulation time were plotted to ensure the systems were sufficiently equilibrated (Supporting Information S2). In the root mean square fluctuation calculations, trajectories were collected after 3ns equilibration.

**Interaction energy calculations**

Comparison of the enzyme-substrate interactions in the inactive and active Michaelis complexes is of interest. The interaction energy (IE) of the enzyme-substrate complex was investigated using MM (Amber 03 Force field) and semiempirical methods (AM1, PM3, and PM6), as well as the MM/GBVI implicit solvent method.

The IE is defined as the energy difference between the enzyme-substrate complex (E-S), and individual enzyme (E) and substrate (S), Eq. (1).

\[
IE = E_{E-S} - (E_{E} + E_{S})
\]

To eliminate the kinetic energy, geometry optimizations were performed with the Amber 03 force field, on snapshots from the MD trajectories (with solvent molecules and Na\(^+\) ions). Single point energy calculations using the Amber force field, the different semiempirical methods, and the MM/GBVI solvent model were then performed on the optimized geometries of the full enzyme-substrate complexes with all water molecules and Na\(^+\) ions removed, and the average values were calculated.

**RESULTS AND DISCUSSION**

**SrtA homology model**

For homology modeling, the sequence identity is the most important criterion when choosing a template. Generally, a higher sequence identity template will give a more reliable model structure. In the initial homology search, two crystal structures of SrtA enzymes were found, namely those of S. pyogenes SrtA (3FN5) and S. aureus SrtA (1T2W). Despite the fact that the S. aureus SrtA is experimentally the most studied SrtA enzyme, S. pyogenes SrtA was chosen as the template to model L. monocytogenes SrtA, because the sequence alignment showed that the catalytic domain sequence identity between L. monocytogenes and S. pyogenes (47.6%) is much higher than that between L. monocytogenes and S. aureus (26.9%).

The obtained L. monocytogenes SrtA homology model superposes well with the template, with the overall Co RMSD being 1.20 Å. The side chains of Cys 188/208, His 127/142, and Arg 197/216, which are known to be the most important residues for the catalytic activity in SrtA enzyme, are almost identical (Supporting Information S3). The most significant difference between the homology model and the template is the B4/H2 loop, which is consistent with the fact that the sequence identity is quite low in this region (Supporting Information S1).

Comparing the molecular surfaces of the homology model and the template, we found that the active sites are very similar (Fig. 2a,b). Region 1 indicates the active
site, which is a long groove in which the pentapeptide substrate binds. Region 2 is a hydrophobic cavity, which is a desirable binding site for the isobutyl sidechain of leucine in the LPXTG sequence. Region 3 represents the position of Arg 197/216, and Region 4 shows the position of His127/142. It should be mentioned that the other important residue Cys 188/208, is not labeled on the molecular surface as it is shielded by other residues.

**Figure 2**
Molecular surfaces of (a) *L. monocytogenes* SrtA homology model; (b) *S. pyogenes* SrtA; (c) Inactive SrtA-LPTTG complex after 30ns simulation; (d) Active SrtA-LPTTG complex after 30ns simulation. Green, blue and pink colors indicate hydrophobic regions, mildly polar regions and hydrogen bonding regions respectively.
MD simulation on SrtA-LPTTG complex

To study the structure of the enzyme-substrate complex, the extended LPTTG pentapeptide was docked into the SrtA homology model. The LPTTG substrate was used because the toxin protein *L. monocytogenes* InlA, which was the first LPXTG protein shown to be a substrate of *L. monocytogenes* SrtA, possesses this motif.

Interestingly, the best docking pose shares similarities with the NMR structure of the *S. aureus* SrtA-LPAT* complex. The isobutyl sidechain of Leucine enters the hydrophobic region 2, attaining greasy contact with Ile 167, Val 172, Val 174, Ile 175, Ile 186 as well as the aliphatic part of Arg 197. The side chains of Met 111 and Ala 125 hold the proline ring. Arg 197 donates two hydrogen bonds to the backbone carbonyl oxygen atoms of the substrate leucine and proline (LP). The peptide bond between the substrate TG, which will be cleaved by the enzyme, stays close to the thiol group of Cys 188 and the imidazole ring of His 127 (Fig. 3a).

MD simulations were performed on this pose with the Cys-His catalytic dyad set to be neutral, which is known to be the inactive form. After 30ns MD simulation, the active site groove has been widened, but the positions of Arg 197 and His 127 are similar to the docking pose. The substrate LP residues maintain in the initial docked positions. However, the TTG part was found to be more flexible, as no specific active site residue is able to lock this part in position. The flexibility of the residues was determined by calculating the root mean square fluctuation (RMSF) of the MD simulation trajectory (Fig. 4).

In the simulation of the active form, the Cys-His catalytic dyad was modified to the zwitterionic form, starting from the last structure of the simulation of the inactive complex. We found that after 30ns simulation Arg 197 had switched its two hydrogen bonds from the LP to the TG residues in the substrate (Fig. 3b). As a result, the flexibility of the TTG part is reduced, as Arg 197 interacts with this part throughout most of the simulation. Comparing the molecular surfaces of the equilibrated structures of the inactive and active SrtA-LPTTG complex (Fig. 2c,d), it is easily noted that Arg 197 has moved towards His 127. To monitor motion of Arg 197, the distance between the center of mass in Arg 197 and His 127, was plotted as a function of time (Fig. 5).

The Arg 197 motion can be divided into three stages, corresponding to three different conformations (Fig. 5). In the beginning, the guanidinium cation donates two hydrogen bonds to the LP backbone carbonyl oxygen atoms.

**Figure 3**
Optimized structures of the last snapshot of 30ns simulation in the (a) inactive and (b) active SrtA-LPTTG complexes. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

**Figure 4**
Comparison of the residue RMSF of the LPTTG substrates in the inactive and active SrtA-LPTTG complexes.

---

B. Tian and L.A. Eriksson

PROTEINS

1568
atoms. This is equivalent to the conformational orientation in the inactive form. After about 2 ns, the guanidinium cation shifts its hydrogen bonding to the carbonyls of the TT part. Twenty nanoseconds later, the hydrogen bonds have switched to the TG residue carbonyls, where it remains for the final 10 ns of the simulation.

The motion of Arg 197 is reversible in that, when we take the last snapshots of the MD simulations on the active enzyme-substrate complex and reassign the Cys-His dyad to the neutral form, Arg 197 will switch its hydrogen bonds from the TG part back to the LP part of the substrate (Supporting Information S4). Both forward and reverse simulations of active and inactive forms were extended to 120 ns; however, no further changes were noted after the first 30 ns simulations, and data hence not shown.

Since there is no experimental data available to support our findings, we also performed similar MD simulations on the NMR structure of the S. aureus SrtA-LPAT* complex.25 Interestingly, the motion of Arg 197 (S. monocyto genes numbering) was reproduced by the MD simulations; these results along with QM/MM studies of the catalytic mechanism in progress and will be reported elsewhere.

The motion of Arg 197 is caused by a rotation of the sidechain guanidinium cation, rather than the motion of the backbone atoms. The most likely reason for this rotation is charge–charge interaction. When Cys 188 exists in its deprotonated form, the thiolate anion attracts the guanidinium cation, and pulls it towards the Cys-His dyad, whereby the reactive Cys-His dyad can be stabilized and the T-G peptide bond will be fixed in the right position. We expect that the thiolate anion will then be able to attack the peptide bond, with Arg 197 still in a favorable position to stabilize the tetrahedral intermediate. Thus, the motion of Arg 197 explains its dual roles in binding the substrates as hydrogen bond donor and in stabilizing the reactive intermediates. However, further calculations, at the quantum level, on activation energy differences between different conformations are required to prove our hypothesis. Such calculations are currently under way.

**Enzyme-substrate interactions**

When investigating the enzyme-substrate interactions, we first optimized the last snapshots of the 30 ns MD
simulations of each of the SrtA-LPTTG complexes. Ramachandran maps were plotted to confirm the reliability of these two structures. Less than 1% outliers are found in either structure, implying the structures are reasonable (Supporting Information S5). Superposing the equilibrated structures of the inactive and active complexes, we observe that besides the motion of Arg 197, the LPTTG substrate in the active SrtA-LPTTG complex has moved away from the active site, compared to the inactive one (Fig. 6). This implies a weaker enzyme-substrate interaction (see below). Differences were also found in the \(\beta_6/\beta_7\) and \(\beta_7/\beta_8\) loops. In the active SrtA-LPTTG complex, the \(\beta_6/\beta_7\) loop has shifted slightly towards the substrate, whereas the \(\beta_7/\beta_8\) loop has moved away from the substrate.

Four points of the inactive and active simulation trajectories of the SrtA-LPTTG complex were used to illustrate the changes in IE (Fig. 7). The first one is for the inactive SrtA-LPTTG complex after 30 ns simulation \((t = 30\text{ ns})\). The other three are from the active Srt-LPTTG simulation trajectory \((t = 0, 15, \text{ and } 30\text{ ns})\). The relative energy and IE for each point in Figure 7 are the average energies of 10 optimized snapshots around these points in the trajectories, except for the second point which is the starting structure \((0\text{ ns})\) of the active SrtA-LPTTG complex after optimization (for details, see Supporting Information S6).

Although the absolute values of interaction energies are different when using different methods, the trends are highly similar (Fig. 7). Due to the displacement in geometry, the overall enzyme-substrate interaction for the active SrtA-LPTTG complex becomes weaker, compared with the inactive one. The calculated difference in interaction energies between the inactive \((t = 30\text{ ns})\) and active \((t = 30\text{ ns})\) forms are 10–23 kcal/mol depending on method used (Fig. 7). It should be mentioned that the solvation and entropy effects were not considered in the semiempirical calculations due to computational limitations, which may cause errors in the estimate of binding affinities. The observed trends are thus more relevant in this respect, rather than the explicit values of the interaction energies.
The variance in IE imply that when the enzyme becomes chemically active, the substrate binding affinity will be affected (lowered). Therefore, when designing inhibitors, one must consider the binding affinity of the inhibitors to the active, rather than the inactive, form of the enzyme. Due to the difficulty in characterizing the active enzyme-substrate structure with experimental techniques (X-ray or NMR), theoretical methods appear essential in the design of L. monocytogenes specific SrtA inhibitors.

**CONCLUSIONS**

In summary, the structures of the L. monocytogenes SrtA-substrate complexes in both inactive and active forms were studied using homology modeling, docking and MD simulation techniques. We found that when the SrtA enzyme converts to the active form, Arg 197 will shift its hydrogen bonds from the LP residues to the TG residues of the LPXTG substrate. The motion of Arg 197 explains its catalytic roles in binding the substrates and stabilizing the reactive intermediates. Due to the changes in geometry, the overall enzyme-substrate interaction becomes weaker in the active SrtA-LPTTG complex, compared with the inactive one, but are more localized to the region about to undergo bond cleavage. We thus conclude that when designing inhibitors against SrtA, the differences in structure and substrate binding affinity between the inactive and active enzyme need be considered.

**REFERENCES**


22. Kudryavtsev KV, Bentley ML, McCafferty DG. Probing of the cis-5-pheryl proline scaffold as a platform for the synthesis of mechanism-based...


Paper 2
Catalytic Mechanism and Roles of Arg197 and Thr183 in the *Staphylococcus aureus* Sortase A Enzyme

Bo-Xue Tian† and Leif A. Eriksson*,†‡

†School of Chemistry, National University of Ireland—Galway, Galway, Ireland
‡Department of Chemistry, University of Gothenburg, 412 96 Göteborg, Sweden

Supporting Information

**ABSTRACT:** The sortase A enzyme, which catalyzes the peptidoglycan cell wall anchoring reaction of LPXTG surface proteins, has been proposed to be a universal target for therapeutic agents against Gram-positive bacteria. The catalytic mechanism of the *Staphylococcus aureus* sortase A enzyme has been systematically studied using molecular dynamics simulations, ONIOM(DFT:MM) calculations, and QM/MM charge deletion analysis. The catalytic roles of Arg197 and Thr183 were analyzed. Our calculations show that Arg197 has several important roles in the mechanism. It is crucial for substrate binding, and is capable of reversible shift of its hydrogen bonds between the LP and TG carbonyls of the LPXTG substrate motif, depending on the protonation state of the catalytic Cys184—His120 dyad. Arg197 stabilizes the catalytic dyad in the active ion pair form but at the same time raises the barrier to acylation by approximately 8 kcal/mol. Thr183 is also essential for the catalytic reaction in that it correspondingly lowers the barrier by the same amount via electrostatic interactions. The catalytic mechanism proceeds via proton transfer from His120, followed by nucleophilic attack from the thiolate anion of Cys184. The data thus supports the proposed reverse protonation mechanism, and disproves the hypothesis of the Arg197 generating an oxyanion hole to stabilize the tetrahedral intermediate of the reaction.

1. INTRODUCTION

The cell wall anchored surface proteins of Gram-positive pathogenic bacteria play vital roles in pathogenicity.1–3 The sortase A (SrtA) enzyme, which catalyzes the peptidoglycan cell wall anchoring reaction of LPXTG (leucine, proline, X, threonine, and glycine, where X is any amino acid) surface proteins, has been proposed to be a universal target for therapeutic agents against Gram-positive bacteria.4–6 As inactivation of SrtA reduces the virulence of pathogens significantly.7,8 The *Staphylococcus aureus* (S. aureus) SrtA is currently the most studied SrtA enzyme due to the severity of Methicillin-resistant *S. aureus* (MRSA) infections, and a number of studies have been carried out in order to search for inhibitors against *S. aureus* SrtA.9,10 However, only a few inhibitors with IC₅₀ values in the submicromolar range are to date identified.11 Knowledge on the enzymatic catalysis can promote the discovery of SrtA inhibitors, and studies on the kinetics,12,13 mutagenesis,14–17 and structures18–22 of the SrtA enzyme have provided important information for understanding the catalytic mechanism. However, no consensus on the catalytic mechanism of SrtA has as yet been reached.

The reaction catalyzed by SrtA is a transpeptidation that undergoes a ping-pong mechanism.12,13 In the first phase, SrtA cleaves the peptide bond between the threonine and glycine (T—G) residues in the LPXTG motif, forming an LPXT—enzyme intermediate. In the second phase, the second substrate pentaglycine (from the outer cell wall cross bridge) is attached to the LPXT moiety, forming the cell wall anchored LPXT surface protein (Scheme 1). According to mutagenesis studies,14–17 Cys184, His120, Arg197, and Thr183 are indispensable, as alanine substitution of any of these four residues results in a more than 1000-fold decrease in enzymatic activity. Several catalytic mechanisms were proposed, including a Cys184—His120 ion pair Cys184 thiol activation mechanism17 which is similar to the enzyme papain (a paradigm cysteine protease23,24), a general base mechanism in which the Cys184 thiol is activated by His120 or Arg19721,25 after the substrate is bound, and a reverse protonation mechanism where a small fraction of the active SrtA is competent for catalysis12 (see ref 12 for a detailed description of the various mechanisms). According to pKₐ measurements, pH dependence experiments, and solvent isotope effect measurements, the Cys184—His120 ion pair mechanism and the general base mechanism have been shown unlikely to occur.12 The reverse protonation mechanism suggests that Cys184 and His120 exist in the thiolate anion and imidazolium...
The role of Thr183 in the catalysis, which has hitherto not been explained, is also discussed.

2. COMPUTATIONAL DETAILS

2.1. Preparation of the Michaelis Complex. The first snapshot of the NMR structure of the *S. aureus* LPXTG−SrtA intermediate (residues 62−206, PDB code: 2KID, RCSB Protein Data Bank) was used as the initial model in the current study. We chose this structure instead of the X-ray structure of the *S. aureus* LPXTG−SrtA complex for several reasons. First, in the NMR structure, the Cys184 and His120 residues are close to each other, whereas in the X-ray structure His120 is far away from Cys184. Mutation studies demonstrated that alanine substitution of His120 results in a 10000-fold decrease in enzyme activity, suggesting that His120 is even more important than Cys184 (1000-fold decrease when mutated to alanine). The X-ray structure is unable to explain the role of His120. Second, in the X-ray structure, Cys184 was mutated to alanine, which may result in undesirable conformation changes. It has been shown that the X-ray structure of the *S. aureus* SrtA−LPXTG complex is significantly different from the structures of other SrtA homologues. Moreover, due to the sparse population of the active form of the enzyme, most structures of SrtA resolved are most likely in the inactive form. The NMR structure is suitable for our purpose on studying the catalytic mechanism of SrtA as it mimics the LPXT−SrtA intermediate, the product of the acylation step. Superposing the LPXTG substrate to the LPXT moiety of the LPXT−SrtA intermediate should be more accurate than docking the substrate to the active site of structures without substrates bound. In order to recover the Michaelis complex, an Nme-LPATG-Ace peptide was built, optimized and superposed on the LPATG moiety of this NMR structure, using the Flexible Alignment module in MOE.

2.2. Molecular Dynamics Simulations. Molecular dynamics (MD) simulations were performed to investigate if the Arg197 motion discovered in our previous study on the *L. monocytogenes* SrtA homology model also applies to the *S. aureus* case. The Amber 03 force field was used for all the MD simulations. The Cys184−His120 dyad was initially set to the active (ion pair) form, while the protonation of the remaining residues was kept the same as in the NMR structure. The obtained structure was then used to explore the proposed multiple roles of Arg197 in SrtA catalysis, including the influence on the relative stabilities of the inactive and active forms of the Cys184−His120 dyad, substrate binding, as well as the effect on the acylation step. The role of Thr183 in the catalysis, which has hitherto not been explained, is also discussed.
The Journal of Physical Chemistry B

solvated by TIP3P water molecules in a periodic box with a buffer distance of 8.0 Å. A Cl⁻ ion was randomly added to satisfy the electroneutrality condition, using the genion module in GROMACS (version 4.0.4) software. The whole system (~20000 atoms) was energy minimized by steepest descent (200 steps) to remove close contacts. A position restrained simulation (2 ns duration, 1.0 fs time step, NPT ensemble, \( T = 298 \text{ K}, P = 1 \text{ bar} \)) was performed, to enable the water molecules to reach more favorable positions. The leapfrog algorithm was used for the integration of the classical Newton equations of motion. Particle-mesh Ewald (PME) summation was used for long-range electrostatics. A 10 Å cutoff was used for both Coulomb and Lennard-Jones interactions. The temperature and pressure were controlled through the Berendsen coupling algorithm, with the time constants 0.1 ps for temperature and 1.0 ps for pressure coupling. All bond lengths were constrained using the LINCS algorithm. In the production run (20 ns duration, 1.0 fs time step, NPT ensemble, \( T = 298 \text{ K}, P = 1 \text{ bar} \)), the temperature was controlled using the Nose–Hoover thermostat with a time constant of 0.1 ps and the pressure was controlled using the Parrinello–Rahman barostat with a time constant 1.0 ps. Other parameters were similar to the position restrained simulations.

Prior to the production run, we performed 1 ns MD simulations using the production run parameters and recorded snapshots at 200 ps, 400 ps, 600 ps, 800 ps, and 1 ns as initial structures for further production runs. After the production run of the active form SrtA–LPXTG complex, we modified the protonation states of the Cys–His dyad to the inactive form and performed another 20 ns MD simulation in order to investigate if the Arg197 side chain can switch to the LP moiety (corresponding to the reverse simulation of our previous L. monocytogenes study). A 20 ns reversibility tests were also performed to confirm the Arg197 motion (for details, see Supporting Information S6-3). All MD simulations were performed using GROMACS.

2.3. QM/MM (ONIOM) Calculations. Representative MD snapshots were further energy minimized with the Amber 03 force field. In the QM/MM calculations, MM parameters for the active Cys–His catalytic dyad were used. Four main systems were investigated, including the last snapshot of the active form from the 20 ns MD trajectory (called Position1), the last snapshot of the inactive form from the 20 ns MD trajectory (called Position2), the R197A mutation of Position1 (called Position1-R197A), and the T183A mutation of Position1 (called Position1-T183A). Mutations were readily obtained using the MOE program. It should be noted that Position1 and Position2 refer to the zwitterionic and neutral forms of the Cy-s184–His120 catalytic dyad. For the R197A and T183A systems, position restrained simulations were performed prior to the MM optimizations, to equilibrate the surrounding water molecules. The substrate-free systems for Position1, Position2, and Position1-R197A were also studied for comparison of the relative stabilities of the inactive and active forms of the enzyme. Position restrained simulations were performed after the removal of the substrates for the three substrate-free systems.

A two-layer ONIOM (QM/MM) scheme was used for the calculations of the SrtA catalytic mechanism. The standard DFT method B3LYP and the two basis sets 6-31G(d) (BS1) and 6-311+G(2d,p) (BS2) were used for treating the QM region. It should be noted that the B3LYP/Amber method has previously been used to study a similar enzyme, papain, and the results are in good accordance with experimental data. Geometries of intermediates and TSs were optimized using the ONIOM(B3LYP/BS1:Amber) method with the ONIOM-ME (mechanical embedding) scheme, followed by frequency calculations at the same level of theory. Atoms far away from the substrate carbonyl group (>15 Å) were kept fixed during the optimizations, and the normal modes of these atoms were excluded from the frequency calculations with the Gaussian 09 keyword notatoms. Single-point energy calculations with the ONIOM(B3LYP/BS2:Amber) method were performed on the optimized geometries with the ONIOM-EE (electronic embedding) scheme. Geometries were not optimized with the ONIOM/EE scheme due to the high computational cost and poor convergence. All ONIOM calculations were performed using the Gaussian 09 software.

The QM/MM partitioning is described in Scheme 2. Three different partitioning schemes were used for studying the acylation step at Position1. Model-1 (blue) contains the ATG moiety of the substrate as well as the side chains of Cys184 and His120. Model-2 (blue + red) added Arg197 and two solvent molecules (SOL-1 and SOL-2) to the QM region, and Model-3 further added Thr183, the Thr183-Cys184 backbone, and another seven solvent molecules to the QM part. Model-1 was also used for the calculations on the Position2, Position1-R197A, and Position1-T183A systems.

2.4. Charge Deletion Analysis and Binding Affinity Calculations. Charge deletion analysis was used to investigate the influence of individual residues as well as solvent molecules on the enzymatic catalysis. This approach has been used in a number of theoretical studies. Deletion of atomic charges of a certain residue in the MM region can result in energy changes in the QM region (\( \Delta E_{QM} \)). If the QM energy changes of a reactant \( \Delta E_{QM}(R) \) and a transition state \( \Delta E_{QM}(TS) \) are different, the reaction barrier is affected. The electrostatic effect of a certain residue on the reaction barrier \( \Delta \Delta E_{QM}(R \rightarrow TS) \) can be defined as in eq 1. A positive sign of \( \Delta \Delta E_{QM}(R \rightarrow TS) \) indicates that the effect of this residue is to lower the barrier height, whereas a negative sign implies that it plays the opposite role. Similarly, the relative stabilities of a reactant and a product (for instance, the relative stabilities of the inactive and active forms of the enzyme) can also be evaluated; see eq 2. All the \( \Delta E_{QM} \) values reported in this study are ONIOM-EE(B3LYP/BS2: Amber) single point energies. To reduce computational cost, every three adjacent residues were combined into a group, and charge deletion analysis was initially used for identifying important
groups (ΔΔE_{QM} > 1 kcal/mol). On the basis of this initial screening, the individual residues in the important groups, as well as all active site residues (not frozen in the ONIOM calculations), were subsequently analyzed (Supporting Information S1).

\[
\Delta \Delta E_{QM}(R\rightarrow TS) = \Delta E_{QM}(TS) - \Delta E_{QM}(R) \tag{1}
\]

\[
\Delta \Delta E_{QM}(R\rightarrow P) = \Delta E_{QM}(P) - \Delta E_{QM}(R) \tag{2}
\]

The binding affinity of the substrate was evaluated using the MM/GBVI approach in the MOE software, which includes implicit solvent effects. The average value of binding affinities from 10 snapshots (with all solvent molecules removed) in the last 1 ns MD trajectory was evaluated (for the Position1-R197A system, only the last snapshot of the position restrained simulation was used).

3. RESULTS AND DISCUSSION

3.1. Arg197 Motion. In our previous study on the L. monocytogenes SrtA homology model, Arg 197 was found capable to change its hydrogen donor interactions within 30 ns MD simulations, from the LP backbone carbonyl groups of the LPXTG substrate in the inactive form to the TG backbone carbonyls in the active form, and that this motion was fully reversible. To explore the Arg197 motion in S. aureus SrtA, we performed a 20 ns MD simulation on the active form, followed by another 20 ns MD simulation with the Cys–His dyad switched to the inactive form.

As expected, Arg197 forms hydrogen bonds to the carbonyl oxygen of the TG peptide bond in the active form, and swings over to form hydrogen bonds to the LP part in the hydrophobic pocket of the active site within the 20 ns MD simulation on the inactive form (five individual 20 ns simulations were performed and showed consistent results). We have shown previously that 20–30 ns simulation is sufficient for this small system, and that extending the simulation time to 100 ns does not substantially influence the results for the Arg197 motion. The hydrogen bonds between Arg197 and the substrate in Position1 and Position2 are illustrated in Figure 1, and a superposed view is shown in Figure 2. In Position1, the side chain of Arg197 forms one direct hydrogen bond with the carbonyl group of the TG peptide bond (H11-OThr, Figure 1). In the simulation of the inactive form of the SrtA–LPXTG complex, the H11-OThr

Figure 1. Hydrogen bonds between Arg197 and substrate in the 20 ns MD simulation of the inactive SrtA–LPXTG complex, starting from the simulated active complex. Position1 and Position2 refer to the preferred positions of Arg197 in the active and inactive SrtA–substrate complexes.

Figure 2. Superposed structures with Arg197 in Position1 (in red) and Position2 (in blue).
hydrogen bond is broken spontaneously, while two new hydrogen bonds are formed (H21 ⋅ ⋅ ⋅ O_leu and HE ⋅ ⋅ ⋅ O_leu, Figure 1). The repositioning of Arg197 depending on the protonation state of the His−Cys dyad may be essential for the SrtA catalysis, for one or more reasons: (1) the Arg197 motion may alter the relative stabilities of the inactive and active forms of the Cys184−His120 catalytic dyad; (2) this motion may increase the substrate binding affinity; (3) Arg197 may form an oxanion hole to stabilize the transition state or tetrahedral intermediate in the acylation step. To elucidate the role of Arg197, different systems were therefore investigated by means of QM/MM (ONIOM) calculations, as outlined above.

3.2. Relative Stabilities of the Inactive and Active Forms.

The relative stability of the inactive and active forms is an important issue in the SrtA enzyme, as pK\textsubscript{aq} measurements\textsuperscript{12,25} have shown that the majority of the enzyme exists in the inactive form. In this respect, SrtA seems to differ from the cysteine proteases such as papain. However, the bell-shaped pH dependencies of kinetic parameters k\textsubscript{cat}/K\textsubscript{m} for SrtA imply that one residue in the Cys−His dyad must be protonated and the other deprotonated for activity.\textsuperscript{12} The reverse protonation mechanism was proposed,\textsuperscript{12} i.e., that a small fraction of SrtA with the Cys−His dyad in the active form catalyzes the reaction.

Recently, Mladenovic et al.\textsuperscript{30} demonstrated that the enzyme environment is capable of altering the relative stabilities of the neutral versus zwitterionic forms of the Cys−His dyad in cathepsin (one of the cysteine proteases)\textsuperscript{22} using QM/MM molecular dynamics simulations, which explained the abnormal pK\textsubscript{a} values of the Cys−His dyad measured in cysteine proteases.\textsuperscript{56−58} The question we posed was whether the Arg197 motion is capable of altering the relative stabilities of the inactive form versus the active form in SrtA. To address this issue, we performed systematic QM/MM calculations on the transformations between the inactive and active forms in Position1, Position1-R197A, and Position2.

Strictly speaking, the substrate may bind to the inactive or active SrtA enzymes, corresponding to the general base mechanism and the reverse protonation mechanism, respectively. We therefore studied the relative stabilities of the inactive and active forms with and without the substrate bound. The protonation states in Position1 (with and without substrate bound) were initially explored using five different simulation snapshots (the last snapshot from five individual MD simulations). In Position1, the active form was found to be more stable than the inactive form (Table 1, entry 1 and 4) by 24−30 kcal/mol. Given the small difference between the five simulation snapshots (Table 1, entry 1), only one starting structure was used in the Position2 and Position1-R197A systems. From Table 1, it is clear that the inactive form becomes more stable than the active form in either Position1-R197A (Position1 with Arg197 mutated to alanine) or Position2 (Arg197 located far away from the Cys−His dyad). In addition, the active form becomes less stable with the substrate bound (Table 1).

The rationale for this is that the substrate consists of hydrophobic residues which prevents solvent molecules from entering the enzyme pocket and that the Cys−His ion pair is less favored in the resulting nonpolar environment. It should be mentioned that the His120 proton forms a stable hydrogen bond (or salt bridge) with the Cys184 thiolate group in all the MD simulations, and we did not observe any water molecules (which could be proton shuttles as described in ref 34) in the middle of the Cys−His ion pair.

Charge deletion analysis of Position1 (with substrate bound) showed that Arg197, Lys134, Lys198, and SOL−1 are essential for the stabilization of the active Cys−His dyad, whereas Thr183, Thr121, Asp186, and SOL−2 play an opposite role (Table 2).

### Table 1. Proton States of the Cys−His Dyad under Different Conditions

<table>
<thead>
<tr>
<th>entry</th>
<th>system</th>
<th>active</th>
<th>inactive</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Position1 without LPXTG</td>
<td>0.0</td>
<td>27.4 ± 3.0\textsuperscript{b}</td>
</tr>
<tr>
<td>2</td>
<td>Position1-R197A without LPXTG</td>
<td>0.0</td>
<td>−10.8</td>
</tr>
<tr>
<td>3</td>
<td>Position2 without LPXTG</td>
<td>0.0</td>
<td>−9.2</td>
</tr>
<tr>
<td>4</td>
<td>Position1 with LPXTG</td>
<td>0.0</td>
<td>2.5</td>
</tr>
<tr>
<td>5</td>
<td>Position1-R197A with LPXTG</td>
<td>0.0</td>
<td>−15.2</td>
</tr>
<tr>
<td>6</td>
<td>Position2 with LPXTG</td>
<td>0.0</td>
<td>−14.1</td>
</tr>
</tbody>
</table>

\textsuperscript{a}The energy of the inactive form relative to the active form is used for each case, and the energies for different systems are not comparable, as the number of atoms used is different. All energies are ONIOM-EE(B3LYP/BS2-Amber) single point energies on Model-3 partitioning (Scheme 2), in kcal/mol. Relative energies of Model-1 are summarized in Supporting Information S2. \textsuperscript{b}Average energy of five snapshots, with error bar. \textsuperscript{c}The N−H distance in His120 in the active form is constrained at 1.05 Å as optimization of the active form spontaneously leads to the inactive form.

### Table 2. Charge Deletion Analysis for Position1 (with LPXTG): Influence of the Environment on the Relative Stabilities of the Inactive and Active forms

<table>
<thead>
<tr>
<th>influence of residues</th>
<th>(\Delta\Delta\text{E}_{\text{QM}}) (active − inactive)\textsuperscript{b}</th>
<th>influence of nearby solvent molecules</th>
<th>(\Delta\Delta\text{E}_{\text{QM}}) (active − inactive)\textsuperscript{b}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg197</td>
<td>23.0</td>
<td>SOL-1</td>
<td>15.8</td>
</tr>
<tr>
<td>Thr183</td>
<td>−10.0</td>
<td>SOL-2</td>
<td>−5.1</td>
</tr>
<tr>
<td>Thr121</td>
<td>−3.1</td>
<td>SOL-3</td>
<td>−0.3</td>
</tr>
<tr>
<td>Lys134</td>
<td>5.9</td>
<td>SOL-4</td>
<td>0.0</td>
</tr>
<tr>
<td>Asp185</td>
<td>1.9</td>
<td>SOL-5</td>
<td>−0.2</td>
</tr>
<tr>
<td>Asp186</td>
<td>−6.0</td>
<td>SOL-6</td>
<td>0.5</td>
</tr>
<tr>
<td>Trp194</td>
<td>−0.2</td>
<td>SOL-7</td>
<td>−0.9</td>
</tr>
<tr>
<td>Lys198</td>
<td>5.5</td>
<td>SOL-8</td>
<td>0.8</td>
</tr>
<tr>
<td>Thr in substrate</td>
<td>−3.7\textsuperscript{a}</td>
<td>SOL-9</td>
<td>1.9</td>
</tr>
</tbody>
</table>

\textsuperscript{a}Positions of the solvent molecules as described in Scheme 2. \textsuperscript{b}As defined in eq 2, in kcal/mol. \textsuperscript{c}Charge deletion analysis on Model-1 with the substrate in the MM region.

### Table 3. Substrate Binding Affinity under Different Conditions

<table>
<thead>
<tr>
<th>entry</th>
<th>system</th>
<th>MM/GBV binding affinity</th>
<th>(\Delta\Delta\text{G}) (active − inactive)\textsuperscript{a}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Position1 active</td>
<td>−55.8 ± 0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>Position1 inactive</td>
<td>−54.0 ± 0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>Position1-R197A active</td>
<td>−49.6 ± 0.9</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>Position1-R197A inactive</td>
<td>−50.9 ± 1.2</td>
<td>0.7</td>
</tr>
<tr>
<td>5</td>
<td>Position2 active</td>
<td>−55.0 ± 0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>6</td>
<td>Position2 inactive</td>
<td>−54.0 ± 0.8</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\textsuperscript{a}In kcal/mol.
The Journal of Physical Chemistry B

Table 4. Properties of the TS in Position1 for Different ONIOM Schemes

<table>
<thead>
<tr>
<th>ONIOM model</th>
<th>barrier heighta</th>
<th>imaginaryb frequency</th>
<th>S_{Cys184···CThr} distance in TS</th>
<th>H_{EHis120···NGly} distance in TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model-1</td>
<td>19.4</td>
<td>-730.4</td>
<td>2.880</td>
<td>1.412</td>
</tr>
<tr>
<td>Model-2</td>
<td>14.8</td>
<td>-700.2</td>
<td>2.907</td>
<td>1.573</td>
</tr>
<tr>
<td>Model-3</td>
<td>17.5</td>
<td>-536.8</td>
<td>3.350</td>
<td>1.190</td>
</tr>
</tbody>
</table>

a In kcal/mol. b In cm⁻¹. c In Å.

Nonpolar residues such as Trp194 have a minor influence on the relative stabilities of the inactive and active forms.

Besides the influence on the relative stability of the Cys–His dyad, the Arg197 motion may also affect the binding affinity of the substrate (Table 3). As expected, the hydrogen bonding between Arg197 and LPXTG is important for substrate binding, as the R197A mutation results in a 4–6 kcal/mol decrease in binding affinity (Table 3, entry 1 vs entry 3 and entry 2 vs entry 4).

However, the binding affinity is not significantly affected by the Arg197 motion (entry 1 vs entry 5 and entry 2 vs entry 6), implying that the substrate does not selectively bind to the active form of the enzyme. The ~5 kcal/mol reduced binding affinity of the substrate also correlates well with the 1000-fold reduction in activity noted for the R197A mutant.17

We note that the general base mechanism remains unlikely to occur, since with the LPXTG bound to the inactive enzyme, the proton transfer from Cys184 to His120 becomes more endothermic (Table 1, entries 3 and 6), unless the Arg197 side chain moves to the TG region. According to our MD simulations, however, the Arg197 side chain stays in the LP region if the Cys–His dyad is in the inactive form. We also performed MD simulations of the active form SrtA without the substrate bound. Unlike the SrtA–substrate complex system, no preferred position of the Arg197 side chain could be identified in the substrate-free system, probably because it becomes more flexible without the substrate to bind to (simulations started from Position1 and Position2, Supporting Information S3). With the substrate bound, Arg197 orientates the side chain hydrogen bonds to the TG moiety of the substrate (Position1), which is the preferred position thereof. Taken together, our calculations favor the reverse protonation mechanism,12 i.e., the LPXTG substrate directly binds to the active form of the enzyme and undergoes acylation, even though the majority of the enzyme is in the inactive form. This is also supported by the fact that the active form is more stable (Arg197 in Position1)/less unstable (Arg197 in Position2), and thus more populated, when the substrate is not bound (Table 1).

3.3. Acylation. Another possible role of Arg197 is to form an oxyanion hole (Position1) that will stabilize the tetrahedral intermediate in the acylation step. However, since we have demonstrated that the Cys–His ion pair, which is the reactant of the acylation step, is significantly stabilized by Arg197, the barrier height of acylation may in fact increase due to the stabilization of the reactant. Therefore, evaluation of the change in barrier height for the acylation step caused by the motion of Arg197 is of interest.

The acylation of cysteine proteases has previously been modeled in a number of theoretical studies.31–33 On the basis of AM1 calculations, Arad et al.33 suggested that in the enzyme papain a proton transfer from histidine (PT-His) occurs prior to, or concerted with, the nucleophilic attack by the thiolate anion (NA-Cys). The papain system was further studied by Harrison et al.32 with the AM1/MM and B3LYP/MM methods, and the concerted pathway without the involvement of a tetrahedral intermediate (TI) was suggested. Recently, Ma et al.33 determined the two-dimensional free energy surfaces of the acylation of human cathepsin by using AM1/MM molecular dynamics simulations. The PT-His and NA-Cys reactions were found highly coupled, and a TI was formed along the NA-Cys pathway. Although these impressive studies provided insights into the catalytic mechanism of cysteine proteases, some limitations are obvious. First, the use of semiempirical methods may not be sufficient for the QM part. Second, no imaginary frequency of the TS was reported. Moreover, alternative QM/MM partitioning schemes were not investigated.

The standard QM/MM method ONIOM(B3LYP:Amber), which has been extensively used for studying enzyme reactions,29 was chosen, and three different QM/MM partitioning schemes were used to model the acylation step in the active form of SrtA in Position1. The deacylation step was not modeled. The second substrate pentaglycine is similar to glycine (the product released in the acylation step) in terms of the structure. It can therefore be expected that the deacylation is highly similar to the reverse process of the current acylation.

Initial scans of two reaction coordinates, i.e., NA-Cys ($S_{Cys184···CThr}$ distance) and PT-His (HE_{EHis120···NGly} distance), were performed separately. In the NA-Cys process, the ONIOM-EE (B3LYP/BS2:Amber) energy increases continuously, whereas, in the PT-His process, the energy starts decreasing when the HE_{EHis120···NGly} distance becomes less than 1.60 Å. Interestingly, the $S_{Cys184···CThr}$ distance is significantly shortened (although not forming a $S_{Cys184···CThr}$ bond) when
the HEHis120 · · · NGly distance becomes less than 1.60 Å (Supporting Information S4). These results suggest that the PT-His reaction coordinate is the major route to reach the TS, and NA-Cys may occur concertedly with or immediately after PT-His. Our results are consistent with the experimental fact that H120A mutation causes a larger reduction in enzyme activity than C184A (~10000-fold decrease for H120A and ~1000-fold decrease for C184A, respectively),17 as well as with previous theoretical studies on cysteine proteases.31–33

The TS for Position1 was then optimized; the results are summarized in Table 4. The obtained barrier heights are in reasonable accordance with the experimental value ~20 kcal/mol12,13 (calculated using the Eyring equation). Checking the normal modes of the imaginary frequency, we found that the TS corresponds to the PT-His reaction coordinate. Interestingly, an IRC-like optimization of the TS (the HEHis120 · · · NGly distance of TS was shortened, followed optimization of the product) gave the final product rather than the protonated substrate or TI, probably because the barrier heights for NA-Cys and the cleavage of the peptide bond are very small (or barrierless) once the proton transfer is completed. Further conformational analysis step), which is essential for SrtA catalysis. Arg197 is capable of reversibly switching its hydrogen donor interactions from the LP backbone carbonyl groups of the LPXTG substrate to the TG backbone carbonyls in the active form, within 20 ns MD simulations. QM/MM calculations and charge deletion analysis show that Arg197 stabilizes the thiolate anion in the Cys184—His120 catalytic dyad (the precursor of the acylation step), which is essential for SrtA catalysis. Arg197 is important for substrate binding, but the Arg197 motion does not affect the binding affinity. On the basis of the MD simulations and QM/MM calculations of the relative stabilities of the inactive and active Cys—His dyad, we suggest that the general base mechanism is unlikely to occur. The reverse protonation mechanism proposed by Frankel et al.12 is the most likely mechanism for SrtA. By analyzing the normal modes of the imaginary frequency for the TS of the acylation step, we conclude that this corresponds to proton transfer from His120 (PT-His), rather than nucleophilic attack by the thiolate anion of Cys184 (NA-Cys). IRC-like optimization of TS yields the final product, due to the barrier heights for NA-Cys and the cleavage of the peptide bond being very small (or lacking) when the PT-His step is completed. Interestingly, with the presence of Arg197, the barrier height of acylation increases, which disproves the previous hypothesis that Arg197 may stabilize the TS by forming an oxyanion hole. To compensate for this, the presence of Thr183 affects the catalysis by lowering the acylation barrier by roughly the same amount via electrostatic interactions.

The reaction energy for the system in Position1 is with the larger Model-3 endothermic by 11.9 kcal/mol (5.9 kcal/mol for the smaller Model-1); Table S5 in the Supporting Information. The barrier for the second half of the reaction, assuming the reaction profile to be symmetric, is thus only 5.6 kcal/mol (13.5 kcal/mol for the smaller model), which enables a rapid formation of the final product once the pentaglycine substrate from the peptidoglycan cell wall is in position to covalently anchor the LPXTG sequence. These energies are consistent with the fact that the acylation step is the rate-limiting step in the overall transpeptidation process.

<table>
<thead>
<tr>
<th>position</th>
<th>acylation barrier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position1</td>
<td>19.4</td>
</tr>
<tr>
<td>Position2</td>
<td>16.8</td>
</tr>
<tr>
<td>Position1-R197A</td>
<td>11.7</td>
</tr>
<tr>
<td>Position1-T183A</td>
<td>28.2</td>
</tr>
</tbody>
</table>

Table 6. Comparison of Acylation Barriers at Different States

4. Conclusions

The catalytic mechanism of the Staphylococcus aureus sortase A enzyme has been systematically studied. Arg197 is capable of reversibly switching its hydrogen donor interactions from the LP backbone carbonyl groups of the LPXTG substrate in the inactive form to the TG backbone carbonyls in the active form, within 20 ns MD simulations. QM/MM calculations and charge deletion analysis show that Arg197 stabilizes the thiolate anion in the Cys184—His120 catalytic dyad (the precursor of the acylation step), which is essential for SrtA catalysis. Arg197 is important for substrate binding, but the Arg197 motion does not affect the binding affinity. On the basis of the MD simulations and QM/MM calculations of the relative stabilities of the inactive and active Cys—His dyad, we suggest that the general base mechanism is unlikely to occur. The reverse protonation mechanism proposed by Frankel et al.12 is the most likely mechanism for SrtA. By analyzing the normal modes of the imaginary frequency for the TS of the acylation step, we conclude that this corresponds to proton transfer from His120 (PT-His), rather than nucleophilic attack by the thiolate anion of Cys184 (NA-Cys). IRC-like optimization of TS yields the final product, due to the barrier heights for NA-Cys and the cleavage of the peptide bond being very small (or lacking) when the PT-His step is completed. Interestingly, with the presence of Arg197, the barrier height of acylation increases, which disproves the previous hypothesis that Arg197 may stabilize the TS by forming an oxyanion hole. To compensate for this, Thr183 significantly reduces the acylation barrier via electrostatic interactions, which explains the unknown role of Thr183 in SrtA catalysis. All results are consistent with available experimental data, and provide a
detailed explanation for the observed features of the sortase A enzymatic mechanism.

**References**

Petersson, G. et al. *Gaussian 09*, revision B.01; Gaussian, Inc.: Wallingford, CT, 2009 (for the complete citation, see Supporting Information S8).


Paper 3
Hydroxylation and Ring-Opening Mechanism of an Unusual Flavoprotein Monooxygenase, 2-Methyl-3-hydroxypyridine-5-carboxylic Acid Oxygenase: A Theoretical Study

Boxue Tian,[a, b] Yaoquan Tu,[a] Åke Strid,[a] and Leif A. Eriksson*[a, b]

Abstract: Hybrid meta-GGA density functional theory (the MPWB1K functional) was used to study the hydroxylation and ring-opening mechanism of 2-methyl-3-hydroxypyridine-5-carboxylic acid oxygenase (MHPCO). This enzyme catalyses the conversion of 2-methyl-3-hydroxypyridine-5-carboxylic acid (MHPC) to \(\alpha\)-(N-acetylamino-methylene)succinic acid (AAMS), which is the essential ring-opening step in the bacterial degradation of vitamin B\(_6\). MHPCO belongs to the flavin-containing aromatic hydroxylases family. However, MHPCO is capable of catalysing a subsequent aromatic ring-cleavage reaction to give acyclic products rather than hydroxylated aromatic ones. Our calculations show that the re-aromatisation of the hydroxylated intermediate occurs spontaneously in aqueous solution; this implies that the ring-opening process occurs inside the enzyme’s active site, in which limited water is available. The instability of the hydroxylated intermediate of MHPCO is the main reason why acyclic products are formed. Previously proposed mechanisms for the ring-opening step were studied, and were shown to be less likely to occur (\(\Delta\Delta G_{298}^{\text{ring}} > 35 \text{ kcalmol}^{-1}\)). Two new pathways with reasonable barrier heights (\(\Delta\Delta G_{298}^{\text{ring}} < 15 \text{ kcalmol}^{-1}\)) are reported herein, which are in accordance with all experimental information present to date.

Keywords: hydroxylation · flavoprotein monooxygenase · molecular modeling · ring-opening reactions · vitamins

Introduction

Flavoprotein monooxygenases (FMOs) are a group of enzymes capable of efficiently and regioselectively catalysing a large number of oxidative reactions.[1–5] Within the FMO family, the aromatic hydroxylases have received much attention over the last three decades, and the protein structures of some members have already been determined.[6–9] The aromatic hydroxylase enzymes are capable of hydroxylating aromatic rings at the ortho position of phenolic substrates via electrophilic aromatic substitution reactions.[10–12]

2-Methyl-3-hydroxypyridine-5-carboxylic acid oxygenase (MHPCO) is a unique member of this aromatic hydroxylase family because the reactions catalysed by the enzyme give acyclic products rather than aromatic ones.[13–18] So far, three compounds have been shown to be substrates of MHPCO, namely, 2-methyl-3-hydroxypyridine-5-carboxylic acid (MHPC), 5-hydroxynicotinic acid (SHN) and N-methyl-5-hydroxynicotinic acid (NMHN); see Equation (1).[19,20]

Investigating the catalytic mechanism of MHPCO is of great importance for the understanding of the aromatic hydroxylase family as well as further applications of such reactions in general.

Previous studies have shown that the monoxygenation process catalysed by aromatic hydroxylases can be divided into two parts: reduction and oxidation. In the reduction part, with NADPH and the substrate already bound, the FAD cofactor is reduced by NADPH. In the oxidation part,
the reduced FADH\(^{-}\) reacts with molecular oxygen to form C-(4a)-flavinhydroperoxide (FIHOOH), which is a short-lived intermediate (Supporting Information, S2).\[^{25}\]

FIHOOH then acts as an electrophile in the oxygenation, and the terminal hydroxyl group is transferred from FIHOOH to the substrate through electrophilic aromatic substitution.\[^{10–12}\]

Previous experimental investigations have provided useful information on the catalytic mechanism of MHPCO. In the experimental studies of Chaiyen et al. several mechanisms were proposed.\[^{19,20}\] Based on the crystal structure, McCulloch et al. recently proposed a new mechanism.\[^6\] Although different pathways are suggested, they share the similarity that hydroxylation (Scheme 1) and ring opening are the two key catalytic steps.\[^{18}\] As with other aromatic hydroxylases, the substrate is first hydroxylated by C-(4a)-flavinhydroperoxide (FIHOOH) through an electrophilic aromatic substitution, which is the rate-limiting step. In the ring-opening step, a water molecule is added to the hydroxylated intermediate either before or after the C2/C0 bond cleaves [Eq. (1)].

So far, no theoretical study on the full catalytic process of MHPCO has been reported. However, there are some theoretical studies modelling the hydroxylation step of p-hydroxybenzoate hydroxylase (PHBH), a paradigm aromatic hydroxylase, which provides insight into the hydroxylation step.\[^{26–33}\] In these studies, the energy barrier of the hydroxylation step was determined to be 20–30 kcal mol\(^{-1}\) by using different theoretical methods. Because the experimental value is 12 kcal mol\(^{-1}\), it was suggested that additional interactions with the enzyme must be considered for the hydroxylation step.\[^{26}\]

Based on previous experimental data, we have herein examined the catalytic mechanism of MHPCO by using quantum chemical methodology in the hybrid meta-GGA density functional theory framework. Previously proposed mechanisms were studied, and two new pathways are presented based on theoretical considerations.

### Computational Methods

The hybrid meta density functional theory (HMDFT) method MPWB1K\[^{34}\] was used to study the reaction pathways because MPWB1K is known to provide accurate barrier heights as well as hydrogen-bonding interactions.

Intermediates and transition states were first optimised in vacuo at the B3LYP/6-31G(d) level.\[^{35–37}\] Further geometry optimisations by using MPWB1K/6-31+G(d,p) were performed on the geometries of the previous step, followed by frequency calculations at the same level of theory to ensure that these were stationary structures on their respective energy surfaces and to extract Gibbs free energy corrections at 298 K. Transition states were tested by using intrinsic reaction coordinate (IRC) calculations to make sure that they connected the correct reactants and products in each step. The integral equation formalism of the polarised continuum model (IEFPCM)\[^{38,39}\] with UAHF radii was used to optimise geometries and calculate frequencies for all conversions that occur in aqueous solution (\(\varepsilon = 78.3\)).

Single-point energy calculations were performed on the optimised structures by using
the higher basis set 6-311+G(2d,2p) in the proton affinity calculation of the hydroxylated intermediates (see below). When evaluating the rotation barrier of different configurations, both transition states in clockwise and anticlockwise rotation were located, and the one with the lower energy barrier was chosen to be the transition state (TS).

The substrate 5HN was used instead of MHPC for all the calculations because 5HN is also a substrate of MHPCO and has fewer atoms; the difference between the two structures is a methyl group in the R1 position [Eq. (1)]. This is, however, expected to influence the energetics of the reactions to a very small extent. All calculations were performed by using the Gaussian 03 program.[40]

**Results and Discussion**

**Hydroxylation**

**Reactant complex:** According to previous titration experiments, the substrates of MHPCO exist in their tripolar forms when binding to the enzyme [Eq. (2)].[25] It should be mentioned that in the related enzyme (PHBH) although the substrate p-hydroxybenzoate should not have the phenolic hydroxyl deprotonated at neutral pH, it has been shown to attain a phenolate form in the enzyme.[45] It was argued that, with the phenolic hydroxyl deprotonated, the substrate is in a higher energy form, thus the energy barrier for hydroxylation is lowered.[26] These enzymes thus activate the substrate by controlling its protonation state.

Comparing the energies and structures of the anion and tripolar forms of 5HN both in vacuo and in water (PCM), the Gibbs free energy of the anion form is found to be 11.9 kcal mol\(^{-1}\) lower than the tripolar form in vacuo, and 7.9 kcal mol\(^{-1}\) lower in water (see Figure 1). Therefore, 5HN exists mainly in the anion form in water. Compared with the anion form, the C2–C3 and C3–C4 bond lengths are increased in the tripolar form (1.384–1.424 Å and 1.381–1.439 Å), whereas the C3–O bond length is decreased (1.363–1.246 Å). Because these geometric features resemble those of the hydroxylated form, conversion of the substrate into the tripolar form predisposes this to facilitate hydroxylation.

In the reactant complex (Complex I), we replaced the ribityl chain with an ethanol group (CH\(_2\)CH\(_2\)OH) to simplify the large system. The complexation energy, that is, the energy difference between the reactant complex (Complex I) and the isolated reactants (FIHOOH and tripolar 5HN), is \(-16\) kcal mol\(^{-1}\). Thus FIHOOH and 5HN readily form a reactant complex. In optimised complex I, the FAD and 5HN planes are essentially parallel to each other (Figure 2a). The hydroperoxy group forms a hydrogen bond with the phenolic oxygen in 5HN, and the H–O distance is 1.630 Å. This interaction can have an effect on the barrier height of the subsequent hydroxyl transfer because in the transition state the H–O distance is 2.990 Å and the interaction thus much weaker (Figure 2b). It should be noted that the phenolic oxygen in 5HN also can form a hydrogen bond with the NH group of FIHOOH (H–O distance 1.939 Å, Figure 2a), but this interaction is not as strong as the interaction with the hydroperoxide group. However, this interaction can possibly affect the hydroxyl transfer because it becomes weaker in the product complex (H–O distance: 2.105 Å).

**TS and product complex:** In TS-I–II, the C2–O and O–O distances are 1.908 and 1.768 Å, respectively (see Figure 2b). The geometry of TS-I–II is similar to previous DFT studies on the corresponding reaction in PHBH.[26,27] However, we note that the imaginary frequency \(v_i = 659.7\) cm\(^{-1}\), obtained herein is higher than \(v_i = 400\) cm\(^{-1}\) reported for the hydroxylation step of PHBH. A likely cause is the use of a different functional in the calculations, but slightly different geometric structures at the TS will also have an impact. In a previous comprehensive study on hydroperoxides, Bach and Dmitrenko[27] found that the imaginary frequency of a transition state can be used to estimate whether the oxidative reaction of the hydroperoxide is a hydroxyl transfer (corresponding to a low imaginary frequency) or a direct oxygen transfer process (corresponding to a high imaginary frequency). According to this result, the current oxygenation reaction with an imaginary frequency \(v_i = 659.7\) cm\(^{-1}\) is still a hydroxyl transfer process. The hydroxyl transfer, which gives complex II, shares similarity with an S\(_{2}\)2 reaction, with FIHO\(^{-}\) as the leaving group.

We obtained an energy barrier of 33.8 kcal mol\(^{-1}\) for the hydroxylation reaction, which is higher than previous DFT studies by using the B3LYP method.[26] Albeit too high for an enzymatic reaction, it is consistent with the fact that B3LYP usually underestimates energy barriers, whereas the newer methods, such as MPWB1K, that are parametrised particularly for accurate evaluation of barrier heights, in...
The too-high barrier found herein can largely be ascribed to the fact that no explicit active-site residues, which might stabilise the TS through hydrogen bonding or electrostatic interactions, were included.

When optimising the product complex, this immediately rearranges to complex III instead of II (Scheme 1 and Figure 2c). It is known that in this type of hydroxyl-transfer reaction, the final product complex of the hydroxylation step is determined by the proton affinity (PA) of FIHO\(^-\) versus the hydroxylated intermediate.\(^{[26]}\) The calculated PA values of FIHO\(^-\) and compound I by using MPWB1K/6-311 + G(2d,2p) show that the PA of I (333 kcal mol\(^{-1}\)) is 12 kcal mol\(^{-1}\) lower than that of FIHO\(^-\) (345 kcal mol\(^{-1}\)) [Eqns. (3) and (4)]. This implies that complex III will be more stable than II, in accordance with the spontaneous rearrangement found.

\[
\text{FIHO}^- + H^+ \rightarrow \text{FIHOH} \quad \text{PA} = 345 \text{ kcal mol}^{-1} \quad (3)
\]

\[
\begin{align*}
\begin{array}{c}
\text{FIHO}^- + H^+ \\
\rightarrow \\
\text{FIHOH} \\
\begin{array}{c}
\text{PA} = 333 \text{ kcal mol}^{-1} \\
\end{array}
\end{array}
\end{align*}
\]

**Reaction mechanism:** Because the hydroxylation step is proposed to be an electrophilic aromatic substitution reaction,\(^{[13]}\) we analysed the atomic charges of the complexes in the hydroxyl-transfer process. The transferred hydroxyl retains a nearly neutral state in the reactant and transition state. An electron is transferred from 5HN to FIHO\(^-\) during the process, and in the transition state, the total atomic charges of the 5HN part and the FIHO\(^-\) part are close to –0.5 e each (Figure 2), in accordance with an electrophilic aromatic substitution mechanism.

However, by using IRC calculations to trace the conversion of the transition state to the product complex, the hydroxyl is found to rotate before binding to the carbon atom of 5HN. In the rotation, the hydroxyl proton is transferred back to FIHO\(^-\) to give product complex III. The energy surface for proton transfer from FIHO\(^-\) to 5HN (i.e., from complex III to II) was scanned, showing that the reaction is strictly endothermic along the reaction coordinate, even as the O–H length in complex II reached 0.98 Å. However, the total energy difference is only 1.5 kcal mol\(^{-1}\), which is much smaller than the PA difference (12 kcal mol\(^{-1}\)). This implies that the product complex is not simply determined by the PA values, and more factors must be considered. Therefore, we suggest that both complex II and III can be the product of the hydroxylation step, with complex III being slightly favoured (Scheme 1).

In the hydroxylation step, due to the SN2-like reaction, a chiral carbon atom is formed. In the enzyme’s active site pocket, the substrate will not be able to flip between different orientations (due to a hydrogen-bonding network stabilising the substrate), so we propose that there will be one stereoselective product in the hydroxylation step. In our calculation, product 1a was formed [Eq. (5)] because the initial orientation was chosen from the crystal structure. Therefore, the R configurations of 1 (1a) and 2 (1a protonated) were used for the calculations below. It should be noted that the stereochemistry of compounds 1 and 2 does not affect the configuration of the product in these model calculations.
The role of enzyme structure: As mentioned above, MHPCO is an unusual member in the aromatic hydroxylase family because after the hydroxylation, there is a further ring-cleavage process to give an acyclic product instead of an aromatic one.

According to Chaiyen et al.,\(^\text{[19,20]}\) the ring-opening process is faster than the rearomatisation so as to prevent the formation of the aromatic product. However, our calculations show that the C2–C3 bond strength in 2 is not that weak (> 40 kcalmol\(^{-1}\)). Moreover, when we optimised the geometry in water (optimisation with PCM model), the rearomatisation occurs spontaneously (Supporting Information, S3). To validate our discovery, we also studied the hydroxylated product of a typical aromatic hydroxylase, PHBH. The hydrogen on C2 readily leaves during the optimisation, and a charge distribution analysis shows that the leaving hydrogen carries a positive charge of 0.86e, that is, a proton. Thus we suggest that once the hydroxylated intermediate is released into aqueous solution, the rearomatisation occurs (Scheme 2). Hence, in the active site of MHPCO there should be a limited amount of water present around the substrate, and the hydroxylated intermediate is not released into water until the aromatic ring is cleaved. The enzyme structure study by McCulloch et al. shows that the active site pocket of MHPCO is significantly different from the other aromatic hydroxylases.\(^\text{[6]}\)

Previous ring-opening pathways: In previous studies by Chaiyen et al., it was suggested that after the hydroxylation, with the FIHO\(^{-}\) moiety leaving, the hydroxylated substrate (similar to 2) undergoes the ring-cleavage reaction with the help of water.\(^\text{[19,20]}\) Recently, McCulloch et al. proposed a new mechanism involving that the hydroxylated intermediate cleaves at the C2–C3 bond to give a ketene intermediate, which is further hydrated to the product.\(^\text{[20]}\) Both mechanisms (Scheme 3) have been evaluated in vacuo (in water, 2 is not stable, as mentioned above).

Our calculations show that the pathway proposed by McCulloch et al. is very difficult to obtain (\(\Delta G^\text{act} = 41.9\) kcalmol\(^{-1}\)). The carboxyl group in 2 is out of the ring plane, implying a poor conjugation (Figure 3a). As we mentioned above, in the PCM optimisation the proton on C2 leaves the molecule directly and carries a net positive charge (rearromatisation). In vacuo, this positive charge is distributed mainly on C2 (\(q = 0.42e\)) and C4 (\(q = 0.68e\)) rather than on the nitrogen atom as described in the Lewis structure.\(^\text{[24]}\) The bond lengths of N1–C2, C2–C3 and C3–C4 are 1.455, 1.526 and 1.451 Å, respectively (Figure 3a), which implies that the ring atoms in compound 2 are not conjugated to a delocalised system. The TS

\[\text{Scheme 3. Previous mechanisms of Chaiyen}^{[19,20]}\text{ and McCulloch}^{[6]}\]
In pathway A, the C2–C3 bond length of TS1–6a is increased to 1.864 Å, after which the ring cleaves to give the ketene intermediate 6a. The energy barrier of the reaction is 14.1 kcal mol⁻¹ (TS1–6a, ν₁ = 795.9 i cm⁻¹), which is much lower than the direct ring-opening reaction in Scheme 3 above (TS2–3, ΔG° = 41.9 kcal mol⁻¹). This implies that it is much easier for 1 to cleave the C2–C3 bond than it is for 2. We suggest that the large difference in barrier heights is due to the difference of electronic structure between 1 and 2. The frontier orbitals of 1 and 2 were studied (Supporting Information, S5). The LUMO orbitals are similar in 1 and 2, but the HOMO orbitals differ a lot. The HOMO orbital in 2 locates on the carboxyl group, so the C2–C3 bond is relatively stable. In 1, however, the HOMO has a large C2–C3 antibonding component, which explains the fact that the ring cleavage of 1 has much lower barrier than 2.

Unlike 2, the removal of the hydroxyl proton essentially makes the pyridine ring atoms and the carboxyl group atoms in 1 lie in the same plane, except for the sp²-hybridised carbon C2. Our geometry optimisation shows that this chiral carbon C2 is above the ring plane (Figure 5a). When using IRC calculations to trace the product from TS1–6a, we obtain the local minimum 6a'. This local minimum can easily reach the global minimum 6a through the C6–N1 single-bond rotation (Supporting Information, S4). The rotation barrier is only 0.2 kcal mol⁻¹, and the free energy difference between 6a' and 6a is −3.9 kcal mol⁻¹. In 6a', the C2–N1–C6–C5 dihedral angle is −63.1°, whereas in 6a, all the atoms are in the same plane (C2–N1–C6–C5 dihedral angle −179.9°) to form a large delocalised system, which enhances the stability of 6a over 6a'.

Current ring-opening pathways: When using the favoured product of the hydroxylation step, compound 1, we find two different pathways leading to the acyclic product (Scheme 4). Compared with Scheme 3, these two pathways display much lower barrier heights.
In pathway B, compound 1 first forms an epoxy transition state (TS1–7, \( \nu_i = 231.5 \text{ cm}^{-1} \)) to give 7 (Figure 5c–d). The energy barrier is 11.1 kcal mol\(^{-1}\), which is slightly lower than that of pathway A. With the weak C3–O bond broken (TS7–6b, \( \Delta G^\ddagger _{298} = 0.5 \text{ kcal mol}^{-1} \), \( \nu_i = 345.0 \text{ cm}^{-1} \)) we obtain the local minimum 6b' because the epoxy transition state constrains the direction of the carbonyl C=O bond of the amide pointing towards the ketene group (Figure 5c). The 3 kcal mol\(^{-1}\) lower barrier of pathway B implies that it is kinetically favoured compared with pathway A.

We found that in epoxy transition states TS1–7, the C2–C3 bond length is shorter than in 1 (1.525 versus 1.538 Å), because during the C3–O bond formation, C2 is pulled back to the pyridine plane (in 1 it is above the plane). Through the rotation of the O=C2–H plane in 1, the hydrogen on C2 moves to a cis position of the hydrogen on N1 (IRC calculation, see Supporting Information S6). This cis configuration is maintained in pathway B. The geometry of TS7–6b (not shown) is almost the same as 7, only except for an increase in the C3–O distance (1.460–1.671 Å). In 6b', the C2-N1-C6-C5 dihedral angle is 64.7°, which is similar to 6a' but in opposite direction. The conversion of 6b to 6b' is similar to that in pathway A (\( \Delta G^\ddagger _{298} = 0.5 \text{ kcal mol}^{-1} \), \( \Delta G^\ddagger _{298} = -4.9 \text{ kcal mol}^{-1} \)).

C2–N1 bond rotation: It should be noted that in the direct ring-opening process (TS1–6a) the two hydrogen atoms on N1 and C2 are in trans positions, whereas through the epoxy
transition state (TS1–7) the two hydrogen atoms are in cis positions (Figure 5b,c). Hence these two pathways reach different configurations of ketene intermediates 6a and 6b.

Compounds 6a and 6b are stoichiometrically identical, and react with water to give the same carboxylic acid product. However, they give two different configurations of the product (A and B) that are observable by 1H NMR spectroscopy because the C2–N1 bond rotation in the product is slower than the 1H NMR time scale and the chemical shifts of the hydrogen atoms in A and B are different. To distinguish between 6a and 6b, we introduce the cis/trans notion, which describes isomers of alkenes; 6a is labelled trans and 6b is labelled cis. The atoms in the amide (OC2-NH-C6) are all in the same plane, again similar to alkenes (Scheme 5).

The lone pair on N1 is conjugated with the C5=C6 bond, the C2=O bond and also the carboxyl group to form a large delocalised system (O=C2–N1–C6–C5–COO−) in 6a and 6b, which restricts the C2–N1 bond rotation. Our calculation shows that the rotation barrier is more than 20 kcal mol−1 in vacuo (6a−6b ΔDG298 = 24.6 kcal mol−1, and 6b−6a ΔDG298 = 20.8 kcal mol−1), implying a slow process. In TS6a−6b, the N1 atom adopts an sp2 configuration rather than the sp3 configurations in 6a and 6b. The O=C2–N1–C6 dihedral angle is −61.9°, and there is little overlap between the π orbitals in the aldehyde and the π orbitals in the rest of molecule (Supporting information, S7).

When 6a and 6b are released into water, products A and B are obtained. We found that the hydrogen attached to nitrogen and the newly formed carboxyl interact by a hydrogen bond giving a seven-membered ring, with an O–H distance of 1.85 Å in both A and B (Supporting Information, S8). As in 6a and 6b, the large delocalised system (O=C2–N1–C6–C5–COO−) is present also in A and B.

Our calculation shows that A is 0.6 kcal mol−1 higher than B in energy, whereas in vacuo 6a is 3.8 kcal mol−1 lower than 6b in energy. This implies that the environment can affect the preference of different configurations. In Chaiy-en’s study, the cis product B is reported to be the major one in D2O, and the ratio of B to A is roughly 3:1, which is explained by the current calculations. By using Boltzmann statistics (e−ΔG/kT), an energy difference of 0.6 kcal mol−1 corresponds to a 2.8:1 ratio between products B and A at T = 298 K.

The C2–N1 bond rotation is as slow as that in the ketene intermediates (A→B ΔDG298 = 21.7 kcal mol−1, and B→A ΔDG298 = 21.2 kcal mol−1). In TS5A–B, the intramolecular hydrogen bond length is shortened (1.859–1.786 Å). The O=C2–N1–C6 dihedral angle is −111.2°, implying poor conjugation between the aldehyde and the C6–C5 double bond.

Alternative ring opening: Besides the two pathways mentioned above, we found that after hydroxylation, the C2–N1 bond in compound 1 can also cleave to give imine product 8 [Eq. (6)].

This conversion is almost barrierless (ΔDG298 = 0.5 kcal mol−1). However, because the energy difference between 1 and 8 is only 6.9 kcal mol−1, compound 8 can also convert back to 1 (ΔDG298 = 7.4 kcal mol−1). We thus propose that in the enzyme this is a reversible reaction because the energies of the ketene intermediates are much lower than 8 (the energy difference between 8 and 6a is 18.2 kcal mol−1, and the energy difference between 8 and 6b is 14.4 kcal mol−1). It should be noted that in principle 8 can also react with water to give α-hydroxyl amine or rearrangement products. Because no such products have been detected, we suggest that water addition probably does not occur in the active site, which is consistent with Scheme 2.

By analysing the reactions of 1 in the presence of FIHOH to mimic the processes in the enzyme active site, we found that FIHOH can stabilise 1. The barrier heights for conversions of 1 to 6a, 7 and 8 are increased (Table 1). The structures of the reactant and TS complexes reveal that the most probable reason is the strong hydrogen bond (O–H distance 1.464 Å) formed in complex III (Figure 2c). In the transition states of the different pathways, this hydrogen bond is much weaker (Supporting Information, S9). Therefore, the hydrogen-bonding interaction is important for maintaining 1, and

![Scheme 5. Product rotations.](image)

Table 1. Comparison of the energy barriers for the ring-opening step in the presence of FIHOH.

<table>
<thead>
<tr>
<th>Conversion</th>
<th>Without FIHOH</th>
<th>With FIHOH</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ΔDG298 [kcal mol−1]</td>
<td>ΔDG298 [kcal mol−1]</td>
</tr>
<tr>
<td>1→6a</td>
<td>14.1</td>
<td>20.7</td>
</tr>
<tr>
<td>1→7</td>
<td>11.1</td>
<td>16.0</td>
</tr>
<tr>
<td>1→8</td>
<td>0.5</td>
<td>7.3</td>
</tr>
</tbody>
</table>


affects the barrier heights for the different ring-opening pathways. In the enzyme, the conversions of 1 to 6a and 6b might be assisted by the other residues in the active site (work in progress).

Conclusions

The hydroxylation and ring-opening mechanism of MHPCO is studied by computational quantum chemistry. In the hydroxylation, a hydroxy group is transferred from the activated flavin (FlHOOH) to the substrate through aromatic electrophilic substitution, which is similar to the other aromatic hydroxylases. A barrier height of 33.8 kcal mol⁻¹ [MPWB1K/6-31+G(d,p)] is obtained for this step, which is higher than the 24.9 kcal mol⁻¹ [B3LYP/6-31+G(d,p)] reported by Bach et al.[26] in a previous study on the PHBH mechanism. The hydroxy proton is spontaneously transferred back between the hydroxylated substrate and the remaining FlHO⁺ moiety. Given the small energy difference, both 1 and 2 [Eq. (4)] can be the product of the hydroxylation step; Our calculations show that complex III (FIHOH and 1) is more stable than II (FIHO⁻ and 2).

The ring opening occurs inside the enzyme active site rather than in water because the rearomatisation is predicted to take place immediately, without any barrier, to give an aromatic product. This has, however, not been detected as the product of MHPCO. The instability of the hydroxylated intermediate of MHPCO is the main reason that the acyclic products are formed.

In the ring-cleavage reaction, compound 1 is more likely to cleave the ring than 2 (ΔAG<sub>298</sub> < 15 kcal mol⁻¹ for 1 versus ΔAG<sub>298</sub> > 35 kcal mol⁻¹ for 2). Two possible ring-opening pathways of 1 are found. In pathway A, the ring opens directly (ΔAG<sub>298</sub> = 14.1 kcal mol⁻¹) to give 6a, the trans configuration of the ketene intermediate. In pathway B, the cis configuration 6b is obtained by way of an epoxy transition state (ΔAG<sub>298</sub> = 11.1 kcal mol⁻¹). The transformation between 6a and 6b is restricted (ΔAG<sub>eq</sub> > 20 kcal mol⁻¹) due to the large conjugated fragment (O=C₂–N₁–C₆=C₅–COO⁻). The transformation between hydrated product A (trans) and B (cis) is slow (ΔAG<sub>298</sub> > 20 kcal mol⁻¹), and can be observed by ¹H NMR spectroscopy.[20] Our calculations show that B is slightly more stable than A, which is in accordance with the fact that B is the major product. The barrier heights found are in line with those reported for ketosteroid isomerase (∼10.3 kcal mol⁻¹), implying that the rate-limiting step for the current system is indeed the initial hydroxylation rather than the subsequent ring-opening reaction.[20]

The C₂–N₁ bond in 1 can also be cleaved to give imine 8 (ΔAG<sub>298</sub> = 0.5 kcal mol⁻¹). Because no other products than those originating from 6a and 6b were detected, and the energy of the ketene products (6a and 6b) are much lower than 8, this must be a reversible reaction in the enzyme. We furthermore propose that the subsequent water addition probably does not occur in the active site because the reaction of 8 with water would give other products than those detected. FIHOH can stabilise I through hydrogen bonding, and the energy barriers are increased for all ring-opening pathways in the presence of FIHOH. This implies that the hydrogen-bonding interaction is important for the ring-opening process.

Acknowledgements

The Swedish Science Research Council (VR), the Faculty of Science and Engineering at Örebro University, and the National University of Ireland—Galway are gratefully acknowledged for financial support.

[32] L. Ridder, J. N. Harvey, I. M. C. M. Rietjens, J. Vervoort, A. J. Mul-
   494.
   211.
[40] Gaussian 03, Revision B.01, M. J. Frisch, G. W. Trucks, H. B. Schleg-
   gel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgom-
   ery, Jr., K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Izengar, J.
   Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega,
   G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R.
   Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao,
   H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross,
   C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev,
   A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K.
   Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, G. Zakrzew-
   ski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K.
   Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz,
   Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G.
   Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox,
   T. Keith, MvA. Al-Laham, C. Peng, Y. A. Nanayakkara, M. Challas-
  combe, P. M. W. Gill, B. Johnson, W. Chen, M. Wong, W. C. Gonzalez,
   266, 17341.
[42] The charges on the carbon atoms are Mulliken atomic charges with
   hydrogen atoms summed into the heavy atoms.
   30, 10849.

Received: August 13, 2009
Published online: January 11, 2010
Paper 4
Catalytic Roles of Active-Site Residues in 2-Methyl-3-hydroxypyridine-5-carboxylic Acid Oxygenase: An ONIOM/DFT Study

Boxue Tian, † Åke Strid, ‡ and Leif A. Eriksson*, †

†School of Chemistry, National University of Ireland—Galway, Galway, Ireland
‡Örebro Life Science Center, School of Science and Technology, Örebro University, 701 82 Örebro, Sweden

Supporting Information

ABSTRACT: The catalytic mechanism of 2-methyl-3-hydroxypyridine-5-carboxylic acid (MHPC) oxygenase (MHPCO) has been systematically studied using DFT and ONIOM(DFT:MM) methods. MHPCO catalyzes the hydroxylation and subsequent ring-opening of the aromatic substrate MHPC to give the aliphatic product α-(N-acetylaminomethylene)succinic acid (AAMS). Our calculations show that the active-site residues Arg211 and Tyr223 have a minor effect on the reaction, while the peptide bond of Pro295-Ala296, the side chain of Tyr82 and several crystal water molecules affect the reaction energy profile considerably. Both DFT and ONIOM calculations show that the ring-opening pathway B, in which an epoxy transition state is formed, is more favored than the direct C2–C3 cleavage pathway A. Different QM/MM partitioning schemes have been used to study the enzymatic reaction, and the results show that both the reaction barriers for the hydroxylation and the ring-opening pathways are sensitive to the QM/MM partitioning.

1. INTRODUCTION

2-Methyl-3-hydroxypyridine-5-carboxylic acid (MHPC) oxygenase (MHPCO) is an FAD-containing monooxygenase involved in the bacterial degradation of vitamin B61–3 (see ref 2 for review). MHPCO catalyzes the hydroxylation and subsequent ring-opening of the aromatic substrate MHPC to give the aliphatic product α-(N-acetylaminomethylene)succinic acid (AAMS; eq 1).

It has been shown that the hydroxylation mechanism is similar to that of aromatic hydroxylases,8 such as p-hydroxybenzoate hydroxylase (PHBH).9 With the help of NADPH and molecular oxygen, the cofactor FAD converts to a reactive intermediate C-(4a)-flavinhydroperoxide (FADHOOH), which then hydroxylates the triprole form of the substrate MHPC via SN2-like electrophilic aromatic substitution.6 Due to the instability of the hydroxylated intermediate, a ring-cleavage reaction follows (Scheme 1), the product of which is different from the normal aromatic products of aromatic hydroxylases.

The hydroxylation step of PHBH has been studied extensively with both QM8,9 and QM/MM9–13 methods. Bach et al.7 investigated the mechanism of C-(4a)-hydroperoxylavin oxidation of a series of heteroatom nucleophiles with several QM methods which provided insight on how to distinguish between oxygen atom transfer and hydroxyl transfer mechanisms. Afterward, they modeled the hydroxylation mechanism of PHBH with inclusion of two active-site residues,13 Arg214 and Tyr201 (in MHPCO, the corresponding residues are Arg211 and Tyr223), which were shown to play minor roles in lowering the reaction barrier. Ridder et al. performed QM/MM calculations on PHBH9,11 and phenol hydroxylase,10 suggesting that the amide oxygen of the active-site proline (Pro293 in PHBH, Pro364 in phenol hydroxylase, and Pro295 in MHPCO) can stabilize the TS by 2–3 kcal/mol.11 Senn et al.12 studied the hydroxylation of PHBH by using QM/MM molecular dynamics methods and analyzed the hydrogen-bonding network in the active site along the reaction coordinate of OH transfer. Interestingly, they found that the hydrogen bond between the Pro293 amide oxygen and the proton in the transferred hydroxyl varied substantially, making the role of the Pro293 backbone carbonyl inconclusive. Recently, Mata et al.13 performed more accurate QM/MM calculations on the PHBH system using higher level methods such as LMP2 and LCCSD for the treatment of the QM region. They obtained a barrier height of ~14 kcal/mol, in accordance with the experimental value (12–15 kcal/mol depending on pH).

It should be mentioned that the monoxygenation catalyzed by aromatic hydroxylases share similarities with that catalyzed by heme-containing monooxygenases (some of the P450 enzymes),
which have been extensively studied by theoretical methods in the past decade.\textsuperscript{14–21} Aromatic hydroxylases and heme-containing monooxygenases are all able to activate molecular oxygen (to form FADHOOH and FeOOH intermediates respectively) and incorporate only one of the oxygen atoms into the substrates. However, differences also exist. For instance, in heme-containing monooxygenases, the important intermediate FeOOH will first release water to form FeO intermediate,\textsuperscript{14} which will further oxygenate the substrate. In the case of aromatic hydroxylases, FADHOOH transfers the hydroxyl cation (via heterolytic cleavage of the O–O bond) to the substrate, prior to releasing water. In addition, aromatic hydroxylases utilize FAD as cofactors, and thus these enzymes are metal-free.

The catalytic mechanism of MHPCO has previously been investigated in our group.\textsuperscript{22} We performed DFT calculations on a small model, containing only FADHOOH and the substrate. We found that, in the hydroxylation, the hydroxyl proton is spontaneously transferred back from the hydroxylated substrate to the remaining FADHO\textsuperscript{−} moiety, to generate the more stable intermediate complex 3 (Scheme 1). This is different from the hydroxylation product of PHBH. Two ring-cleavage pathways (pathway A and B) of 3 were found, the barriers of which are more reasonable than those obtained using previously proposed pathways by Chaiyen\textsuperscript{3,6} and McCulloch.\textsuperscript{23}

However, several issues remain to be resolved regarding the MHPCO catalytic mechanism. First, in our previous study, the obtained free energy of activation for the rate-limiting step (the hydroxylation step) was too high, compared to the experimental value (33.8 kcal/mol vs 16.3 kcal/mol, using rate constant from ref 4). One possible reason is the small model system, containing only the substrate and FADHOOH, and that the calculations were performed in the gas phase without considering the effect of the protein environment. Therefore, it is of interest to investigate if the energy barrier of the hydroxylation step will be lowered when the protein environment is taken into account through QM/MM calculations.

Second, the recent crystal structure of MHPCO shows that the binding of MHPC in the active site is different from PHBH.\textsuperscript{23} It was suggested that the active-site residues Tyr82, Tyr223, Arg211, and crystal water molecules may be important for the catalysis. Hence, an additional aim of the current study is to provide theoretical evidence to identify important active-site residues for the enzymatic reaction.

Moreover, in previous gas-phase calculations, the barrier heights for the two ring cleavage pathways were shown to be very close. Therefore, it is of significance to distinguish which pathway is more favorable by performing more accurate calculations as a means to better understand the catalytic mechanism.

As a continuation of our previous study, we have to this end performed extensive DFT and ONIOM(DFT/MM) calculations to address the above issues.

2. COMPUTATIONAL DETAILS

2.1. Preparation of the Model System. The crystal structure of MHPCO with the substrate bound was obtained from the RCSB Protein Data Bank (PDB code 3GMC\textsuperscript{23}). Hydrogen atoms were added to the crystal structure using the Molecular Operating Environment (MOE) software.\textsuperscript{24} FAD was manually modified to FADHOOH (S configuration), and the substrate MHPC was set to its tripolar form, in accordance with the experimental evidence.\textsuperscript{6} The protonation states of the rest of the system was determined through the MOE Protonate 3D function at pH = 7.0. The Gaussian 09 program\textsuperscript{25} was used for all subsequent calculations.

Two model systems were investigated (Table 1). The first model system (S\textsubscript{1}), which is used for determining the QM/MM
**Scheme 2. Active Site of MHPCO**

All 149 atoms in the figure were included in the $S_i$ model, DFT-only calculations, and the red-labeled atoms were included as the QM part in the largest ONIOM model calculation (96 QM atoms).

To evaluate the energetic contribution of a certain active-site residue to the reaction, we removed that residue from $S_1$ (or set those atoms as dummy atoms) and recalculated the single-point energy without geometry optimizations, so as to compare the difference in energy profile between the $S_i$ system and the system without that residue ($S_i$-Res). The mean unsigned error (MUE) between the relative energies of $S_i$ and $S_i$-Res was used to assess the importance of the residue to the reaction energy profile. For example, in the $S_1$ system, we located all intermediates and TSs in Scheme 1, and the relative energies are $E_1$, $E_{TS-2}$, $E_3$, ..., $E_{TS-8}$, $E_9$. When we remove the one residue, for example Arg211, we obtain the system $S_1$-Arg211, with the relative energies $E_1'$, $E_{TS-2}'$, $E_3'$, ..., $E_{TS-8}'$, $E_9'$. The MUE of residue Arg211 is then evaluated according to eq 2:

$$\text{MUE}_{\text{Arg211}} = \frac{|E_{TS-2}' - E_{TS-2}| + |E_3' - E_3| + ... + |E_{TS-8}' - E_{TS-8}|}{8}$$

The residues that have larger MUE will hence be more important than those with smaller MUE, in terms of the contribution to the reaction energy profile. Given the size of the system, BSSE corrections were not included. It should be noted that this strategy has not been reported before.

**2.3. Computational Details for the $S_i$ System: ONIOM Calculations on the Full Enzyme.**

The $S_i$ system was initially optimized with the Amber 03 force field. Missing parameters of the substrate MHPC and the cofactor FADHOOH were derived by Antechamber, and the partial atomic charges were assigned using the AM1-BCC charge method. The charges on the OOH moiety of FADHOOH are OS ($-0.3280$), OH ($-0.2121$), and HO (0.2132); for details, see Supporting Information, S2.

A two-layer ONIOM (QM/MM) scheme, in which the interface between the QM and MM regions is treated by hydrogen link atoms, was then used for the calculations on the $S_i$. The ONIOM-ME (mechanical embedding) scheme was used for geometry optimizations and frequency calculations, and the ONIOM-EE (electronic embedding) scheme was used for single-point energy calculations on ONIOM-ME optimized geometries. Geometries were not optimized with the ONIOM-EE scheme due to its high computational cost. Geometries of intermediates and TSs in Scheme 1 were optimized using the ONIOM(DFT/BS1:Amber) method (Table 1), followed by frequency calculations at the same level of theory. Atoms far away from the substrate (MHPC) atoms (>15 Å) were kept fixed during the optimization. Single-point energy calculations with the BS2 basis set were performed on the optimized geometries with both ONIOM-ME and ONIOM-EE schemes. Eight different QM/MM partitioning schemes were used, and the largest QM part contains 96 QM atoms, as described in Scheme 2. The partitioning schemes are summarized in Table 2. For the ONIOM-I calculations, both B3LYP and M062X methods were used for the QM layer, while for the other partitioning schemes, only the B3LYP functional was used to treat the QM part.

**3. RESULTS AND DISCUSSION**

**3.1. QM Calculations, $S_i$ Model: Active-Site Residue Decomposition Analysis.** Before presenting the results, we will briefly summarize the results of our previous DFT model study. The catalytic process of MHPCO can be divided into two stages, hydroxylation and ring cleavage. The hydroxylation step is...
Table 2. Different ONIOM Partitioning Schemes Employed

<table>
<thead>
<tr>
<th>ONIOM calculations</th>
<th>QM region</th>
<th>QM atoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>ONIOM-1</td>
<td>FAD, MHPC, Pro295→Ala296, Tyr82, six water molecules</td>
<td>96</td>
</tr>
<tr>
<td>ONIOM-2</td>
<td>FAD, MHPC, Pro295→Ala296, six water molecules</td>
<td>81</td>
</tr>
<tr>
<td>ONIOM-3</td>
<td>FAD, MHPC, Tyr82, six water molecules</td>
<td>88</td>
</tr>
<tr>
<td>ONIOM-4</td>
<td>FAD, MHPC, six water molecules</td>
<td>73</td>
</tr>
<tr>
<td>ONIOM-5</td>
<td>FAD, MHPC, Pro295→Ala296, Tyr82</td>
<td>78</td>
</tr>
<tr>
<td>ONIOM-6</td>
<td>FAD, MHPC, Pro295→Ala296</td>
<td>63</td>
</tr>
<tr>
<td>ONIOM-7</td>
<td>FAD, MHPC, Tyr82</td>
<td>70</td>
</tr>
<tr>
<td>ONIOM-8</td>
<td>FAD, MHPC</td>
<td>55</td>
</tr>
</tbody>
</table>

Table 3. Mean Unsigned Error (MUE)\textsuperscript{a} of Active-Site Residues and Waters

<table>
<thead>
<tr>
<th>residue/water</th>
<th>MUE (B3LYP/BS1)</th>
<th>MUE (B3LYP/BS2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pro295→Ala296</td>
<td>10.4</td>
<td>8.8</td>
</tr>
<tr>
<td>Tyr82</td>
<td>12.0</td>
<td>8.1</td>
</tr>
<tr>
<td>Arg211</td>
<td>2.2</td>
<td>2.1</td>
</tr>
<tr>
<td>Tyr223</td>
<td>1.0</td>
<td>0.6</td>
</tr>
<tr>
<td>W654</td>
<td>7.2</td>
<td>5.5</td>
</tr>
<tr>
<td>W410</td>
<td>2.6</td>
<td>1.7</td>
</tr>
<tr>
<td>W435</td>
<td>4.9</td>
<td>3.9</td>
</tr>
<tr>
<td>W493</td>
<td>4.9</td>
<td>3.7</td>
</tr>
<tr>
<td>W546</td>
<td>1.8</td>
<td>1.5</td>
</tr>
<tr>
<td>W547</td>
<td>3.3</td>
<td>2.7</td>
</tr>
<tr>
<td>W633</td>
<td>5.4</td>
<td>4.0</td>
</tr>
<tr>
<td>W642</td>
<td>2.7</td>
<td>2.2</td>
</tr>
</tbody>
</table>

\textsuperscript{a}As defined in eq 2, in kcal/mol.

similar, but not identical, to the one in PHBH. In the hydroxylation, we found that the obtained intermediate transferred the hydroxyl proton back to the FADHO\textsuperscript{−} moiety (Scheme 1), resulting in intermediate 3. Furthermore, we found that intermediate 3 is more likely to cleave the C2−C3 bond than in the case of its protonated form (ΔΔG\textsubscript{298} > 15 kcal/mol vs ΔΔG\textsubscript{298} > 35 kcal/mol). It should be mentioned that, in our previous study,\textsuperscript{22} the substrate 5-hydroxyisocitric acid (SHN) was used instead of the native substrate MHPC. The difference between MHPC and SHN is the substituent at the C2 position, which in MHPC is a methyl group and in SHN a hydrogen atom. In the previous study, the smaller SHN substrate was used as the emphasis was on explaining why the product of MHPCO is acyclic rather than aromatic. We reported two different ring-cleavage pathways (Scheme 1). In pathway A, the C2−C3 bond cleaves directly to get S, while in pathway B, a seven-member ring intermediate 7 is obtained by way of a C2−O−C3 epoxy transition state TS-6, followed by the cleavage of the C3−O bond. It should be noted that S and 9 are stoichiometrically identical, and can convert between each other by rotation of the C2−N1 bond, as observed by \textsuperscript{1}H NMR spectroscopy.\textsuperscript{6} The ketene (S or 9) will then be hydrated to give the product AAMS.

Through the QM calculations on the S\textsubscript{1} system, the aim is to identify residues that have a major effect on the potential energy profile along the reaction paths and treat those residues quantum mechanically in the ONIOM calculations. In a recent study, Hu et al.\textsuperscript{15} discussed how residues may be selected when modeling enzyme reactions with QM-only methods. They suggested that to accurately model the influence of the surroundings on enzymatic reaction energies, a detailed and unbiased atomic account of the surroundings needs to be included (at least up to 10−15 Å from the active site). In our S\textsubscript{1} model, the residues attaining hydrogen bond to the substrate MHPC were all included. However, the hydrophobic residues were excluded to our model, as they are not expected to substantially affect the reaction pathways. Moreover, the main objective of the S\textsubscript{1} model is to identify important residues rather than accurately calculate the reaction energies (the S\textsubscript{2} model should be more accurate than the S\textsubscript{1} model in terms of reaction energies), plus that inclusion of the hydrophobic residues will significantly increase the computational demand.

All intermediates and TSs in Scheme 1 were optimized (Cartesian coordinates see Supporting Information, S4). We removed each residue and recalculated the energy of S\textsubscript{1,Res} as specified above, and then calculated the MUE (eq 2). The MUE of the active-site residues are summarized in Table 3. According to Figure 1 and Table 3, we found that the magnitudes of the MUE values for the BS2 basis set were smaller than the results with the BS1 basis set for all the cases investigated, implying that the use of different basis sets can influence the relative energies considerably. Hence, the larger basis set is more suitable for the purpose of this study, since errors on relative energies are generally reduced (becomes more stable) when a higher basis set is employed (Table 3).

Comparing the relative energy profiles of S\textsubscript{1} and S\textsubscript{1,Res} the peptide bond of Pro295→Ala296 and the side chain of Tyr82 are found to affect the energy profile significantly (MUE values 8.8 and 8.1 kcal/mol, respectively, with the BS2 basis set). All intermediates and TSs were stabilized (lower relative energy) with Pro295→Ala296 removed, while Tyr82 was shown to play an opposite role (Figure 2). Arg211 affects the energy profile in a minor scale (MUE 2.1 kcal/mol, BS2), while Tyr223 has nearly no effect on the reaction (MUE 0.6 kcal/mol, BS2). In addition, several water molecules, such as W654, W435, W493, and W633 (MUE values are 5.5, 3.9, 3.7, and 4.0 kcal/mol, respectively), might be important for the reaction, although the errors obtained are of such a magnitude that they may also be attributed to the loss of hydrogen bonding. It should be mentioned that Pro295→Ala296 and Tyr82 affect not only the relative energies of various points on the PES (MUE), but also the barrier heights (see below).

The obtained results are consistent with the earlier DFT study on PHBH,\textsuperscript{6} namely that Arg211 and Tyr 223 may be important for substrate binding, but not for the reaction process. Accordingly, based on the MUE values and the study by Bach and Dmitrenko,\textsuperscript{8} in the following ONIOM calculations, Arg211 and Tyr223 as well as two water molecules (W410 and W546, MUE values are 1.7 and 1.5 kcal/mol) which have minor effects on the reaction will not be treated as part of the QM region. It should be noted that, in the ONIOM calculations, the side chain atoms of
Pro295 and Ala296 were not included in the QM part (Scheme 2), as the corresponding atoms have no direct hydrogen-bonding interactions with neither MHPC nor FAD.

According to our calculations, the hydroxylation barrier was lowered from 31.5 kcal/mol for SI to 26.4 kcal/mol when Pro295-Ala296 was removed, suggesting that the hydrogen bond between the amide oxygen and the peroxide (OOH) proton stabilizes the reactant more than the TS. This does not support the QM/MM study by Ridder et al., who suggested that the conserved proline peptide bond can stabilize the TS of the hydroxylation step. A possible reason is that the MHPC system is different from the PHBH one in terms of substrate, active-site interactions, and residue orientations. In the active site of MHPC, there are more water molecules than in PHBH, and the substrate interacts with these water molecules instead of a direct interaction with residues such as Tyr82, Tyr223, and Arg211.

Besides the hydroxylation step, Pro295-Ala296 affects the barrier height of TS-8 significantly; however, TS-8 was not found in the ONIOM calculations (probably the conversion of 7 to 9 is barrierless). Tyr82 has minor effect (less than) on the hydroxylation step, but can affect the barrier heights of the ring-cleavage pathways (3 and 5 kcal/mol for TS-4 and TS-6 respectively).

### 3.2. ONIOM Calculations, SII Model

#### 3.2.1. Geometries of ONIOM-1 Calculations

We will focus on the ONIOM(B3LYP:Amber) geometries in the ONIOM-1 calculations (96 QM atoms), even though geometries were also optimized with the ONIOM(M062X:Amber) method. The Cα-rmsd between the crystal structure and the structure optimized at the ONIOM-(B3LYP:Amber) level is 0.61 Å, and the geometry differences are mainly in the loop regions and terminals, far away from the active site. In Figure 3, we can see that due to the modification of FAD to FADHOOH, the Pro295-Ala296 backbone moves toward the peroxide proton and forms a hydrogen bond after optimization (Figure 3); this constitutes the most significant change of the active-site residues. As noted in section 3.1, this H-bond was found also in the S1 model and was concluded to stabilize the reactants.

In the hydroxylation reaction, the peroxide proton forms a hydrogen bond in 1 with the amide oxygen of Pro295-Ala296 (H⋯O distance 1.750 Å, Figure 4a); a hydrogen bond that remains in TS-2 (H⋯O distance 1.780 Å, Figure 4b). W654 links FAD and MHPC through hydrogen bonds. Tyr82 forms an indirect hydrogen bond with the N1 proton in the substrate. In TS-2, the hydroxyl lies in the middle between FAD and MHPC (O⋯C and O⋯O distances being 1.940 and 1.958 Å, respectively. It should be mentioned that, in a very accurate model for the PHBH enzyme, the average O⋯C and O⋯O distances in the transition state are 2.10 and 1.83 Å (B3LYP/GROMOS geometry). A likely reason for the geometry differences (O⋯C and O⋯O distances) between TS-2 in our study and the transition state reported by Mata et al. could be that the MHPCO system is not the entirely the same as the...
The Journal of Physical Chemistry B

The hydrogen bond between W654 and the negatively charged oxygen in the oxygenated intermediate.

3.2.2. Potential Energy Profiles of ONIOM-1 Calculations.

The ONIOM-EE (BS2) energies, in which the electrostatic interactions from the protein are incorporated into the QM calculations, were used to characterize the potential energy profile along the reaction coordinates (Figure 6). The ONIOM-ME (BS1 and BS2) energies are listed in Supporting Information, S3. From Figure 6, we note that the ONIOM-1 calculations with B3LYP:Amber and M062X:Amber do not show any significant differences, except for TS-2. The M062X:Amber obtains the energy barrier 47.1 kcal/mol for the hydroxylation, while for B3LYP: Amber, the value is much lower (32.2 kcal/mol, ONIOM-EE, BS2). It should be noted that the in the gas-phase model which contains only MHPC and FADHOOH, the energy barrier for the hydroxylation is lower than in the ONIOM model for both methods using BS2 (B3LYP: 24.7 kcal/mol and M062X 28.8 kcal/mol, Supporting Information, S5). As the experimental value is only 16.7 kcal/mol, there must be additional factors that are not considered in our system, such as the dynamics of the enzyme, or that the single configuration (crystal structure) of the protein available to date may not be sufficiently accurate to describe the MHPCO reaction.

Comparing the ONIOM(B3LYP:Amber) calculations with the QM (B3LYP) calculations, the activation energies for the hydroxylation are similar (32.2 kcal/mol vs 31.5 kcal/mol), suggesting that the MM layer has minor effect on the hydroxylation step. However, for the ring-opening pathways, the results show several differences. First, according to the QM calculations, the intermediate 5 (product of pathway A) is more stable than 9 (product of pathway B); however, in the ONIOM calculations, 9 becomes more stable. Second, the energy barriers obtained from QM (B3LYP) calculations (22.0 kcal/mol for pathway A and 16.6 kcal/mol for pathway B) are considerably higher than those from ONIOM-1 (B3LYP:Amber) calculations (14.3 kcal/mol for pathway A and 5.6 kcal/mol for pathway B), suggesting that the ring-opening process becomes easier in the presence of the PHBH system in terms of both substrate and protein environment. Moreover, a back-transfer process of the hydroxyl proton was observed in this study, which was not reported in PHBH. When the hydroxyl attaches to C2 of MHPC, the hydroxyl proton transfers back to the FAD moiety and forms a hydrogen bond with the transferred hydroxyl oxygen in intermediate 3 (H···O distance 1.769 Å, Figure 4c), consistent with the gas-phase calculations. The hydrogen bond between W654 and the phenolic oxygen of MHPC becomes weaker in 3 compared with in 1 (1.652 Å vs 2.016 Å), suggesting that this oxygen is less negatively charged in the oxygenated intermediate.

In the ring-opening pathway A, the C2···C3 bond in MHPC cleaves directly to give intermediate 5. In TS-4, the C2···C3 distance increases to 2.058 Å, and the hydrogen bond between W654 and phenolic oxygen becomes much weaker than in 3 (H···O distance 2.920 Å vs 2.016 Å, Figure 5a). The breaking of this hydrogen bond may increase the barrier height for this path, and will be discussed in more detail later. In the ketene intermediate 5, W654 forms a strong hydrogen bond (H···O distance 1.663 Å) with the newly formed carbonyl group (Figure 5b), which in turn stabilizes this intermediate. The hydrogen bond of Tyr82 is unchanged between 3 and TS-4.

In ring-opening pathway B, the oxygen on C2 moves toward C3 and forms an epoxy-like transition state TS-6 (O···C3 distance 1.717 Å, Figure 5c), followed by cleavage of the C2···C3 bond to give intermediate 9 (Figure 5d). It should be mentioned that we were unable to locate intermediate 7, probably because the conversion from 7 to 9 is barrierless in the protein environment (the barrier for this conversion is <1 kcal/mol in our previous gas-phase study). In TS-6, W654 forms a strong hydrogen bond with the phenolic oxygen in MHPC (H···O distance 1.698 Å, Figure 5c), this being much weaker in both 3 and 9 (H···O distances are 2.016 and 2.047 Å, respectively) which implies the oxygen is more negative in the TS than in either reactant or intermediate. The hydrogen bond between Tyr82 and the substrate in 3 and TS-6 again does not change significantly, implying that the Tyr82 does not affect the ring-opening energy barriers via forming and breaking hydrogen bonds. Therefore, it is likely that Tyr82 affects the reactions through electrostatic interactions.

Figure 4. Geometries of ONIOM-1 calculations: (a) TS-2, and (b) and (c) 3.
The barrier difference between the two pathways (5.4 kcal/mol in QM vs 8.7 kcal/mol in ONIOM) becomes larger in the ONIOM calculations, and pathway B ($\Delta E^T = 5.6$ kcal/mol) becomes even more favored than pathway A ($\Delta E^T = 14.3$ kcal/mol). Hence, according to the ONIOM calculations, pathway B is more likely to occur, both thermodynamically and kinetically. The crystal waters may play an important role in controlling the ring-opening paths. In intermediate 3, W654 forms hydrogen bond with the phenolic oxygen of MHPC (Figure 4c). To reach TS-4 of pathway A, this hydrogen bond must be broken (Figure 5a), whereas for TS-6 of pathway B, this hydrogen bond becomes even stronger (Figure 5c). It should be mentioned that, in our previous gas-phase study,22 pathway B was also kinetically more favored than pathway A, even though the energy difference between TS-4 and TS-6 was only 3 kcal/mol. Therefore, despite the fact that the hydrogen-bonding network between crystal water and the substrate may be modified in the real protein (as we did not study the dynamics of the protein), we still believe pathway B to be the dominant one.

3.2.3. Effect of Using Alternative QM/MM Partitioning Schemes. Besides the main ONIOM model calculations (ONIOM-1, 96 QM atoms), we also performed other ONIOM calculations using several different QM/MM partitioning schemes (Table 2), in order to explore if the results are affected by alternative QM/MM partitioning schemes. The activation energies for TS-2, TS-4, and TS-6 in the different QM/MM partitioning schemes are summarized in Table 4. In the calculations, we mainly consider the effect of the backbone of Pro295-Ala296, the side chain of Tyr82, and the crystal water molecules. From Table 4, we can see that the energy barriers are very sensitive to different QM/MM partitioning schemes (compare, e.g., the values for ONIOM-1 and ONIOM-7). According to Table 4, when the backbone of Pro295-Ala296 is added to the QM part, the activation energy for the hydroxylation and ring-opening pathway B increases, while the activation energy for ring-opening pathway A decreases (ONIOM-1 vs ONIOM-3; ONIOM-2 vs ONIOM-4; ONIOM-5 vs ONIOM-7; and ONIOM-6 vs ONIOM-8). Inclusion of the side chain of Tyr82 to the QM part can affect the barrier height for the hydroxylation step when the Pro295-Ala296 backbone is absent in the QM part (ONIOM-3 vs ONIOM-4; ONIOM-7 vs

Figure 5. Geometries from ONIOM-1 calculations of the ring-opening pathways: (a) TS-4, (b) TS-5, (c) TS-6, and (d) TS-9.

Figure 6. Relative energy profiles of $S_1$ (QM, B3LYP, in black) and $S_\Pi$ (ONIOM-EE, B3LYP:Amber in red and M062X:Amber in blue) calculations.
The hydrogen bond between Tyr82 and the substrate changed insignificantly for all pathways investigated, suggesting that Tyr82 is likely to affect the reaction barriers via electrostatic interactions instead of hydrogen bond.

Comparing the S1 model calculations with the SII model, we found that the MM layer has minor effect on the hydroxylation step, but can affect the ring-opening reactions. The activation energies obtained for the hydroxylation step are 32.2 kcal/mol for ONIOM-1(B3LYP/BS2:Amber) model and 31.5 kcal/mol for the QM model (B3LYP/BS2), which are much higher than the experimental value (16.7 kcal/mol). Therefore, additional factors must be present that were not considered in our system, such as the dynamics of the enzyme, and that a single conformation (crystal structure) of the protein may not be sufficient to model the MHPCO reaction. In addition, our calculations show that the ring-opening pathway B, in which an epoxy like transition state is formed, is much more favored than the direct C2–C3 cleavage pathway A, in the presence of the protein environment.

Eight different QM/MM partitioning schemes have been used to study the corresponding reactions, and we conclude that the energy barriers are very sensitive to the QM/MM partitioning scheme, especially for the treatment of the peptide bond in Pro295–Ala296. When this peptide bond is added to the QM region, the barrier heights of the hydroxylation and the ring-opening pathway B increase, while the barrier for the ring-opening pathway A decreases.

## ASSOCIATED CONTENT

S Supporting Information. MM parameters, electronic energies, and Cartesian coordinates are available. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION

Corresponding Author
E-mail: leif.eriksson@nuigalway.ie.

## ACKNOWLEDGMENT

The National University of Ireland—Galway is gratefully acknowledged for financial support.

## REFERENCES

Catalytic Mechanism of Porphobilinogen Synthase: The Chemical Step Revisited by QM/MM Calculations

Bo-Xue Tian,† Edvin Erdtman,‡ and Leif A. Eriksson*,§

† School of Chemistry, National University of Ireland—Galway, Galway, Ireland
‡ School of Engineering, University of Borås, 501 90 Borås, Sweden
§ Department of Chemistry and Molecular Biology, University of Gothenburg, 412 96 Göteborg, Sweden

Supporting Information

ABSTRACT: Porphobilinogen synthase (PBGS) catalyzes the asymmetric condensation and cyclization of two 5-aminolevulinic acid (5-ALA) substrate molecules to give porphobilinogen (PBG). The chemical step of PBGS is herein revisited using QM/MM (ONIOM) calculations. Two different protonation states and several different mechanisms are considered. Previous mechanisms based on DFT-only calculations are shown unlikely to occur. According to these new calculations, the deprotonation step rather than ring closure is rate-limiting. Both the C–C bond formation first mechanism and the C–N bond formation first mechanism are possible, depending on how the A-site ALA binds to the enzyme. We furthermore propose that future work should focus on the substrate binding step rather than the enzymatic mechanism.

1. INTRODUCTION

Tetrapyrrole derivatives, such as porphyrin, chlorine, and corrin, are essential to most life forms.1 Porphobilinogen synthase (PBGS) catalyzes the asymmetric condensation and cyclization of two 5-aminolevulinic acid (5-ALA) substrate molecules to give porphobilinogen (PBG) and is known as the first common step in the biosynthesis of the tetrapyrroles (Scheme 1).2−4 PBGSs from most organisms utilize metal ions (Zn2+, Mg2+, K+, and Na+) as cofactors. The metal ions of PBGS reside at two sites, that is, the active site and the allosteric site.5 Most PBGSs have Zn2+ or Mg2+ in the active site,5−9 albeit some do not use any active site metal ion at all.10 PBGSs that have active site Zn2+ are commonly found in metazoan (including human), archaea, and yeast organisms. We herein focus on the yeast PBGS to illustrate the catalytic process of the enzyme (Scheme 2).

The two substrate molecules bind to the enzyme separately. To distinguish the two ALA molecules, the labels A (acetyl-) and P (propionyl-) are used throughout this study (Scheme 1), referring to the length of the resulting carboxylic acid tail in the final PBG molecule. When the active site lid is open, the P-site substrate binds to the enzyme first and forms a Schiff base with the P-site lysine (Lys263, yeast numbering; Scheme 2).6,7,11 A water molecule is released and the second substrate enters the active site. With the A-site ALA bound, the active site lid becomes closed to prevent further solvent molecules from entering the active site.12 Binding of the A-site ALA is controversial, as the intermediate with both substrates bound is short-lived and has never been observed.12 Crystal structures of

© XXXX American Chemical Society
PBGS-inhibitor complexes\textsuperscript{13−15} have shown that the A-site ALA might also form a Schiff base with the A-site lysine (Lys210). Mutagenesis studies suggest that the P-site lysine is more important than the A-site lysine in terms of decrease in enzymatic activity.\textsuperscript{16} It is argued that the A-site ALA should bind in a geometry that will support the highly exothermic reaction and block access to the bulk solvent, irrespective of whether an A-site Schiff base is formed or not.\textsuperscript{12}

In the current quantum mechanics/molecular mechanics (QM/MM) study, we have considered several starting structures and two different protonation states for the Michaelis complex, aiming to fill the "blank" in Scheme 2. Given the short distance between the nitrogen atom of Lys210 and that of P-site bound ALA (∼4 Å in crystal structure 1H7O,\textsuperscript{6} Scheme 2), we suggest that the A-site ALA may bind directly to the P-site ALA via Schiff base formation to get PS1-S or PS2-S (Scheme 3); such Michaelis complexes were not considered previously (PS2-S is an intermediate in our earlier DFT study\textsuperscript{17}). The Schiff base formation as well as Schiff base exchange reactions are not considered in this study because multiple groups (water molecules, metal ions as well as active site residues) are probably involved,\textsuperscript{18,19} along with conformational changes of the enzyme.\textsuperscript{12}

The role of the active site zinc ion has been discussed in a number of studies,\textsuperscript{5,7,12,18,20−27} but no consensus has been reached. In the crystal structure of the yeast PBGS-PBG* intermediate,\textsuperscript{7} Zn\textsuperscript{2+} binds to three cysteine residues (Cys133, Cys135, and Cys143) as well as the amino nitrogen of the A-site ALA. It is suggested that Zn\textsuperscript{2+} is important for facilitating the binding and reactivity of the A-site ALA, as well as maintaining the stability of the enzyme.\textsuperscript{7,12,25,26} It has also been hypothesized that the active site Zn\textsuperscript{2+} is important for removal of the hydroxyl group during the Schiff base formation and for controlling various proton transfer processes via electrostatic interactions.\textsuperscript{18} Recently, the crystal structure of the yeast PBGS-PBG* intermediate (which has an active site Zn\textsuperscript{2+}) and the crystal structure of Toxoplasma gondii PBGS-PBG complex\textsuperscript{10} (Tg-PBGS, which has no active site metal ion) were resolved, which reveal that the orientation of the product, as well as the positioning of most of the conserved residues, are almost identical (Scheme 4).\textsuperscript{10} We therefore base our models on the assumption that active site metal ion is used for binding the A-site ALA but not for the chemical step.

Another issue is how the pyrrole ring of PBG is formed.\textsuperscript{28,29} Generally speaking, there are two routes to reach the pyrrole ring, either C3−C4 bond formation first or C5−N1 bond formation first (numbering cf. Scheme 1). Recent crystal structures\textsuperscript{6,7,13,15} of PBGS-inhibitor complexes as well as isotope experiments\textsuperscript{18,30} support the C3−C4 bond formation first pathway,\textsuperscript{18} hereafter defined as Mechanism 1 (Scheme 5). In contrast, small model DFT calculations favor the C5−N1 bond formation first pathway,\textsuperscript{17} which is defined as Mechanism 2 (Scheme 6). It should be noted that the protonation state of Mechanism 2 is different from that of Mechanism 1. The favored mechanisms might hence be different for different protonation states.\textsuperscript{31}

In this study, various mechanisms for the chemical step are modeled using the yeast PBGS. Mechanisms 1 and 4, which we show are the most likely ones, are also modeled using the Tg-PBGS structure, to validate the assumption that the active site metal ion is not involved in the chemical step. We wish to address the following questions: (1) whether Schiff base formation of A-site ALA with A-site Lys210 is necessary for catalysis; (2) which protonation state or starting structure is
preferred; (3) which bond (C3−C4 or C5−N1) is formed first in the reaction.

2. COMPUTATIONAL DETAILS

The crystal structures of the yeast PBGS-PBG* intermediate (PDB code 1OHL) and the Toxoplasma gondii PBGS-PBG complex (PDB code 3OBK) were obtained from the RCSB Protein Data Bank. Monomers of the two structures were used. Hydrogen atoms were added using the Molecular Operating Environment (MOE) software (version 2010.10). Protonation states of the active site atoms are summarized in Scheme 4. The complex with charged Lys263 is called Protonation State 1 (protonation state used by Goodwin et al.18) and that with Lys263 neutral is called Protonation State 2 (protonation state used by Erdtman et al.17). In yeast PBGS, the three active site cysteines are set to the deprotonated form, as the Zn−S distances in the crystal structure suggest this pattern (the Zn−S distances in 1OHL are 2.33, 2.33, and 2.30 Å, respectively). If the cysteines are protonated, the Zn−S distance should be ∼2.70 Å. The Zn−S distances (2.25−2.35 Å) obtained from our QM/MM optimizations are consistent with those found in the crystal structure.

The QM region in the QM/MM calculations of the two systems is described in Scheme 4. Starting from the crystal structures of the PBGS-PBG*7 and PBGS-PBG10 complexes, the different Michaelis complexes were traced backward. The results for the forward and reverse reactions are consistent. The systems were initially optimized using the AMBER 03 force field. Missing parameters of PBG were derived using Antechamber35,36 and the partial atomic charges were assigned using the RESP charge method (Supporting Information, S1). QM/MM methods have been extensively used for modeling enzymatic
we have herein used a two-layer ONIOM (QM/MM) scheme with electronic embedding (EE). Geometries were optimized with the ONIOM-EE (B3LYP/AMBER) method. The SDD basis set was used for Zn$^{2+}$ and 6-31G(d) for the other atoms (Table 1), a combination that has been successfully employed in a number of QM/MM studies on zinc-containing enzymes.

**Table 1. Summary of Computational Details**

<table>
<thead>
<tr>
<th></th>
<th>yeast PBGS</th>
<th>Tg-PBGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>protonation state</td>
<td>PS1 and PS2</td>
<td>PS1</td>
</tr>
<tr>
<td>mechanism studied</td>
<td>all mechanisms</td>
<td>Mechanisms 1 and 4</td>
</tr>
<tr>
<td>geometry optimization</td>
<td>B3LYP/BS1/AMBER$^a$</td>
<td>B3LYP/BS3/AMBER$^c$</td>
</tr>
<tr>
<td>single point energy</td>
<td>B3LYP/BS2/AMBER$^b$</td>
<td>B3LYP/BS4/AMBER$^d$</td>
</tr>
</tbody>
</table>

$^a$BS1: SDD for Zn$^{2+}$, 6-31G(d) for other elements. $^b$BS2: SDD for Zn$^{2+}$, 6-311+G(2d,p) for other elements. $^c$BS3: 6-31G(d). $^d$BS4: 6-311+G(2d,p).

3. **RESULT AND DISCUSSION**

**Overview.** For the yeast PBGS system, two different protonation states are used (Scheme 4) and several different mechanisms considered. The most likely mechanisms obtained, that is, Mechanisms 1 and 4, are further modeled in the Tg-PBGS. As a lot of intermediates and transition states (TSs) are involved, we divide each pathway into three stages, that is, (1) deprotonation of A-site ALA, (2) ring closure, and (3) post-PBG formation steps. In the main text, we report relative energies for the key intermediates; detailed energies and geometries can be found in the Supporting Information.

**Mechanism 1.** To date, Mechanism 1 is the most favored one for PBGS.\textsuperscript{6,7,13,15,18} According to this mechanism, both ALA molecules bind to the enzyme in the protonated Schiff base form (PS1-D-1 in Scheme 5). HR at the A-site ALA is abstracted by its own carboxyl group to give the enamine intermediate (PS1-D-2), followed by C3−C4 bond formation. Isotope effect experiments\textsuperscript{18} show that deuterium substitution of HR has a larger isotope effect on $V_{\text{max}}$ than what HS has, but the effect on $V_{\text{max}}/K_M$ is identical for deuterium substitution at both HR and HS, which suggests that deprotonation might not be the rate-limiting step.\textsuperscript{18} PS1-D-3 is unstable and converts to the ring-closed intermediate PS1-D-4.\textsuperscript{15} The steps from PS1-D-4 to the final product are relatively easy. According to our QM/MM calculations, a minor change to the previous mechanism is observed, in that Asp131/Ser179 can abstract a proton from the N1 atom of the P-site ALA and stabilize PS1-D-4, which was not considered previously.

\textsuperscript{a}Reference 17, with small modification.
The energies for the three stages of Mechanism 1 are summarized in Figure 1. Our results indicate that the overall rate-limiting step is the deprotonation stage, which has a barrier of 22.8 kcal/mol. The geometry of PS1-D-TS1-2a (yeast system) is shown in Figure 2. This is somewhat higher than the phenomenological barrier height of ~18 kcal/mol, derived from the experimental turnover rate30 using the Eyring equation. The difference between theoretical results and experimental data might be due to the neglect of entropic contributions and quantum tunneling effects. In addition, deprotonation might occur in the substrate binding process and directly give the deprotonated intermediates. As the geometries are only optimized locally (cf. Computational Details), the transition from the open form to the closed form or other large conformational changes are also not considered in this work.

The ring closure stage can start from two different conformations, PS1-D-2a and PS1-D-2b (Figure 3). The difference between these is whether the Schiff base proton of Lys263 forms a hydrogen bond with the N1 atom (Figure 3 and Scheme 5) or not. This hydrogen bond stabilizes PS1-D-2a and increases the barrier height for the C3–C4 bond formation ($\Delta E^\ddagger = 25.5$ kcal/mol in PS1-D-2a vs $\Delta E^\ddagger = 19.0$ kcal/mol in PS1-D-2b). Interestingly, the conversion PS1-D-2a $\rightarrow$ PS1-D-4 is concerted (Supporting Information, S2). Although intermediate PS1-D-3 can be found in the stepwise pathway from PS1-D-2b (Scheme 5), the conversion PS1-D-3 $\rightarrow$ PS1-D-4 is barrierless (Figure 1), implying that PS1-D-3 is unstable. As the stepwise ring closure from PS1-D-2b ($\Delta E^\ddagger = 19.0$ kcal/mol, Figure 4) is more favorable than the concerted ring closure from PS1-D-2a ($\Delta E^\ddagger = 25.5$ kcal/mol), we suggest that before the C3–C4 bond formation, PS1-D-2a must convert to PS1-D-2b by breaking the hydrogen bond between N1 and the iminium proton of Lys263 through a rotation of the $N-C3-C2-N1$ dihedral angle. The PBG formation stage is fast and we did not observe any step that has a barrier height of more than 10 kcal/mol (Figure 1). The key conversions PS1-D-1 $\rightarrow$ PS1-D-2a and PS1-D-2b $\rightarrow$ PS1-D-3 are also modeled in the Tg-PBGS system, and the results are consistent with those in the yeast PBGS system. This suggests that the active site metal ion is presumably not involved in the chemical step.

Mechanisms 2 and 3. Mechanisms 2 and 3 were previously studied using DFT-only (B3LYP/6-31G(d)//IEFPCM) calculations, and Mechanism 2 was shown to be more favorable than Mechanism 3. However, in our QM/MM calculations on the PS2 system (Scheme 6), the barrier heights for deprotonation are dramatically different from the DFT-only calculations. It should be noted that the protonation state PS2 is used in both Mechanism 2 and Mechanism 3, albeit there is strong experimental evidence that the P-site Schiff base is protonated when the first ALA (P-site ALA) binds (Scheme 2).11
In the previous B3LYP/6-31G(d)//IEFPCM calculations, the barrier heights for conversions PS2-S-1 → PS2-S-3 and PS2-D-1 → PS2-D-2 were 12.1 and 19.4 kcal/mol, respectively. However, in the current ONIOM calculations, the barrier heights for these two steps are considerably higher, 37.5 and 42.0 kcal/mol, suggesting that such conversions are unlikely to occur in PBGS. Such large differences between QM and QM/MM calculations may be due to the easier rotation of the N≡C double bond in the Lys263 imine group (the nitrogen then abstracts the C4 proton HS) in the QM-only calculations, which have more degrees of freedom. In addition, the overall electrostatic effects may also influence the barriers of these two conversions. An alternative pathway is also considered, in which Lys210 acts as a proton shuttle (Scheme 6), but the barrier height is still very high (PS2-S-TS1-2, ΔE‡ = 30.0 kcal/mol). We therefore suggest that the mechanisms in the DFT-only calculations\textsuperscript{17} might be problematic and that neglecting the enzymatic environment is not suitable for modeling these enzymatic reactions. Interestingly, the barriers for the subsequent ring closure stage are actually not high (15.5 and 16.9 kcal/mol, respectively), and if there is some other pathway that gives PS2-S-3 or PS2a-D-2, the ring closure should be straightforward (Tables 2 and 3).

**Mechanism 4.** Mechanism 4 is similar to Mechanism 2, with the main variation that the P-site Schiff base is now

---

**Table 2. Relative Energies of Key Intermediates and Transition States in Mechanism 2**

<table>
<thead>
<tr>
<th>stage</th>
<th>intermediate or TS</th>
<th>relative energy\textsuperscript{a}</th>
</tr>
</thead>
<tbody>
<tr>
<td>deprotonation</td>
<td>PS2-S-1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS1-2</td>
<td>30.0</td>
</tr>
<tr>
<td></td>
<td>PS2-S-2</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS2-3</td>
<td>16.3</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS1-3</td>
<td>37.5</td>
</tr>
<tr>
<td>ring closure</td>
<td>PS2-S-3</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS3-4</td>
<td>23.0</td>
</tr>
<tr>
<td>post PBG formation</td>
<td>PS2-S-4</td>
<td>-8.1</td>
</tr>
<tr>
<td></td>
<td>PS2-S-TS4-5</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td>PS2-S-5</td>
<td>-22.3</td>
</tr>
</tbody>
</table>

\textsuperscript{a}ONIOM(B3LYP/BS2:AMBER)//ONIOM(B3LYP/BS1:AMBER) energies, yeast PBGS system, in kcal/mol.

---

**Table 3. Relative Energies of Key Intermediates and Transition States in Mechanism 3**

<table>
<thead>
<tr>
<th>stage</th>
<th>intermediate or TS</th>
<th>relative energy\textsuperscript{a}</th>
</tr>
</thead>
<tbody>
<tr>
<td>deprotonation</td>
<td>PS2a-D-1</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>PS2a-D-TS1-2</td>
<td>42.0</td>
</tr>
<tr>
<td></td>
<td>PS2a-D-2\textsuperscript{b}</td>
<td>-13.3</td>
</tr>
<tr>
<td></td>
<td>PS2a-D-TS2-3</td>
<td>3.6</td>
</tr>
</tbody>
</table>

\textsuperscript{a}ONIOM(B3LYP/BS2/AMBER)//ONIOM(B3LYP/BS1/AMBER) energies, yeast PBGS system, in kcal/mol. \textsuperscript{b}Conformation similar to PS1-D-2b is used.

---

**Scheme 7. Mechanism 4**

In the previous B3LYP/6-31G(d)//IEFPCM calculations, the barrier heights for conversions PS2-S-1 → PS2-S-3 and PS2-D-1 → PS2-D-2 were 12.1 and 19.4 kcal/mol, respectively. However, in the current ONIOM calculations, the barrier heights for these two steps are considerably higher, 37.5 and 42.0 kcal/mol, suggesting that such conversions are unlikely to occur in PBGS. Such large differences between QM and QM/MM calculations may be due to the easier rotation of the N≡C double bond in the Lys263 imine group (the nitrogen then abstracts the C4 proton HS) in the QM-only calculations, which have more degrees of freedom. In addition, the overall electrostatic effects may also influence the barriers of these two conversions. An alternative pathway is also considered, in which Lys210 acts as a proton shuttle (Scheme 6), but the barrier height is still very high (PS2-S-TS1-2, ΔE‡ = 30.0 kcal/mol). We therefore suggest that the mechanisms in the DFT-only calculations\textsuperscript{17} might be problematic and that neglecting the enzymatic environment is not suitable for modeling these enzymatic reactions. Interestingly, the barriers for the subsequent ring closure stage are actually not high (15.5 and 16.9 kcal/mol, respectively), and if there is some other pathway that gives PS2-S-3 or PS2a-D-2, the ring closure should be straightforward (Tables 2 and 3).

**Mechanism 4.** Mechanism 4 is similar to Mechanism 2, with the main variation that the P-site Schiff base is now...
protonated (Scheme 7). Although deprotonation in Mechanism 4 is still as difficult as in Mechanism 1 (22.4 kcal/mol vs 22.8 kcal/mol), the ring closure becomes much easier (13.0 kcal/mol vs 19.0 kcal/mol; Figure 5). The geometries of PS1-S-TS1 and PS1-S-TS2 are shown in Figure 6. Given only 0.4 kcal/mol difference for the rate-limiting step, we suggest that both Mechanism 1 (C3−C4 bond formation first) and Mechanism 4 (C5−N1 bond formation first) are possible. It should be noted that the K210A mutation causes 1000-fold decrease in PBGS activity; however, this mutation also significantly affects the Schiff base formation of P-site ALA.16 Therefore, mutagenesis is unlikely to tell whether Mechanism 1 or 4 is actually the preferred one. We propose that whether Mechanism 1 or 4 occurs in PBGS is determined by the binding mode of the A-site ALA, which might be different in different PBGS enzymes.

Other Pathways and General Consideration of the PBGS Mechanism. Besides the four mechanisms discussed above, we also modeled other possible pathways, for example, using PS2b-D as starting structure (Schemes 3 and 8). The common feature of all the mechanisms is that the deprotonation of A-site ALA is rate-limiting. Once a deprotonated intermediate is generated, the ring closure is straightforward (ΔE‡ = 15−19 kcal/mol in the various mechanisms). However, in all our attempts to find alternative mechanisms for PBGS, the rate-limiting (deprotonation stage) barriers are always higher than those of Mechanisms 1 and 4 (Table 4). It seems that none of the base groups (Scheme 8) considered in the current study are strong enough to abstract the C4 protons on A-site ALA (Table 4). Even in Mechanisms 1 and 4, the activation energy ~22 kcal/mol is relatively high for an enzymatic reaction. Therefore, we suggest that future work should focus on the binding mode of A-site ALA and the deprotonation of the C4 protons, rather than which bond (C3−C4 or C5−N1) is formed first in the chemical step.

4. CONCLUSIONS

The chemical step of PBGS is studied. Extensive and systematic QM/MM calculations are performed on the yeast PBGS and Tg-PBGS using different Michaelis complexes. We propose that A-site ALA can form a Schiff base with either the A-site lysine or the P-site ALA, and that the Schiff base between A-site ALA and A-site lysine might not be necessary for the ring closure. Our calculations show that Mechanisms 1 and 4 are the most likely ones among all tested. Mechanism 2, which is favored in DFT-only calculations, is shown unlikely to occur. According to our calculations, the preferred protonation state is protonation state 1 in which both Schiff base nitrogens are protonated. The question whether C3−C4 or C5−N1 forms first in the reaction might be system-dependent, given that two competing mechanisms (Mechanisms 1 and 4) exist and that the ring closure stage is presumably not rate-limiting. The rate-limiting barrier of ~22 kcal/mol for Mechanisms 1 and 4 is higher than that obtained from experimental data ~18 kcal/mol, and possible explanations are that neither entropic nor quantum tunneling effects are considered in this work, and that deprotonation of the A-site ALA might occur already in the substrate binding process where large conformational changes are involved. Unfortunately, the current methodology is not able to evaluate such changes. MD simulations may be employed to study the conformational changes induced by the binding of A-site ALA, but are obviously nontrivial. To better understand the PBGS mechanism, we propose that future focus should be on the binding process of A-site ALA and the deprotonation of the C4 protons.

ASSOCIATED CONTENT

Supporting Information
S1: AMBER parameters for PBG; S2: 3D relative energy surface near PS1-D-TS2a-4; S3: Complete citation for ref 50; S4: Cartesian coordinates (QM region) and QM/MM energies
of key intermediates and TSs. This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

**Corresponding Author**
*E-mail: leif.eriksson@chem.gu.se.*

**Notes**
The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**
The National University of Ireland—Galway and the Faculty of Science of the University of Gothenburg are gratefully acknowledged for financial support. The SFI/HEA Irish Centre for High-End Computing (ICHEC) is acknowledged for the provision of computational facilities and support.

**REFERENCES**

(3) Jordan, P. M. *Biosynthesis of Tetrapyroles*; Elsevier: Amsterdam, 1991; Vol. 19.
Paper 6
Exploring Enzyme Substrate Specificity with Docking and QM/MM Methods: A Benchmarking Study on the Stereo-specificity of D-glucarate Dehydratase

Bo-Xue Tian†, Frank Wallrapp‡, Chakrapani Kalyanaraman†, Suwen Zhao§, Leif A. Eriksson*§, Matthew P. Jacobson*‡

†School of Chemistry, National University of Ireland – Galway, Galway, Ireland
‡Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94158, USA
§Department of chemistry and molecular biology, University of Gothenburg, 412 96 Göteborg, Sweden

Keywords: Enzyme specificity, stereochemistry, enzyme function prediction, QM/MM, induced fit docking, MM/GBSA.

Supporting Information

ABSTRACT: The stereo-specificity of D-glucarate dehydratase (GlucD) is explored by combined docking and QM/MM calculations. According to our calculations, both the substrate binding and the chemical steps of GlucD contribute to substrate specificity. The current approach will be used for assisting enzyme function assignment.

Enzymatic reactions with high stereo-specificity usually cannot be achieved by synthetic approaches. Understanding enzyme specificity at the atomic level is important for enzyme function prediction and redesign of enzymes. An enzymatic reaction consists of three common steps, i.e. substrate binding, chemical reaction and product release. The first two steps are important for substrate specificity, and $k_{cat}/K_M$ is often used to experimentally determine the best substrate for an enzyme. Nowadays, modeling enzyme-ligand systems with computer simulations has evolved to a routine methodology. To evaluate $k_{cat}/K_M$ with computational approaches, one needs to model both substrate binding and chemical steps. However, most computational studies only focus on one of these aspects.

In most cases, enzyme (in vitro) function prediction problem is equivalent to studying enzyme substrate specificity, where the key is to find a ligand with high $k_{cat}/K_M$ values. Therefore, combining properties for the substrate binding step (e.g. binding affinity) and for the chemical step (e.g. activation energy) should give more reasonable results in determining whether or not a ligand is a substrate. Docking-based approaches have been successfully used in enzyme function prediction. However, an obvious limitation of docking-based approaches is that no information about the chemical step can be obtained.

We herein report a benchmarking study on the stereo-specificity of D-glucarate dehydratase (GlucD) based on combined docking and quantum mechanical/molecular mechanical (QM/MM) methods. GlucD catalyzes the dehydration of D-glucarate or L-idarate to give 5-keto-4-deoxy-D-glucarate (5-KDG), as well as the interconversion between D-glucarate and L-idarate (Scheme 1). In the dehydration of D-glucarate, an active site histidine (H339) abstracts a proton from C5 to give the enediolate intermediate I1. The protonated H339 then transfers the proton to the hydroxyl on C4 to give the enol intermediate I2, which subsequently converts to the final product 5-KDG (Scheme 1). Dehydration of L-idarate is initialized by K207, and then undergoes the same intermediates I1 and I2.

Scheme 1. Catalytic mechanism of GlucD

GlucD is highly stereo-specific in the non-reactive regions, and as a result, none of m-allarate, D-mannarate or D-altrarate is a substrate of GlucD (Scheme 2). In the current work, docking-
based methods were used for studying the GlucD specificity at the binding step. Computational details can be found in SI. Protonation states of H339 and K207 were assigned according to the Michaelis complex for D-glucarate (Scheme 1), as the reactive regions of the four ligands in Scheme 2 are identical. Glide-SP, Glide-XP and induced fit docking (IFD) methods were used to get the binding poses of the four ligands in Scheme 2, and the MM/GBSA method to evaluate binding affinities (Table 1). All three docking methods are able to reproduce the binding pose of D-glucarate as seen in the crystal structure. D-glucarate is the best binder in the IFD/MM/GBSA calculations and is the second best binder in the Glide-SP/MM/GBSA and the Glide-XP/MM/GBSA calculations (Table 1). As the natural substrate D-glucarate is expected to be the best binder, this suggests that the binding affinity at the IFD/MM/GBSA level is appropriate for qualitatively distinguishing substrate and non-substrate for GlucD. It should be mentioned that in Table 1, MM/GBSA calculations are performed on the reactive poses of each ligand, i.e. the pose whose reactive region is similar to the ligand in the crystal structure of GlucD. We choose reactive poses because only these poses can undergo the chemical steps, which will be discussed later in the text. The reactive poses are obtained from the top 10 docking poses, and in most cases, the reactive pose actually has the highest MM/GBSA affinity, details see Table S2.

Scheme 2. Structures of D-glucarate and three non-substrates

Table 1. Relative binding affinities of the reactive poses of D-glucarate and the three non-substrates

<table>
<thead>
<tr>
<th>Ligands</th>
<th>Glide-SP</th>
<th>Glide-XP</th>
<th>IFD</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-glucarate</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>m-allarate</td>
<td>-9.3</td>
<td>11.4</td>
<td>47.8</td>
</tr>
<tr>
<td>D-mannarate</td>
<td>9.5</td>
<td>34.0</td>
<td>24.2</td>
</tr>
<tr>
<td>D-altrarate</td>
<td>8.0</td>
<td>-21.0</td>
<td>18.0</td>
</tr>
</tbody>
</table>

*binding affinities are calculated at the MM/GBSA level (see SI for details), in kcal/mol

The QM/MM [ONIOM (B3LYP:AMBER)] method was used for modeling the chemical step of GlucD, with the QM layer as described in Scheme 3. All the QM/MM energies reported here are relative energies at the [B3LYP/6-311+G(2d,p):AMBER] level. Although there is experimental evidence to support the catalytic mechanism of GlucD (as described in Scheme 1), no computational study on this mechanism has yet been reported. According to QM/MM calculations on two related enzymes, i.e. mandelate racemase (MR) and enolase (in the following text, EL means the enzyme enolase, and EN stands for the enolase superfamily), the catalytic mechanisms of the EN superfamily are not always the same, depending on binding modes of the substrates and the protonation states of the active site residues. MR, EL and GlucD all belong to the EN superfamily, and GlucD shares more similarities with EL in terms of substrate binding mode, in which the carboxylate group of the substrate is a bidentate ligand to the Mg\(^{2+}\) (Scheme 3).

Computational details can be found in SI. According to our calculations on D-glucarate substrate, TS2 is the rate-determining state (Figure 1), and the obtained energy barrier 15.1 kcal/mol is consistent with the phenomenological activation energy \(\sim15\) kcal/mol (At 295.15K, \(k_a\) value is 35 s\(^{-1}\) for D-glucarate\(^{1a}\)). The barrier height for the proton abstraction (PA) step is also significant (TS1, 12.9 kcal/mol, Figure 1). Although no experimental results are available to verify what the rate-limiting step is, it has been suggested that the PA step is at least partially rate-limiting\(^{7,8}\), because all the enzymes in the EN superfamily share this common step.

Scheme 3. QM region (~100 atoms) in the QM/MM calculations

Figure 1. Energy profile (in kcal/mol) along the reaction coordinate for D-glucarate

We thus focus on both transition states (TS1 and TS2) in order to study the specificity at the chemical step. D-glucarate is the best substrate in terms of energies for both TSs (Table 2). In TS1 and TS2, the energy difference between D-glucarate and the three non-substrates is in the range of 2-7 kcal/mol (Table 2), which is large enough for us to conclude that specificity is encoded in the chemical step. Comparison of the energy profiles and the optimized geometries of the four ligands can be found in Figure S4-S6. Besides the energies of the TSs, we notice that the atomic charges derived from electrostatic potentials (ESP) relate to the specificity at the PA step. In the Michaelis complexes of the four ligands, the charge difference between the proton on C5 and the nitrogen acceptor on H339 (qH-qN, c.f. Scheme 3) is in qualitatively accordance with the energies of...
TS1 (Table 2). As the optimized geometries at the reactive region are very similar for the four ligands, we suggest that the energies of TS1 are mainly affected by the variations of the overall electrostatic environment. As optimizations of the Michaelis complexes are computationally more efficient than optimizations of both Michaelis complex and the transition state, the ESP atomic charges of the Michaelis complex may thus be used for excluding poor substrate candidates when the computational resources are limited.

Table 2. Relative Energies of TS1 and TS2 for D-glucurate and three non-substrates

| Ligands          | $\Delta E_{TS1}$ | $\Delta\Delta E_{TS1}$ | $\Delta E_{TS2}$ | $\Delta\Delta E_{TS2}$ | $q_H$/|$q_N$ |
|------------------|-------------------|-------------------------|------------------|-------------------------|------|
| D-glucarate      | 12.9              | 0.0                     | 15.1             | 0.0                     | 1.10 |
| m-allarate       | 16.1              | 3.2                     | 17.3             | 2.2                     | 0.64 |
| D-mannarate      | 15.5              | 2.6                     | 18.3             | 3.2                     | 0.98 |
| D-altarate       | 19.3              | 6.4                     | 20.2             | 5.1                     | 0.33 |

$^a$B3LYP/6-311+G(2d,p); $^b$AMBER energies, in kcal/mol

In conclusion, docking and QM/MM calculations are performed to study the GlucD stereo-specificity at the non-reactive regions. Our results show that GlucD specificity is encoded in both the substrate binding and the chemical steps. We emphasize that enzyme substrate specificity is a two parameter ($k_a$/$k_b$) problem, and computational studies should investigate properties for both steps (possibly using mathematics to integrate these properties in the future). The method presented here may be used to assist in enzyme function prediction.

ASSOCIATED CONTENT

Supporting Information. Computational details; docking results; ESP charges; Images for intermediates and TSs; Cartesian coordinates and energies. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author
matt.jacobson@ucsf.edu

Funding Sources
Any funds used to support the research of the manuscript should be placed here (per journal style).

Notes
Any additional relevant notes should be placed here.

ACKNOWLEDGMENT

The National University of Ireland–Galway is gratefully acknowledged for financial support. The SFI/HEA Ireland Centre for High-End Computing (ICHEC) is acknowledged for the provision of computational facilities and support.

ABBREVIATIONS

GlucD, D-glucarate dehydratase; 5-KDG, 5-keto-4-deoxy-D-glucarate; SP, Standard Precision; XP, Extra Precision; IFD, induced fit docking; PA, proton abstraction; QM/MM, quantum mechanical/molecular mechanical; ESP, electrostatic potential.

REFERENCES

Paper 7
Catalytic Mechanism and Product Specificity of Oxidosqualene-lanosterol Cyclase: A QM/MM study

Bo-Xue Tian\textsuperscript{1} and Leif A. Eriksson\textsuperscript{2}\textsuperscript{*}

\textsuperscript{1}School of Chemistry, National University of Ireland – Galway, Galway, Ireland.
\textsuperscript{2}Department of Chemistry and Molecular Biology, University of Gothenburg, 412 96 Göteborg, Sweden

Abstract

Oxidosqualene-lanosterol cyclase (OSC) is a key enzyme in the biosynthesis of cholesterol. The catalytic mechanism and the product specificity of OSC have herein been studied using QM/MM calculations. According to our calculations, the protonation of the epoxide ring of oxidosqualene is rate-limiting. Wild type OSC (which generates lanosterol), and the mutants H232S (which generates parkeol) and H232T (which generates protosta-12,24-dien-3-beta-ol) were modeled, in order to explain the product specificity thereof. We show that the product specificity of OSC at the hydride/methyl-shifting stage is unlikely to be achieved by the stabilization of the cationic intermediates, as the precursor of lanosterol is in fact not the most stable cationic intermediate for wild type OSC. The energy barriers for the product-determining conversions are instead found to be related to the product specificity of different OSC mutants, and we thus suggest that the product specificity of OSC is likely to be controlled by kinetics, rather than thermodynamics.
**Keywords:** Oxidosqualene-lanosterol Cyclase, catalytic mechanism, product specificity, QM/MM, kinetic control

*corresponding author: leif.eriksson@chem.gu.se

**Introduction**

Oxidosqualene-lanosterol cyclase (OSC), which catalyzes the conversion from oxidosqualene to lanosterol, is a key enzyme in the biosynthesis of cholesterol\(^1\). OSC inhibitors are drug candidates against hypercholesterolemia\(^2-4\). The proposed catalytic mechanism of OSC is shown in Scheme 1\(^3\). Protonation of the epoxide ring of oxidosqualene by Asp455 triggers ring-forming reactions, yielding the protosterol cation 4. A series of 1,2-hydride and 1,2-methyl group shifts then occurs, followed by a final deprotonation step leads to the product lanosterol (Scheme 1)\(^3\). In the crystal structure of the OSC-lanosterol complex, His232 and Tyr503 (human OSC numbering) are in favorable positions for the final deprotonation step (Scheme 2)\(^3\). Mutation of the active site histidine His232 to different amino acids leads to diverse products, some of which are shown in Scheme 2\(^5\).

Although multiple cationic/olefin intermediates are formed during the catalytic process of wild type OSC, the sole product lanosterol is obtained. We divide the OSC specificity into two stages, i.e. before or after the formation of protosterol cation 4. In the first stage, the 4-ring chair-boat-chair scaffold protosterol cation is the dominant intermediate formed in wild type OSC, whilst various byproducts from 1-ring to 3-ring are observed in OSC mutants\(^5-14\). In the second stage, although the hydride and methyl shifts may results
in many different 4-ring products (e.g. P1, P2 and P3; Scheme 2), only lanosterol is formed in wild type OSC.

**Scheme 1. Proposed mechanism of OSC**

To date, no computational study on the overall mechanism of OSC has been reported, except that the first step (conversion \( \text{R}\rightarrow\text{1} \)) was studied by QM calculations on small models\(^{15} \). The mechanism of squalene-hopene cyclase (SHC), a closely related enzyme, has been investigated in a number of computational studies\(^{16-23} \), where the ring formation steps have been discussed in great detail. In the current work, we are interested in the
product specificity at the hydride/methyl-shifting stage, which was never studied previously. In particular, two representative mutants were considered, i.e. H232S and H232T, which generate sole product Parkeol (P2) and protosta-12,24-dien-3-beta-ol (P3), respectively (Scheme 2)\textsuperscript{5}.

**Scheme 2.** OSC mutants and their corresponding products
Computational details

DFT calculations

The DFT functionals B3LYP\textsuperscript{24-26} and M06-2X\textsuperscript{27} were used to study the energies of the key intermediates and products. Geometries were optimized \textit{in vacuo} using B3LYP with the 6-31G(d) basis set (BS1), followed by frequency calculations at the same level of theory. Single-point energy calculations were performed on the optimized structures using the higher basis set 6-311+G(2d,p) (BS2). All DFT calculations were performed using the Gaussian 09 program\textsuperscript{28}.

ONIOM calculations

The crystal structure of the OSC-lanosterol complex (PDB code 1W6K\textsuperscript{3}) was obtained from the RCSB Protein Data Bank. Hydrogen atoms were added using the Molecular Operating Environment (MOE) software (Version 2010.10)\textsuperscript{29}. Lanosterol, Asp455, Tyr503 and His232 were included in the QM region in the QM/MM calculations. In the OSC-lanosterol complex, Asp455 was set to the deprotonated form, and Tyr503 was set to the protonated form (in the corresponding reactant complex, Asp255 is protonated and Tyr503 is deprotonated, cf. Scheme 1). Three different protonation states for His232 were considered in the wild type OSC, i.e. HID, HIE and HIP (the HID model is shown in Scheme 2, which is favored in the MOE protonation 3D prediction, the other models are shown in Supporting Information S1). It should be noted that HID232 might be the base for the deprotonation step in wild type OSC, but in the two mutants in Scheme 2, Tyr503 is the only base suitable for deprotonation. As the deprotonation by Tyr503 is easier than that by HID232 (data not shown), we only discuss the Tyr503 residue as the base for
deprotonation. Intermediates and TSs in the proposed mechanism (Scheme 1) from the product complex (OSC-lanosterol) to the reactant complex (OSC-oxidosqualene) were optimized using wild type OSC (HID form). Key intermediates 6-9 were then optimized using the H232S and H232T mutants, as well as HIE and HIP forms of wild type OSC. The mutants were built using a similar approach as previously reported\textsuperscript{30}. The ONIOM method\textsuperscript{31} was used for all QM/MM optimizations. The systems were initially optimized using the AMBER 03 force field\textsuperscript{32}. Missing parameters of lanosterol were derived using Antechamber\textsuperscript{33,34} with the RESP charge method (Supporting Information S2). A two-layer ONIOM (QM/MM) scheme with electronic embedding (EE) was used. Geometries were optimized at the ONIOM-EE (B3LYP/BS1:AMBER) level. Atoms far away from the lanosterol atoms (>10 Å) were held fixed during the QM/MM optimizations. Frequency calculations were performed on the QM atoms only (using the ‘geom.= readfreeze’ option in G09) to confirm the correct normal modes of the imaginary frequency in those TS’s. Full frequency calculations were not performed due to shortage of memory. Single-point energy calculations were performed on the optimized geometries with ONIOM-EE (B3LYP/BS2:AMBER). All ONIOM calculations were performed using the Gaussian 09 software\textsuperscript{28}.

**Result and Discussion**

*DFT calculations on key intermediates and products*

The two DFT methods B3LYP and M06-2X were used to analyze the gas phase energies of the cation intermediates in Scheme 1 and Scheme 2, and the results are summarized in
Table 1. Recently, Kürti et al.\textsuperscript{16} studied the energies of various cations involved in the OSC catalysis using quantum mechanics calculations. They suggest that OSC utilizes a nonstop process, and that product specificity of OSC is probably achieved by the stabilization of the cationic intermediates\textsuperscript{16}. According to their hypothesis, to selectively produce lanosterol, intermediate 8 must have lower energy than 6 and 9 (Scheme 2). However, our DFT calculations disprove this, in that intermediate 8 was is less stable than 6 or 9 in terms of gas phase energy (Table 1). The gas phase energies of the three products in Scheme 2 were also compared (Table 2). According to our calculations, the energies of P1 and P3 are close, while P2 seems to be thermodynamically much less stable. Taken together, the gas phase calculations on the key intermediates and products do not favor P1 as the sole product. To investigate the product specificity of OSC, it thus appears that one must take the enzymatic environment into account, rather than simply look at the gas phase energies of the isolated intermediates.

Table 1. Gas phase energies of various cation intermediates

<table>
<thead>
<tr>
<th>Cationic intermediate\textsuperscript{a}</th>
<th>$E_{\text{B3LYP}}$\textsuperscript{b}</th>
<th>$E_{\text{M06-2X}}$\textsuperscript{c}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>-9.1</td>
<td>-22.3</td>
</tr>
<tr>
<td>4</td>
<td>-8.5</td>
<td>-36.3</td>
</tr>
<tr>
<td>5</td>
<td>-14.0</td>
<td>-41.3</td>
</tr>
<tr>
<td>6</td>
<td>-16.8</td>
<td>-44.3</td>
</tr>
<tr>
<td>7</td>
<td>-12.8</td>
<td>-40.6</td>
</tr>
<tr>
<td>8</td>
<td>-15.3</td>
<td>-43.5</td>
</tr>
<tr>
<td>9</td>
<td>-17.2</td>
<td>-46.0</td>
</tr>
</tbody>
</table>
aThe structures of intermediates 1-9 are described in Scheme 1 and Scheme 2

brelative energies at the B3LYP/BS2//B3LYP/BS1 level

crelative energies at the M06-2X/BS2//B3LYP/BS1 level

**Table 2.** Gas phase energies of the three major products

<table>
<thead>
<tr>
<th>Product</th>
<th>$E_{B3LYP}^a$</th>
<th>$E_{M06-2X}^c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>P2</td>
<td>11.8</td>
<td>11.6</td>
</tr>
<tr>
<td>P3</td>
<td>0.5</td>
<td>-1.0</td>
</tr>
</tbody>
</table>

a product structures are described in Scheme 2

b relative energies at the B3LYP/BS2//B3LYP/BS1 level

c relative energies at the M06-2X/BS2//B3LYP/BS1 level

**ONIOM calculations on wild type OSC**

The proposed mechanism of OSC (Scheme 1) was subsequently studied using QM/MM (ONIOM) methodology. The QM/MM method has been employed in a number of mechanistic studies on enzymes. In a recent QM/MM study by Rajamani et al., the catalytic mechanism of SHC, a related enzyme, was systematically analyzed. The two major findings were 1) the proposed ring expansion process from initial 5-membered to 6-membered ring is unlikely for either C- or D-ring formation, instead 6-membered ring is formed directly; 2) the distribution of two key products, the 6,6,6,5-tetracyclic carbon skeleton and the 6,6,6,6,5-pentacyclic hopanoids, is controlled by kinetics. In the current study on OSC, we are interested in the product specificity at the hydride/methyl-shifting stage after the formation of the 6,6,6,5-tetracyclic intermediate 4. Intermediates
and the TSs linking these were thus optimized by using ONIOM calculations (Scheme 1). As mentioned above, His232 was set to the HID protonation isomer.

**Figure 1.** Energy profile of the proposed mechanism of the wild type OSC (H232 is in the HID form).

The first step is the protonation of \( \mathbf{R} \), which is concerted with A ring formation, to give \( \mathbf{1} \). As this step has the highest barrier in the catalytic process (23.0 kcal/mol), we suggest that it is the rate-limiting step of OSC. It should be noted that the obtained barrier height 23.0 kcal/mol is much lower than a previous small model QM study\(^{15}\) (40 kcal/mol with B3LYP/BS1). The optimized geometry of **TSR-1** is shown in Figure 2a. The B ring formation is relatively easy, whereby intermediate \( \mathbf{2} \) is obtained. The 6,6,6-tricyclic intermediate \( \mathbf{3} \) was found to be unstable, which is consistent with the previous QM/MM study on SHC\(^{21}\) as well as the fact that 6,6,6-tricyclic products are never observed. The
optimized geometry of **TS2-4** is shown in Figure 2b. It should be mentioned that in SHC, the C-ring formation may lead to 6,6,5-tricyclic, 6,6,5,6-tetracyclic and 6,6,6,5-pentacyclic intermediates. In the current study on OSC, 6,6,5-tricyclic and 6,6,6,6,5-pentacyclic intermediates were not considered, as our focus is on the product specificity at the hydride/methyl-shifting stage. The conversion of 4 to 8 is fast and exothermic (Figure 1), whereafter 8 converts to lanosterol.
Figure 2. Optimized geometries of a) TSR-1; b) TS2-4

ONIOM calculations on OSC mutants

Intermediate 8 is the most stable one of those shown in Scheme 1 (Figure 1), which seems to support the hypothesis that product specificity of OSC is achieved by the stabilization of the cationic intermediates $^5,^{11-14,16}$. However, our QM and QM/MM calculations show that intermediate 9, which is the precursor of Parkeol (P2), is more stable than 8 (the precursor of P1). The relative energies of key intermediates 6, 8 and 9 in the wild type OSC and those in the H232S and H232T mutants were thus compared (Table 3, columns 2, 5 and 6). The trends of the relative energies of these intermediates are the same. The relative energies of the three products in different mutants were also studied, and the results show that P2 is more stable than P1 (Table 3). Hence, according
to our calculations, \( P_2 \) is thermodynamically more favorable than \( P_1 \) for all the systems investigated in this study.

**Table 3.** Relative energies of the key intermediates and products in the QM/MM calculations

<table>
<thead>
<tr>
<th>Intermediates or Products</th>
<th>Relative Energy(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wild type HID</td>
</tr>
<tr>
<td>6</td>
<td>24.5</td>
</tr>
<tr>
<td>8</td>
<td>16.0</td>
</tr>
<tr>
<td>9</td>
<td>12.1</td>
</tr>
<tr>
<td>( P_1 )</td>
<td>0.0</td>
</tr>
<tr>
<td>( P_2 )</td>
<td>-6.6</td>
</tr>
<tr>
<td>( P_3 )</td>
<td>1.5</td>
</tr>
</tbody>
</table>

\(^a\) ONIOM-EE(B3LYP/BS2:AMBER) energies, relative to \( P_1 \), in kcal/mol

In the previous QM/MM study by Rajamani et al.\(^{21}\), the product specificity of SHC was found to be controlled by kinetics. We thus investigate whether this holds true also for OSC. In the current work, only the barrier heights for the conversions in the last few steps were studied, in order to explain why \( P_1 \), \( P_2 \) and \( P_3 \) are selectively generated in wild type OSC, H232S and H232T mutants, respectively (Table 4 and Figure 3). In the wild type OSC and the H232S mutant, the formation of 7 is faster than \( P_3 \) (Table 4 and Figure 3), and thus \( P_3 \) is not generated. In the H232T mutant, however, the barrier for formation of \( P_3 \) is lower than that leading to formation of 7 (Table 4 and Figure 3). This hence explains why \( P_3 \) is the product in the H232T mutant. Similarly, in wild type OSC, as the
formation of P1 from 8 is faster than the formation of 9 (Table 4 and Figure 3), P1 is the preferred product, although P2 is thermodynamically more stable.

For the H232S mutant, finally, the reaction from 8 to 9 is associated with a lower barrier then 8 to P1, which thus in this case allows the system to proceed towards the observed product P2.

**Figure 3.** Energy profiles of wild type OSC (in black), H232S (in red) and H232T (in green) mutants

**Table 4.** Barrier heights for the key conversions that determine product specificity

<table>
<thead>
<tr>
<th>conversion</th>
<th>Barrier heights$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wild type HID</td>
</tr>
<tr>
<td>6→7</td>
<td>6.4</td>
</tr>
<tr>
<td>6→P3</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---</td>
</tr>
<tr>
<td>8→9</td>
<td>5.8</td>
</tr>
<tr>
<td>8→P1</td>
<td>0.8</td>
</tr>
<tr>
<td>9→P2</td>
<td>2.0</td>
</tr>
</tbody>
</table>

aONIOM-EE(B3LYP/BS2:AMBER) energies, in kcal/mol

Conclusions

The catalytic mechanism and the product specificity of OSC were investigated using QM/MM calculations. The first step, i.e. protonation of the epoxide ring of oxidosqualene, was found to be rate-limiting. According to our DFT and ONIOM calculations, intermediate 9, which is the precursor of parkeol, is more stable than 8 (the precursor of lanosterol). Therefore, we suggest that the relative stability of the cationic intermediates may to some extent contribute to the OSC product specificity, but is not the main factor that governing this. The barrier heights at the hydride/methyl-shifting stage were found to be consistent with the product profiles of different OSC mutants. Hence, we suggest that the product specificity of OSC is likely to be controlled by kinetics. Mutagenesis results in changes of the electrostatic environment of the enzyme, and thus affects the reaction barriers at the hydride/methyl-shifting stage. When the base group deprotonates the carbon cation intermediate, the product is generated irreversibly. Our calculations are consistent with all the experimental data available to date.

Acknowledgements
The National University of Ireland-Galway and the Faculty of Science of the University of Gothenburg are gratefully acknowledged for financial support. The SFI/HEA Irish Centre for High-End Computing (ICHEC) is acknowledged for the provision of computational facilities and support.

Supporting Information

Illustration of HID, HIE and HIP models; AMBER parameters for lanosterol; Cartesian coordinates (QM region) and QM/MM energies of key intermediates and TS; Complete citation for Ref 28. Supporting Information is available free of charge via the Internet at http://pubs.acs.org.

References
