TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM School of Life Sciences

Application of QM/MM methods to Protein/Ligand binding

Chen Zheng

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Dmitrij Frischmann

Prüfer der Dissertation:

1. Prof. Dr. Aphrodite Kapurniotu
2. Prof. Dr. Martin Zacharias

Die Dissertation wurde am 29.11.2021 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 30.03.2022 angenommen.

Abstract

Serine proteases are one type of enzymes that are able to cleave peptide bonds in proteins, in which serine serves as a nucleophile at the active site. They are widespread in eukaryotes, bacteria, archaea, and viruses. Because of the important physiological role of serine proteases, they are popular targets for the antivirulence treatment of bacterial or viral infections. A complete catalytic cycle of serine protease includes two separate steps: acylation and deacylation. According to the generally accepted reaction mechanism, a tetrahedral intermediate is formed in both acylation and deacylation reactions. Understanding the reaction mechanism of serine proteases with inhibitors and main factors which influence the kinetics of reactions at a molecular level can help us to design novel drugs. In this dissertation, QM/MM calculations were performed to study the reaction mechanism and to calculate (free) energy barriers for two model systems of serine proteases, which are *Staphylococcus aureus* caseinolytic protease P (SaClpP) and hepatitis C virus (HCV) NS3/4A protease, respectively.

In the first part of this dissertation, both acylation and deacylation reactions of SaClpP with three inhibitors and one fluorescent substrate were simulated. In the beginning, several semi-empirical QM methods were tested to simulate the nucleophilic attack step of the acylation reaction. No single semi-empirical QM method can properly describe all model systems. Then QM/MM calculations were conducted by DFT methods. As automatic proton hopping to the ligand was observed for β-lactones, a second reaction coordinate which controls the progress of proton transfer from the histidine residue in the catalytic triad to the ligand was required. Thus, two-dimensional potential energy surfaces were calculated for all systems. Our calculation results at the B3LYP/def2-TZVP level agree with experimental values very well. Furthermore, the results reveal that there is no tetrahedral intermediate and the acylation reaction follows a one-step reaction mechanism for two β-lactones and phenyl ester ML90. For the modelling of the deacylation step, only DFT methods were used. Simulation of the deacylation step is more difficult than the acylation step since one water molecule is involved in the reaction. Our calculation results reveal that water dynamics play a pivotal role in successfully modelling the deacylation reaction. By iteratively improving to the calculation setup, our best protocol, which combines QM/MM free energy perturbation calculations and including more water molecules into the QM region, basically reproduced experimental values. In addition, the ability of the leaving group seems the major factor which determines the rate constants of ClpP for both acylation and deacylation reactions except for the systems with ring-opening.

The second part of this dissertation focused on QM/MM simulations of acylation reactions for HCV NS3/4A protease variants bound to mitochondrial antiviral-signaling protein (MAVS) and the NS4A/B substrate. Recent experimental findings pointed out a new mutation in HCV protease, Q41R, responsible for a significant enhancement of the enzyme's reactivity towards MAVS. The Q41R mutation is located rather far from the active site, and its involvement in the overall reaction mechanism is thus unclear. Classical molecular dynamics and QM/MM were used to study the acylation reaction of HCV NS3/4A protease variants bound to MAVS and the NS4A/4B substrate and uncovered the indirect mechanism by which the Q41R mutation plays a critical role in the efficient cleavage of the substrate. Our simulations reveal that there are two major conformations of the MAVS H1'(p) residue for the wild type protease and only one conformation for the Q41R mutant. The conformational space of H1'(p) is restricted by the Q41R mutation due to a π-π stacking between H1'(p) and R41 as well as a strong hydrogen bond between the backbone of H57 and the side chain of R41. Further QM/MM calculations indicate that the complex with the conformation ruled out by the Q41R substitution is a non-reactive species due to its higher free energy barrier for the acylation reaction. Based on our calculations, we propose a kinetic mechanism that explains experimental data showing an increase of apparent rate constants for MAVS cleavage in Q41R mutants. Our model predicts that the non-reactive conformation of the enzyme-substrate complex modulates reaction kinetics like an uncompetitive inhibitor.

Zusammenfassung

Serinproteasen sind Enzyme, die Peptidbindungen in Proteinen spalten können, wobei Serin als Nukleophil am aktiven Zentrum dient. Sie sind in Eukaryoten, Bakterien, Archaeen und Viren weit verbreitet. Aufgrund der wichtigen physiologischen Rolle der Serinproteasen sind sie beliebte Ziele für die Therapie von bakteriellen oder viralen Infektionen. Ein vollständiger Katalysezyklus der Serinprotease umfasst zwei separate Schritte: Acylierung und Deacylierung. Nach dem allgemein anerkannten Reaktionsmechanismus wird sowohl bei Acylierungs- als auch bei Deacylierungsreaktionen eine tetraedrische Zwischenstufe gebildet. Das Verständnis des Reaktionsmechanismus von Serinproteasen mit Inhibitoren und Hauptfaktoren, die die Kinetik der Reaktionen auf molekularer Ebene beeinflussen, kann uns bei der Entwicklung neuer Medikamente helfen. In dieser Dissertation wurden QM/MM-Rechnungen durchgeführt, um den Reaktionsmechanismus zu untersuchen und (freie) Energiebarrieren für zwei Modellsysteme von Serinproteasen zu berechnen, nämlich die caseinolytische Protease P von *Staphylococcus aureus* (SaClpP) und die NS3/4A-Protease des Hepatitis-C-Virus (HCV).

Im ersten Teil dieser Dissertation wurden sowohl Acylierungs- als auch Deacylierungsreaktionen von SaClpP mit drei Inhibitoren und einem fluoreszierenden Substrat simuliert. Zu Beginn wurden mehrere semi-empirische QM-Methoden getestet, um den nukleophilen Angriffsschritt der Acylierungsreaktion zu simulieren. Keine einzelne semi-empirische QM-Methode kann alle Modellsysteme richtig beschreiben. Dann wurden QM/MM-Berechnungen mit DFT-Methoden durchgeführt. Da für β-Lactone ein automatisches Protonen-Hopping zum Liganden beobachtet wurde, war eine zweite Reaktionskoordinate erforderlich, die den Fortschritt des Protonentransfers vom Histidinrest in der katalytischen Triade zum Liganden steuert. Daher wurden für alle Systeme zweidimensionale Potentialhyperflächen berechnet. Unsere Berechnungsergebnisse auf dem B3LYP/def2-TZVP-Niveau stimmen sehr gut mit experimentellen Werten überein. Darüber hinaus zeigen die Ergebnisse, dass es keine tetraedrische Zwischenstufe gibt und die Acylierungsreaktion einem einstufigen Reaktionsmechanismus für zwei β-Lactone und Phenylester ML90 folgt. Für die Modellierung des Deacylierungsschritts wurden ausschließlich DFT-Methoden verwendet. Die Simulation des Deacylierungsschritts ist schwieriger als der Acylierungsschritt, da ein Wassermolekül an der Reaktion beteiligt ist. Unsere Berechnungsergebnisse zeigen, dass die Wasserdynamik eine entscheidende Rolle bei der erfolgreichen Modellierung der Deacylierungsreaktion spielt. Durch die iterative Verbesserung des Berechnungsaufbaus reproduzierte unser bestes Protokoll, das QM/MM-Störungsberechnungen der freien Energie kombiniert und mehr Wassermoleküle in die QM-Region einbezieht, im Wesentlichen experimentelle Werte. Außerdem scheint die Fähigkeit der Abgangsgruppe der Hauptfaktor zu sein, der die Geschwindigkeitskonstanten von ClpP sowohl für Acylierungs- als auch für Deacylierungsreaktionen mit Ausnahme der Systeme mit Ringöffnung bestimmt.

Der zweite Teil dieser Dissertation konzentrierte sich auf QM/MM-Simulationen von Acylierungsreaktionen für HCV-NS3/4A-Proteasevarianten, die an das mitochondriale antivirale Signalprotein (MAVS) und das NS4A/B-Substrat gebunden sind. Jüngste experimentelle Ergebnisse wiesen auf eine neue Mutation in der HCV-Protease Q41R hin, die für eine signifikante Verbesserung der Reaktivität des Enzyms gegenüber MAVS verantwortlich ist. Die Q41R-Mutation befindet sich relativ weit vom aktiven Zentrum entfernt und ihre Beteiligung am gesamten Reaktionsmechanismus ist daher unklar. Klassische Molekulardynamik und QM/MM wurden verwendet, um die Acylierungsreaktion von HCV-NS3/4A-Proteasevarianten, die an MAVS und das NS4A/4B-Substrat gebunden sind, zu untersuchen und den indirekten Mechanismus aufzudecken, durch den die Q41R-Mutation eine entscheidende Rolle für die effiziente Spaltung des Substrats spielt. Unsere Simulationen zeigen, dass es zwei Hauptkonformationen des MAVS H1'(p)-Restes für die Wildtyp-Protease und nur eine Konformation für die Q41R-Mutante gibt. Der Konformationsraum von H1'(p) wird durch die Q41R-Mutation aufgrund einer π-π-Stapelung zwischen H1'(p) und R41 sowie einer starken Wasserstoffbrücke zwischen dem Rückgrat von H57 und der Seitenkette von R41 eingeschränkt. Weitere QM/MM-Berechnungen zeigen, dass der Komplex mit der durch die Q41R-Substitution ausgeschlossenen Konformation aufgrund seiner höheren freien Energiebarriere für die Acylierungsreaktion eine nicht reaktive Spezies ist. Auf der Grundlage unserer Berechnungen schlagen wir einen kinetischen Mechanismus vor, der experimentelle Daten erklärt, die einen Anstieg der scheinbaren Geschwindigkeitskonstanten für die MAVS-Spaltung in Q41R-Mutanten zeigen. Unser Modell sagt voraus, dass die nicht-reaktive Konformation des Enzym-Substrat-Komplexes die Reaktionskinetik wie ein nicht-kompetitiver Inhibitor moduliert.

List of Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
AAA+	ATPases associated with diverse cellular activities
ACE	acetyl
AM1	Austin model 1
AM1-BCC	Austin model 1 with bond charge correction
AMBER	Assisted Model Building with Energy Refinement
AO	atomic orbital
B3LYP	Becke 1988, 3-parameter, Lee-Yang-Parr functional
B3PW91	Becke 1988, 3-parameter, Perdew-Wang 1991 functional
B88	Becke 1988 functional
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BO	Bond Order
cc-PVQZ	correlation consistent polarized valence quadruple zeta
CGTO	contracted Gaussian-type orbital
CHARMM	Chemistry at Harvard Macromolecular Mechanics
ClpP	caseinolytic protease P
CNDO	complete neglect of differential overlap
COX	cyclooxygenase
DFT	density functional theory
dsRNA	double-stranded Ribonucleic Acid
DZ	double-zeta

EA-VTST/MT	ensemble averaged variational transition state theory/ multidimensional tunnelling
ESP	electrostatic potential
FDA	U.S. Food and Drug Administration
FEP	free energy perturbation
ff	force filed
GAFF	general amber force field
GB	generalized Born
GGA	generalized gradient approximation
GHO	generalized hybrid orbital
GTF	Gaussian-type function
GTO	Gaussian-type orbital
HCV	hepatitis C virus
HEG	homogeneous electron gas
HF	Hartree-Fock
HTS	high-throughput screening
IFN	interferon
INDO	intermediate neglect of differential overlap
Kcat	substrate turnover rate
KM	Michaelis-Menten constant
KS	Kohn-Sham
l.h.s	left-hand side
LA	link-atom
L-BFGS	limited-memory Broyden-Fletcher-Goldfarb-Shanno
LDA	local density approximation
LSCF	local self-consistent field
LSDA	local spin-density approximation
LYP	Lee-Yang-Parr

MAM	mitochondrion-associated membrane
MAVS	mitochondrial antiviral-signalling protein
MD	molecular dynamics
MDA5	melanoma differentiation-associated gene 5
MM	molecular mechanics
MNDO	modified neglect of diatomic overlap
MO	molecular orbital
MP2	second-order Møller-Plesset perturbation theory
N/A	not applicable
NDDO	neglect of diatomic differential overlap
NF-κB	nuclear factor kappa-light-chain-enhancer of activated B cells
NME	N-methylamide
NPT	constant number of particles, constant pressure and constant temperature
NS	non-structural
N-terminal	amino-terminal
NVE	constant number of particles, constant volume and constant energy
NVT	constant number of particles, constant volume and constant temperature
P86	Perdew 1986 functional
PBE	Perdew-Burke-Ernzerhof
PDB	Protein Data Bank
PES	potential energy surface
PI	protease inhibitor
PM3	parametric method number 3
PM6	parametric method number 6
PM7	parametric method number 7

PW91	Perdew-Wang 1991 functional
QM	quantum mechanics
QM/MM	hybrid quantum mechanics/molecular mechanics
r.h.s	right-hand side
RESP	restrained electrostatic potential
RI	resolution of the identity
RIG-I	retinoic acid-inducible gene I
RLRs	RIG-I-like receptors
RM1	Recife model 1
RNA	ribonucleic acid
SaClpP	*Staphylococcus aureus* caseinolytic protease P
SCC-DFTB	self-consistent charge density functional tight-binding
SCF	self-consistent field
SD	Slater determinant
SPC/E	extended simple point charge
STO	Slater-type orbital
Suc-LY-AMC	N-Succinyl-Leu-Tyr-7-amido-4-methylcoumarin
SVP	split valence polarization
TIP3P	transferable intermolecular potential with 3 points
TLR-3	Toll-like receptor 3
TZ	triple-zeta
TZVP	valence triple-zeta polarization
UEG	uniform electron gas
VDZ	valence double-zeta
VWN	Vosko-Wilk-Nusair
WT	wildtype
ZDO	zero differential overlap
Z-LY-CMK	carbobenzyloxyleucyl-tyrosine chloromethyl ketone

Contents

Chapter 1

Introduction

1.1 Serine proteases play an important role in life processes

Proteins are a class of macromolecules composed of one or more chains of polypeptides. They play pivotal roles in physiological processes. Life processes cannot proceed without proteins. Functions of proteins include catalyzing metabolic reactions, cell signalling, providing structure to cells, tissues and organisms, storage of metal ions or amino acids, transporting ions or molecules from one location to another and et cetera. Enzymes are a type of proteins that catalyze biochemical reactions. Without enzymes, most chemical reactions in organisms would take place too slowly to support life processes. Enzyme inhibitors are molecules that slow or halt an enzymatic reaction by binding to the enzyme. Nowadays, a lot of drug molecules are enzyme inhibitors. For example, aspirin inhibits the cyclooxygenase (COX) enzyme to suppress the production of prostaglandins and thromboxanes. This will cause the relief of pain and the prevention of clotting. Based on the type of catalyzed reactions, enzymes can be classified into oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. Proteases, also called peptidases, are a kind of hydrolases that catalyze the cleavage of proteins into smaller polypeptides or even single amino acids by breaking the peptide bonds. Proteases are involved in a lot of biological functions such as digestion of ingested proteins, degradation of misfolded or damaged proteins and cleavage of polyproteins into individual mature proteins. They can be found in all forms of lives and viruses. Proteases can be further classified into serine proteases, cysteine proteases, threonine proteases, aspartic proteases, glutamic proteases, metalloproteases and asparagine peptide lyases based on the key catalytic residue [1]. Serine protease contributes more than one-third of known protease [2]. The classic serine proteases rely on a “charge-relay” system, or called catalytic triad, to conduct reactions [3]. The catalytic triad consists of three residues of aspartate, histidine, and serine. Except for the catalytic triad, the oxyanion hole is also important for the catalysis of serine proteases. It contains the backbone NHs of two residues near the active site. These atoms form a pocket that interacts with the carbonyl group of the peptide bond and stabilizes the negatively charged tetrahedral intermediate. The generally accepted reaction mechanism of serine protease catalysis includes two major steps: acylation and deacylation [2]. In the acylation reaction, serine attacks the carbonyl group of the peptide substrate, assisted by histidine as a general base, to form a tetrahedral intermediate (Figures 3-1 and 4-1). The positively charged histidine residue is stabilized by a hydrogen bond to aspartate. The oxyanion of the intermediate is stabilized by interactions with the backbone NHs of the oxyanion hole. The tetrahedral intermediate collapses with the proton transfer from histidine to the substrate and the departure of the leaving group to yield an acyl-enzyme intermediate. The deacylation reaction essentially repeats the above process: water as a nucleophile attacks the acyl-enzyme forming a second tetrahedral intermediate (Figure 3-2). This intermediate collapses with the release of the product of carboxylic acid and the regeneration of the serine residue.

Although the catalysis reactions of serine proteases have been well studied, some details of the reactions are still not very clear. Firstly, there is no undeniable evidence to demonstrate the existence of the tetrahedral intermediate. The concept of tetrahedral intermediate was inferred from solution chemistry. In the case of good leaving groups such as p-nitrophenol, the lifespan of the tetrahedral intermediate is less than a vibration, the reaction proceeds as a concerted manner and the intermediate must be considered a transition state [4–6]. The hydrolysis of p-nitrophenyl acetate by chymotrypsin is proofed to be a concerted reaction [7]. In some literature, the terms “tetrahedral intermediate” and “transition state” are often used indiscriminately [2]. Thus it is formally possible that acylation and deacylation reactions of a serine protease with its substrate don’t undergo through tetrahedral intermediates. In addition, one theoretical study also suggests that there is no clear intermediate state for the acylation reaction of HCV NS3/NS4A protease with the NS5A/5B substrate [8]. Secondly, despite the fact that several lattice water molecules were observed, it is not very clear how hydrolytic water molecule approaches the acyl-enzyme and starts the deacylation reaction.

Understanding the reaction mechanism at the molecular level can provide direction for the rational design of potential pharmacological inhibitors. In this dissertation, we focused on the finding of major factors which determines the reaction kinetics of both acylation and deacylation reactions. Two model systems of serine proteases were studied by theoretical methods. The first one is *Staphylococcus aureus* ClpP (SaClpP) which is discussed in Chapter 3. The second one is the NS3/4A protease of hepatitis C virus (HCV) and this is described in Chapter 4.

1.1.1 *Staphylococcus aureus* ClpP is a novel target for the treatment of bacterial infection

Caseinolytic protease P (ClpP) is a highly conserved serine protease. Its homologs exist not only in most bacteria but also in eukaryotes and even in *Homo sapiens* [9, 10]. ClpP is important in protein quality control, stress response, and some regulation events [11, 12]. It is also one of the major mechanisms involved in protein degradation [13]. In some pathogenic bacteria, ClpP has been further attributed to functions related to virulence regulation [14, 15]. ClpP alone only shows moderate and unspecific proteolytic activity [16]. It requires AAA+ ATPases (ATPases associated with diverse cellular activities) such as ClpA, ClpC, or ClpX to form a proteolytic active complex [17–21]. The chaperons recognize, unfold, and thread proteins into the inner chamber of ClpP where they are degraded [22]. ClpP is a barrel-shaped homotetradecamer with an inner chamber. It consists of two separate heptameric rings which are stacked upon each other (Figure 1-1). SaClpP has three major conformations: extended, compact and compressed conformations [23–26]. The extended conformation of SaClpP is the only active state, since in the latter two conformations, the catalytic triad doesn't align properly.

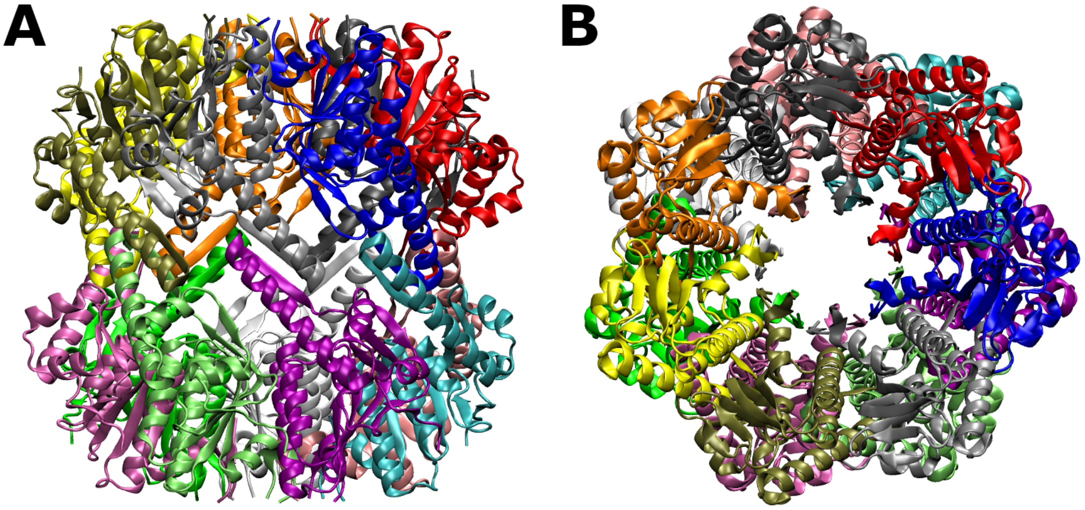


Figure 1-1. (A) The side view and (B) the top view of the structure of *Staphylococcus aureus* ClpP taken from PDB file 3V5E [25]. Chains are shown in different colors.

With the excessive use of traditional antibiotics, multidrug resistance in bacteria has emerged under selective pressure [27]. Bacteria infectious pose once again a severe threat to public health. Therefore, developing a new type of antibiotics is an urgent task for human beings. Since ClpP plays an important role in *Staphylococcus aureus* virulence regulation [14], it is an attractive target for the treatment of bacterial infection. β-lactone was the first type of small-molecule inhibitors discovered for SaClpP [28–30]. The key structural feature of β-lactone is a four-membered ring (Figure 1-2A,B). Later phenyl ester type of inhibitors (Figure 1-2C) was found out by high-throughput screening (HTS) [31]. Except for the blocking of the active site by acylation reaction, compound-induced deoligomerization of protease and dehydroalanine formation of active serine are two other mechanisms of ClpP inhibition [32].

Although ClpP has been extensively researched by biochemical and structural methods [13, 15, 20, 23, 33–36], the detailed mechanism of small-molecule inhibition of this serine protease is still poorly understood. Understanding the inhibition mechanism at the molecular level will provide insights into designing better inhibitors. In Chapter 3, we mainly study the reaction mechanism of directly blocking the active sites of SaClpP and do not consider the other two inhibition mechanisms. Subsequent deacylation reactions were also studied. Since β-lactones and phenyl esters are important inhibitors of SaClpP, three different ligands were investigated. Lig24 and Lig25 are S,S-configured beta-lactones. The former one has a long aliphatic side chain at C-3 position while the latter one has a very short side chain (Figure 1-2A, B). We simulated both ligands to understand the influence of length of the aliphatic side chain on the kinetics behavior of both acylation and deacylation reactions. ML90 is a phenyl ester type of inhibitor which has a methyl group at the α position of the ester group (Figure 1-2C). It shows enhanced acyl-enzyme stability with a moderate reduction of potency compared to its original compound AV170 [31]. In addition, one fluorescent substrate Suc-LY-AMC (Figure 1-2D), which is a popular fluorescent substrate for measuring the activity of proteases, was also simulated.

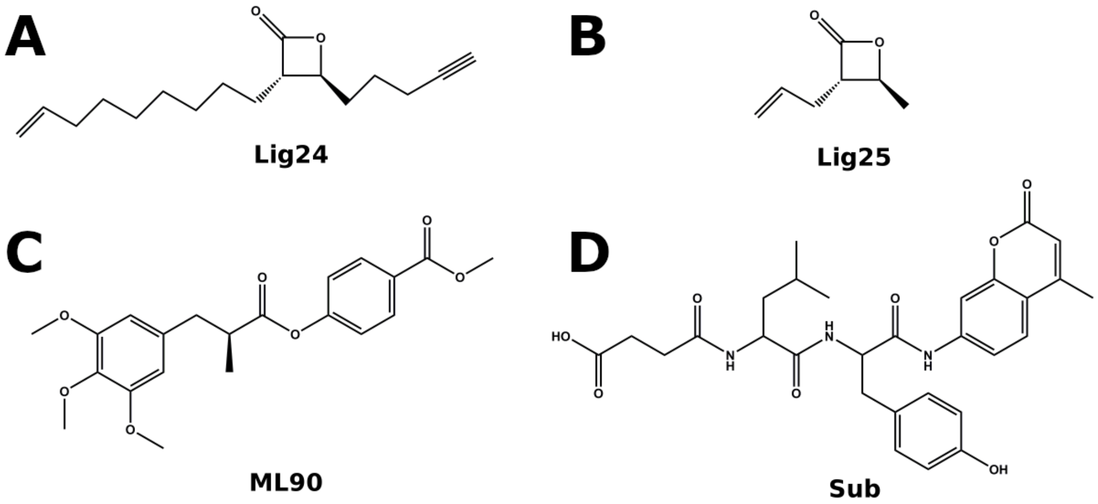


Figure 1-2. Chemical structures of four ligands simulated in acylation and deacylation reactions of SaClpP.

1.1.2 The HCV NS3/4A protease is an important target for drug discovery

WHO estimates that 71 million persons were living with hepatitis C virus (HCV) infection worldwide, accounting for 1% of the population with approximately 0.4 million deaths as the result of complications of chronic HCV infection in 2015 [37]. Every year, 1.75 million people are infected with hepatitis C virus [37]. HCV-related liver disease will progress in a sneaking manner over decades. The advance form of the disease includes liver cirrhosis, which is a critical stage of chronic liver disease. Without antiviral treatment, about 67 to 91% of patients with liver cirrhosis caused by HCV die due to hepatocellular carcinoma or liver failure [38]. In addition, HCV infections increase the likelihood of lymphoma and liver cancer in humans [39, 40]. These facts demonstrate that hepatitis C is still a major public health challenge. HCV is a positive-sense single-stranded RNA virus, its genome encoding both structural proteins and non-structural proteins. The latter include the proteins p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B, which are important for viral maturation and replication. The NS3/4A serine protease, a complex of two of these proteins, is one of the major drug targets for existing antiviral therapeutics. [41, 42]. The protease is a heterodimer of NS3, a 67 kDa protein, and the NS4A cofactor, which is a 54 amino acid membrane-anchored protein. The viral polyprotein translated from HCV's RNA requires cleavage by proteases to mature. The primary function of the NS3/4A protease is the cleavage of viral polyprotein at four junctions of non-structural proteins (NS3/4A, NS4A/4B, NS4B/5A, and NS5A/5B).

In addition to maturation of the viral polyprotein, HCV employs a strategy to evade the immune system by cleaving MAVS ("mitochondrial antiviral-signaling protein", also known as IPS-1, VISA, and Cardif) [43–46], an essential protein for antiviral innate immunity on the mitochondrion-associated membrane (MAM) [46–48]. MAVS-dependent antiviral signaling is initiated after viral double-stranded RNA (dsRNA) is sensed by either Toll-like receptor 3 (TLR-3) or one of two RIG-I-like receptors (RLRs), namely retinoic acid-inducible gene I (RIG-I) and melanoma differentiation-associated gene 5 (MDA5), in the infected cell [49–56]. Although these two pathways differ in their initiating stimuli (different types of dsRNAs) and downstream effectors, they both eventually activate the transcriptional factors, nuclear factor κB (NF-κB), and interferon regulatory factors. This leads to the rapid production of pro-inflammatory cytokines and type I interferons (IFN-α and -β) which promote the subsequent development of adaptive antiviral immunity [54, 57, 58].

Due to its significant role in the life cycle of HCV, the NS3/4A protease is an important target for drug discovery. Two NS3/4A protease inhibitors (PIs), boceprevir and telaprevir [59–62], were first approved by FDA in 2011. These were followed by the subsequent approvals of several noncovalent protease inhibitors including linear (asunaprevir and narlaprevir) [63, 64] and macrocyclic compounds (simeprevir, paritaprevir, vaniprevir, and grazoprevir) [65–68]. However, effective therapy is impeded by the emergence of PI-resistant NS3/4A mutants under drug pressure, such as D168A, R155K, and Q41R. In order to develop inhibitors not susceptible to these resistance mutations, a thorough structural understanding of the protease's catalytic mechanism and interactions with its substrates, namely the viral polyprotein and MAVS, is desirable.

In Chapter 4, we studied acylation reactions of the wild type HCV NS3/4A protease and its Q41R, R155K, D168A, and Q41R-D168A mutants bound to MAVS and the NS4A/4B substrate using both classical molecular dynamics and the QM/MM method (Figure 1-3) [69]. The main focus of our simulations was to uncover the major factors that determine the energy barriers of the acylation reaction and rationalize the mechanism by which PI-resistance mutants influence the kinetics of the reaction.

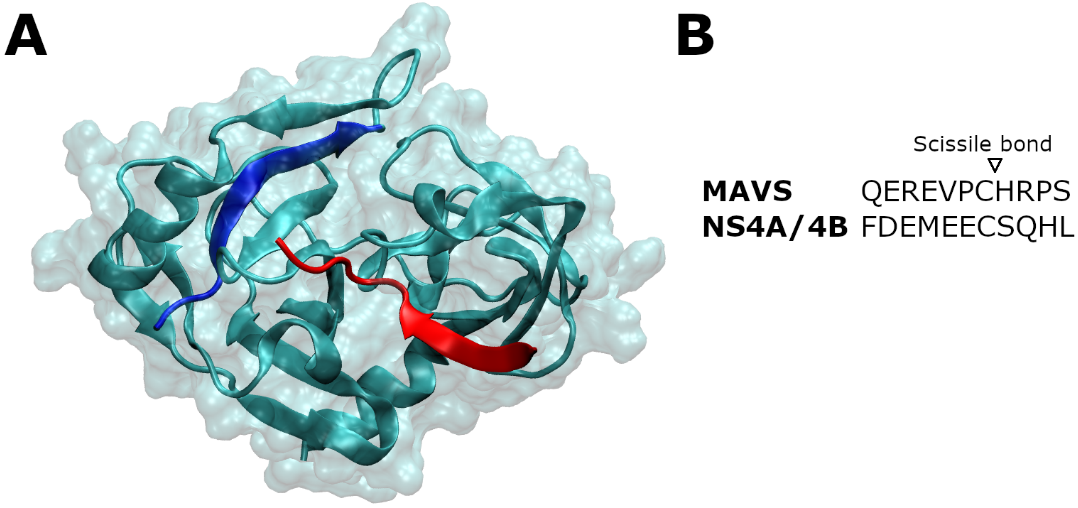


Figure 1-3. (A) The structure of the wildtype HCV NS3/4A serine protease with MAVS modified from PDB file 3RC5 [70]. The NS3 protein is shown in cyan, NS4A cofactor in blue and MAVS in red. (B) Peptide sequences of the NS4A/4B substrate and MAVS simulated in the calculations.

1.2 QM/MM Method is a powerful tool to study enzyme-catalyzed reactions

Computational biochemistry is a subject that uses non-experimental tools to study and analyze systems of biomolecules. By establishing reasonable computational models and checking with experimental data, some properties, functions, and reactions of biomolecules can be explained and predicted. One big advantage of theoretical study is that people can solve some problems which are difficult to be treated by experimental methods such as the reaction path of a reaction, structures of short-lived transition states or intermediates. As technology advances, more and more sophisticated models are built to answer very complicated problems. For large biomolecules, force-field-based molecular mechanics (MM) is a proper method to describe biological macromolecules systems. However, it is not capable of dealing with enzyme-catalyzed reactions because bonds making and breaking are normally involved in the reactions. In this case, the quantum mechanics (QM) method is a good choice. But the QM method has its own limitations. Restricted by computational resources, QM methods are only applied up to a certain amount of atoms. Therefore, the QM/MM method is a compromising method to model the enzyme-catalyzed reactions. The idea of the QM/MM method is fairly simple. The active site of an enzyme is treated by quantum mechanics and the surroundings are dealt with molecular mechanics (Figure 1-4). The concept was originally introduced by Warshel and Levitt in 1976 [71]. Later it has been quickly developed and applied to solve many biomolecular problems. The technical details of different QM methods, the MM method and the QM/MM method are discussed in Chapter 2.

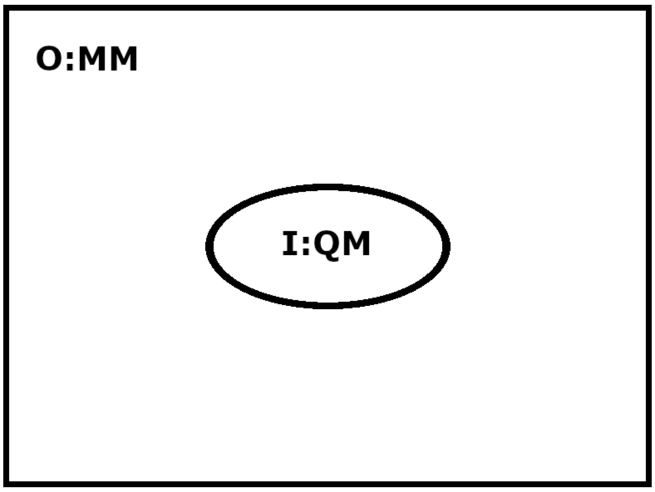


Figure 1-4. The basic concept of the QM/MM method. The entire system is partitioned into inner (I) and outer (O) subsystems. The inner system is treated by the QM method and the outer system is dealt with the MM method.

A number of theoretical papers have investigated the acylation reaction of different serine protease with substrates or inhibitors by using QM/MM methods (Table 1.1) [8, 72–85]. Among these studies, HCV NS3/4A protease is a popular target. Miguel González and co-workers were pioneers in this specific topic. They first studied the reaction mechanism of the HCV NS3/4A protease with the NS5A/5B substrate at the

AM1/CHARMM level. HF and MP2 single point corrections were performed on the QM region. The roles of D79, R109, K136, G137, and R155, which stabilize the electrostatic structure of the transition state, were explained [76]. Later, they used the same system to compare two different boundary treatment methods, the link-atoms (LA) method and the generalized hybrid orbital (GHO) method. In addition, the free energy profile of the acylation reaction at AM1/MM level was also calculated [77]. Afterwards, they further extended the system to include another two substrates, the NS4A/4B and the NS4B/5A substrates [80]. Finally, when SCC-DFTB was used as the QM theory and ensemble averaged variational transition state theory/multidimensional tunnelling (EA-VTST/MT) methods were employed, it was found that there is no clear intermediate state for the acylation reaction of the NS3/4A protease with the NS5A/5B substrate [8]. Jindal et al. used the EVB method to calculate activation free energies for wild type and several protease mutations with the NS5A/5B substrate [86].

Table 1.1. Summary of some previous QM/MM studies on serine proteases

Protein	Ligand	QM level	MM level	Calculation type	Refs
Acetylcholin-esterase	acetylcholine	HF, MP2, B3LYP	AMBER	EM, MD	[72, 81]
Trypsin	Peptide (CPRIWM), MCTI-A	HF, MP2, B3LYP	AMBER	FEP, US(MD)	[73, 75, 82, 84]
Neutrophil elastase	Peptidyl α-ketoheterocyclic inhibitors	PM3, B3LYP	AMBER	EM	[74]
HCV NS3/4A protease	NS5A/5B (EDVVCCSMSY) NS4A/4B (DEMEECSQHL) NS4B/5A (ECTTPCSGSW)	AM1, HF, MP2 SCC-DFTB	CHARMM	EM, US(MD)	[8, 76, 77, 80]
Furin H5N1	hemagglutinin (RERRRKKRGL)	PM3, B3LYP	CHARMM	EM	[78]
Lipase	(S)-MPP acetate (M)-MPP acetate	BLYP	AMBER	Metadynamics	[79]
DENV NS2B-NS3pro	CH3NH-RRPV-COCH3	PM3	AMBER	US(MD)	[83]
Subtilisin	Chymotrypsin inhibitor 2	B3LYP	AMBER	US(MD)	[85]

Theoretical studies summarized in Table 1.1 used either AMBER or CHARMM force fields which are two very popular forced fields. Some of these studies used traditional static energy minimization to obtain reaction profiles. Other studies applied umbrella sampling to calculate free energy profiles. In addition, there are only a few papers simulating deacylation reactions of serine proteases [81, 85, 87]. This step is difficult to model because one water molecule is involved in the reaction. Besides, all these studies used traditional one-dimensional reaction coordinate (equations (3-1), (3-4), and (4-1)) except for the one that used metadynamics (they used another set of reaction coordinates). Although some studies calculated so-called 2D potential energy surfaces, they simply divided the 1D reaction coordinate into two parts. In Chapter 3, we established a solid framework for calculating two-dimensional potential energy surfaces by introducing a second reaction coordinate which controls the proton transfer from histidine to the ligand to study the entire reaction cycle of SaClpP with four different ligands in details. QM/MM calculations with semi-empirical QM methods and DFT methods were also compared. To the best of our knowledge, no group has ever studied enzyme reactions of ClpP with either inhibitors or substrates by QM/MM methods until now. In Chapter 4, we used the protocol developed in Chapter 3 to simulate acylation reactions of HCV NS3/4A protease variants with the NS4A/4B substrate and MAVS. Our model successfully reproduces experimental data [88] and give an explanation at the molecular level.

Chapter 2

Theoretical background

2.1 Quantum Mechanics and Wave Functions

2.1.1 Schrödinger Equation

Quantum mechanics is one of the most successful mechanics to describe microscopic systems, for example, a molecule. The fundamental postulate of quantum mechanics is that the state of a system can be described by a wave function. The wave function contains all information about the system. Well-behaved wave functions require single-valued, continuous and quadratically integrable functions. Another important postulate in quantum mechanics is that each physically observable property corresponds a linear Hermitian operator. The only possible values that can be obtained from measurements for the physical observable B are the eigenvalues b_i in the equation

$$\hat{B}g_i = b_i g_i \tag{2-1}$$

where $\hat{B}$ is the corresponding operator to the physically observable property B and g_i are the eigenfunctions. When the target physical observable is energy of atoms and molecules, then the operator in equation (2-1) is the Hamitonian H. In this case, it becomes the well-known time-independent Schrödinger equation [89]:

$$\hat{H}\psi = E\psi \tag{2-2}$$

where ψ is the wave function of the system. If we treat electrons and the nuclei as point masses and neglect spin-orbit and relativistic effects, then the molecular Hamiltonian operator can be written as

$$\begin{aligned}\hat{H} = &-\frac{\hbar^2}{2}\sum_{\alpha}\frac{1}{m_\alpha}\nabla_\alpha^2 - \frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 + \sum_\alpha\sum_{\beta>\alpha}\frac{Z_\alpha Z_\beta e^2}{4\pi\varepsilon_0 r_{\alpha\beta}} - \sum_\alpha\sum_i \frac{Z_\alpha e^2}{4\pi\varepsilon_0 r_{i\alpha}} \\ &+ \sum_j\sum_{i>j}\frac{e^2}{4\pi\varepsilon_0 r_{ij}}\end{aligned} \tag{2-3}$$

where $\hbar$ is Planck's constant divided by 2π, m_α is the mass of nucleus α, m_e is the mass of the electron, ∇^2 is the Laplacian operator, e is the electron charge, Z_α is atomic number of nucleus α, and r_{ab} is the distance between two particles a and b. In this equation, Greek letters refer to nuclei and Latin letters refer to electrons.

2.1.2 The Born-Oppenheimer Approximation and Potential Energy Surface

Accurate wave functions for many-particle molecular systems are extremely difficult to solve because of the correlation of particles. The Hamiltonian in equation (2-3) contains pairwise terms of attraction and repulsion, which means no particle is moving independently of others. Nuclei are more massive than electrons and masses of nuclei and electrons appear in the denominators of the kinetic energy terms in equation (2-3). This means, under typical physical conditions, the electrons of molecular systems are moving much faster than the nuclei. It can be considered that electrons relax with respect to nuclear motion instantaneously. As such, it is convenient to separate these two motions and compute electronic energies with fixed positions of nuclei. That is, the kinetic energy term of nuclei is independent of the electrons. Correlation in the electron-nuclei attraction term is eliminated and the nuclei-nuclei repulsion term becomes a constant for a given geometry. The approximation of decoupling electronic and nuclear motions is called the Born-Oppenheimer Approximation [90]. Thus, the electronic Schrödinger equation can be written as

$$(\hat{H}_{el} + V_{NN})\psi_{el} = U\psi_{el} \tag{2-4}$$

where ψ_{el} is the electronic wave function and purely electronic Hamitonian $\hat{H}_{el}$ is

$$\hat{H}_{el} = -\frac{\hbar^2}{2m_e}\sum_i \nabla_i^2 - \sum_\alpha \sum_i \frac{Z_\alpha e^2}{4\pi\varepsilon_0 r_{i\alpha}} + \sum_j \sum_{i>j} \frac{e^2}{4\pi\varepsilon_0 r_{ij}} \tag{2-5}$$

The nuclei repulsion term V_{NN} is

$$V_{NN} = \sum_\alpha \sum_{\beta>\alpha} \frac{Z_\alpha Z_\beta e^2}{4\pi\varepsilon_0 r_{\alpha\beta}} \tag{2-6}$$

With the Born-Oppenheimer Approximation, we can establish the concept of potential energy surface. For each possible nuclear configuration, one can solve the electronic Schrödinger equation (2-4) and get a series of electronic wave functions with corresponding electronic energies. For a given geometry, each solved electronic wave function corresponds to a different electronic state. The state with lowest electronic energy is called ground state. The electronic energy of the ground state is thus a function of nuclear positions. Furthermore, ground-state electronic energies over all possible nuclear coordinates define the potential energy surface (PES) of a molecular system. As we are only interested in the ground state of protein systems, in the later part of this dissertation, all electronic energies refer to the energies of the ground state for simplicity. Since the large number of variables, electronic energy is actually a "surface" in a space of $3N - 6$ dimensions (for a non-linear molecule and N being the number of atoms). In practice, the potential energy surface of a reaction is usually defined by one or two variables, which are also called reaction coordinates. In our cases, potential energy depends on two variables, then a plot of electronic energy and two variables in three dimensions forms a surface in traditional three-dimensional space. The concept of potential energy surface is the fundamental of quantum chemistry. With the idea of PES, we can further define the transition state and intermediate of a chemical reaction on the PES.

2.1.3 The Variation Theorem

In most cases, the electronic Schrödinger equation (2-4) cannot be solved analytically. One way to avoid solving the electronic Schrödinger equation is using the variation method to approximate ground-state electronic energy. For a given Hamiltonian of a system, if ϕ is an arbitrary normalized, well-behaved function satisfying the boundary conditions of the problem, it can be proven that

$$\int \phi^* \hat{H} \phi \, d\tau \geq E_0 \tag{2-7}$$

where E_0 is the ground-state energy of the system and $d\tau$ denotes integration over all coordinates of a system. The function ϕ is also called a trial variation function. The integral of the left-hand side (l.f.s) of equation (2-7) is called variational integral. The significance of the variation theorem is that it provides a standard to judge the quality of the trial function. The lower the outcome of variational integral, the better the trial function is.

2.2 Quantum Mechanics Methods

There are several approaches to obtain the potential energy of a system as a function of nuclear coordinates. These methods mainly include *ab initio* methods, semi-empirical methods, density functional theory and force field (molecular mechanics) methods. They will be discussed in sections 2.2 and 2.3.

2.2.1 The Hartree-Fock Method

The Hartree-Fork SCF method [91–93] is one of the most basic *ab initio* methods. In this dissertation, open-shell configurations are not involved. Thus, we will only consider the situation of closed-shell configurations. Closed-shell means all electrons spin-paired and there is no orbital with single electron. For simplicity, all equations in this section are shown in atomic unit.

An electron is a fermion, which means the wave function of electrons should be antisymmetric. The HF method uses one single Slater Determinant [94]

$$\psi_{\text{SD}} = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \dots & \chi_n(1) \\ \chi_1(2) & \chi_2(2) & \dots & \chi_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(n) & \chi_2(n) & \dots & \chi_n(n) \end{vmatrix} \tag{2-8}$$

as an ansatz of electronic wave function. In equation (2-8), n is the number of electrons and χ refers to a spin-orbital of an electron, which is the product of a spatial orbital and an electron spin function (α or β). The HF energy E_{HF} is calculated by

$$E_{\text{HF}} = \left\langle \psi_{\text{SD}} | \hat{H}_{el} + V_{NN} | \psi_{\text{SD}} \right\rangle = 2\sum_{i=1}^{n/2} H_{ii}^{\text{core}} + \sum_{i=1}^{n/2}\sum_{j=1}^{n/2}(2J_{ij} - K_{ij}) + V_{NN} \tag{2-9}$$

$$H_{ii}^{\text{core}} = \langle \phi_i(1) | H^{\text{core}}(1) | \phi_i(1) \rangle \tag{2-10}$$

$$\hat{H}^{\text{core}}(1) = -\frac{1}{2}\nabla_1^2 - \sum_{\alpha} \frac{Z_\alpha}{r_{1\alpha}} \tag{2-11}$$

$$J_{ij} = \left\langle \phi_i(1)\phi_j(2) \left| \frac{1}{r_{12}} \right| \phi_i(1)\phi_j(2) \right\rangle \tag{2-12}$$

$$K_{ij} = \left\langle \phi_i(1)\phi_j(2) \left| \frac{1}{r_{12}} \right| \phi_j(1)\phi_i(2) \right\rangle \tag{2-13}$$

where H_{ii}^{core} are the one-electron integrals, ϕ_i represent spatial orbitals, $\hat{H}^{\text{core}}$ is the one-electron core Hamiltonian, J_{ij} and K_{ij} are the Coulomb integrals and the exchange integrals respectively. The classic analogue of J_{ij} is the Coulomb repulsion between electrons i and j. However, there is no analogue of exchange interactions in classical mechanics. Exchange only occurs between two electrons with the same spin. The source of exchange interactions is the Pauli exclusion principle.

By using variational principle, the HF method looks for a set of molecular orbitals (MOs) ϕ_i minimizing the variational integral. For computational convenience molecular orbitals are taken to be orthonormal. This is because one can always obtain a Slater

Determinant of orthonormal MOs from another Slater Determinant of non-orthonormal MOs by Gram-Schmidt Orthonormalization or other procedure. Given a purely electronic Hamiltonian (equation (2-5)), it can be further proven [95] that the minimization condition of variation integral is equivalent to the set of MOs satisfying

$$\hat{F}(1)\phi_i(1) = \varepsilon_i \phi_i(1) \tag{2-14}$$

where ε_i is the energy of the i-th orbital. The one-electron Fock operator $\hat{F}$ is

$$\hat{F}(1) = \hat{H}^{\text{core}}(1) + \sum_{j=1}^{n/2} \left[2\hat{J}_j(1) - \hat{K}_j(1)\right] \tag{2-15}$$

where $\hat{J}_j$ is the coulomb operator and $\hat{K}_j$ is the exchange operator. They are defined by

$$\hat{J}_j(1)f(1) = f(1)\int |\phi_j(2)|^2 \frac{1}{r_{12}} d\tau_2 \tag{2-16}$$

$$\hat{K}_j(1)f(1) = \phi_j(1)\int \frac{\phi_j^*(2)f(2)}{r_{12}} d\tau_2 \tag{2-17}$$

where f is an arbitrary function.

In practice, each spatial orbital is usually written as a linear combination of a set of pre-defined basis functions (see more details in section 2.2.4):

$$\phi_i = \sum_{\nu=1}^{b} c_{\nu i}\,\chi_\nu \tag{2-18}$$

Substitution of equation (2-18) into the Hartree-Fock equation (2-14) with multiplication by χ_μ^* and integration gives

$$\sum_{\nu=1}^{b} c_{\nu i}\left(F_{\mu\nu} - \varepsilon_i S_{\mu\nu}\right) = 0, \quad \mu = 1, 2, \ldots, b \tag{2-19}$$

$$F_{\mu\nu} = \langle \chi_\mu \,|\, \hat{F} \,|\, \chi_\nu \rangle \tag{2-20}$$

$$S_{\mu\nu} = \langle \chi_\mu \,|\, \chi_\nu \rangle \tag{2-21}$$

where $F_{\mu\nu}$ are Fork matrix elements and $S_{\mu\nu}$ are the overlap integrals. The equations (2-19) are called Roothaan equations [96], which form a set of linear homogeneous equations with b unknowns $c_{\nu i}$. For a nontrivial solution, the following determinant must be 0:

$$\begin{vmatrix} F_{11}-\varepsilon S_{11} & F_{12}-\varepsilon S_{12} & \dots & F_{1n}-\varepsilon S_{1n} \\ F_{21}-\varepsilon S_{21} & F_{22}-\varepsilon S_{22} & \dots & F_{2n}-\varepsilon S_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ F_{n1}-\varepsilon S_{n1} & F_{n2}-\varepsilon S_{n2} & \dots & F_{nn}-\varepsilon S_{nn} \end{vmatrix} = 0 \tag{2-22}$$

The equation (2-22) is called secular equation and the roots of the equation give the orbital energies ε_i. Fock matrix element $F_{\mu\nu}$ can be calculated by

$$F_{\mu\nu} = \hat{H}^{\text{core}}_{\mu\nu} + \sum_{\lambda=1}^{b}\sum_{\sigma=1}^{b} P_{\lambda\sigma}\left[(\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\lambda|\nu\sigma)\right] \tag{2-23}$$

$$\hat{H}^{\text{core}}_{\mu\nu} = \langle\chi_\mu(1)|\hat{H}^{\text{core}}(1)|\chi_\nu(1)\rangle \tag{2-24}$$

$$P_{\lambda\sigma} = 2\sum_{j=1}^{n/2} c^*_{\lambda j}\, c_{\sigma j}, \quad \lambda = 1,2,\dots,b, \quad \sigma = 1,2,\dots,b \tag{2-25}$$

where $P_{\lambda\sigma}$ are called density matrix elements and the two-electron repulsion integral $(\mu\nu|\lambda\sigma)$ is defined as:

$$(\mu\nu|\lambda\sigma) = \iint \frac{\chi^*_\mu(1)\chi_\nu(1)\chi^*_\lambda(2)\chi_\sigma(2)}{r_{12}}\, d\tau_1\, d\tau_2 \tag{2-26}$$

The Hartree-Fock energy E_{HF} can also be calculated from Fock, density and core Hamiltonian matrix elements:

$$E_{\text{HF}} = \frac{1}{2}\sum_{\mu=1}^{b}\sum_{\nu=1}^{b} P_{\mu\nu}\left(F_{\mu\nu} + H^{\text{core}}_{\mu\nu}\right) + V_{NN} \tag{2-27}$$

The Roothaan equations are mostly solved by matrix methods. The matrix form of Roothaan equations is

$$\mathbf{FC} = \mathbf{SC}\boldsymbol{\varepsilon} \tag{2-28}$$

where **F** is the Fock matrix, **C** is the matrix formed by coefficients $c_{\nu i}$, **S** is the overlap matrix and $\boldsymbol{\varepsilon}$ is a diagonal matrix in which diagonal elements are the orbital energies ε_i. The Roothaan equations must be solved iteratively since Fock matrix elements depend on the MOs ϕ_i, which depend on the unknown coefficients $c_{\nu i}$. One has to guess an initial set of coefficients $c_{\nu i}$ and iteratively calculate the density matrix **P** until the convergence condition is met. This method is also called self-consistent field (SCF) method.

A typical SCF HF calculation contains the following steps [97]:

1. Choose a basis set.
2. Evaluate the one-electron integrals, two-electron repulsion electrons and overlap

integrals.

3. Perform an Orthonormalization procedure to obtain the **A** matrix of coefficients $a_{\mu\nu}$ that produces a new set of orthonormal basis functions.
4. Guess initial coefficients $c_{\nu i}$ and calculate the density matrix $\mathbf{P}_0$.
5. Calculate Fock matrix elements $F_{\mu\nu}$ from one-electron integrals, two-electron integrals and the density matrix $\mathbf{P}_n$.
6. Calculate the matrix **F'** by **F'** $=$ $\mathbf{A}^{\mathbf{T}}\mathbf{FA}$.
7. Diagonalize the matrix **F'** to obtain the eigenvalue matrix ε and eigenvector matrix **C'** of **F'**.
8. Calculate the coefficient matrix **C** by **C** = **AC'**.
9. Calculate a new density matrix $\mathbf{P}_{n+1}$ from **C** by $\mathbf{P}^* = 2\mathbf{C}\mathbf{C}^{\mathbf{T}}$.
10. Compare the matrix $\mathbf{P}_{n+1}$ with the preceding matrix $\mathbf{P}_n$. If they differ negligibly based on the convergence condition, the calculation has converged. The converged SCF wave function is used to calculate molecular properties. Otherwise, go back to step 5 and calculate a new Fork matrix from the current density matrix $\mathbf{P}_{n+1}$ and then do the succeeding steps.

Although the HF method can give accurate exchange interactions, the main limitation of the method is the lack of electron correlation. There are a lot of post-Hartree-Fock *ab initio* methods which take electron correlation into account. Since these methods are beyond the scope of this dissertation, they will not be discussed here.

2.2.2 Semi-empirical Methods

The main cost of performing HF calculations is the calculation of two-electron repulsion integrals. The main idea of semi-empirical methods is to reduce the number of these integrals. The first step is to only consider the valence electrons of atoms. The core electrons are treated together with the nucleus. In addition, a minimal basis set (see section 2.2.4.2) of Slater AOs is used for the valence electrons. The key assumption of semi-empirical methods is the assumption of Zero Differential Overlap (ZDO), which neglects the product of basis functions of the same electron centered on different atoms:

$$\chi_\mu(1)\chi_\nu(1) = 0 \tag{2-29}$$

where χ_μ and χ_ν are centered on different atoms. Note that in equation (2-29) it is the product of the two basis functions instead of an integral over such a product. The consequences of the ZDO approximation include: 1) The overlap matrix **S** becomes a unit matrix. 2) Three-center one-electron integrals (one from the operator) are neglected.

3) Three-center and four-center two-electron integrals vanish.

There are several semi-empirical QM methods, which differ based on the amount and the method of simplifications. The neglect of diatomic differential overlap (NDDO) method [98] only uses the approximation mentioned above. The intermediate neglect of differential overlap (INDO) method [99] neglects all two-center two-electron integrals which are not of the Coulomb type in addition to the NDDO approximation. And the two-center two-electron integral only depends on the types of the two atoms. The complete neglect of differential overlap (CNDO) method [98, 100] further simplifies the one-center two-electron integrals. All one-center repulsion integrals on atom A have the same value of γ_{AA}. All these semi-empirical methods still use the SCF procedure to obtain the energy of the system. The aim of INDO of CNDO methods is using less computational resource to reproduce HF results with minimal basis set. On the contrary, the aim of NDDO methods is try to obtaining energies of chemical systems with chemical accuracy instead of reproducing the HF results. In some cases, the NDDO methods can give better results than the HF method because the choice of the proper parameters can compensate the missing of electron correlation in the HF method. Since CNDO and INDO methods are rarely used nowadays, only NDDO methods are discussed in the following sections.

2.2.2.1 MNDO

Modified neglect of diatomic overlap (MNDO) method [101] developed by Dewar and Thiel is a NDDO method. The diagonal element of the Fock matrix is written as

$$\begin{aligned} F_{\mu\mu} = U_{\mu} - \sum_{B \neq A} C_B(\mu\mu|s_B s_B) + \sum_{\nu \in A} P_{\nu\nu}\left[(\mu\mu|\nu\nu) - \frac{1}{2}(\mu\nu|\mu\nu)\right] \\ + \sum_{B}\sum_{\lambda \in B}\sum_{\sigma \in B} P_{\lambda\sigma}(\mu\mu|\lambda\sigma) \end{aligned} \tag{2-30}$$

where μ is the basis function centered on atom A. U_{μ} is the atomic orbital ionization potential. The second term on the right-hand side (r.h.s) of equation (2-30) reflects the attraction of electron on orbital μ to the other nuclei. C_B is the core charge on atom B, which is equal to the atomic number minus the number of core electrons of atom B. s_B refers to a valence s orbital of atom B. The third term represents the Coulomb and exchange interactions between two electrons on atom A and the last term represents Coulomb interactions between one electron on atom A and another electron on atom B.

An off-diagonal element of the Fock matrix for two basis functions centered on atom

A is written as

$$\begin{aligned} F_{\mu\nu} = &- \sum_{B \neq A} C_B(\mu\nu|s_B s_B) + P_{\mu\nu}\left[\frac{3}{2}(\mu\nu|\mu\nu) - \frac{1}{2}(\mu\mu|\nu\nu)\right] \\ &+ \sum_{B} \sum_{\lambda \in B} \sum_{\sigma \in B} P_{\lambda\sigma}(\mu\nu|\lambda\sigma) \end{aligned} \tag{2-31}$$

Other matrix element is written as

$$F_{\mu\nu} = \frac{1}{2}(\beta_\mu + \beta_\nu)S_{\mu\nu} - \frac{1}{2}\sum_{\lambda \in A}\sum_{\sigma \in B} P_{\lambda\sigma}(\mu\lambda|\nu\sigma) \tag{2-32}$$

where μ and ν centered on atom A and B respectively. The first term on the r.h.s of equation (2-32) is the resonance integral which reflects one-electron kinetic energy and attraction to the nuclei. β_μ and β_ν are two-center one-electron resonance integrals for atomic orbitals μ and ν respectively. $S_{\mu\nu}$ is the element of overlap matrix which is not consistent with the ZDO approximation. This is why the MNDO method is called "modified".

If an atom has only s and p orbitals, there are five unique one-center two-electron integrals:

$$\begin{aligned} (ss|ss) &= G_{ss} \\ (ss|pp) &= G_{sp} \\ (pp|pp) &= G_{pp} \\ (pp|p'p') &= G_{pp'} \\ (sp|sp) &= H_{sp} \end{aligned} \tag{2-33}$$

These values are taken from spectroscopic data. The nuclear repulsion energy of MNDO method is calculated by

$$V_{AB}^{\text{MNDO}} = C_A C_B (s_A s_A | s_B s_B)\left(\tau e^{-\alpha_A r_{AB}} + e^{-\alpha_B r_{AB}}\right) \tag{2-34}$$

$$V_N = \sum_{B>A}^{\text{nuclei}} C_A C_B (s_A s_A | s_B s_B)\left(1 + \tau e^{-\alpha_A r_{AB}} + e^{-\alpha_B r_{AB}}\right) \tag{2-35}$$

where C_A and C_B are core charges. α is a parameter for each atom type. τ is equal to 1 unless two atoms A and B are O–H or N–H pair. In MNDO, there are six optimized parameters fitted to experimental data for each kind of atom. These parameters include one-center one-electron integral U_{ss} and U_{pp}, the STO orbital exponent ζ ($\zeta_s = \zeta_p$ for MNDO), two-center one-electron resonance integrals β_s and β_p, and α.

2.2.2.2 AM1

AM1 (Austin Model 1) is an improved version of MNDO developed by Dewar and co-workers [102]. The main difference between MNDO method and AM1 method is the nuclear repulsion term. The nuclear repulsion energy of AM1 method is computed by

$$V_{AB}^{\mathrm{AM1}} = V_{AB}^{\mathrm{MNDO}} + \frac{C_A C_B}{r_{AB}} \sum_k \left[a_{Ak} e^{-b_{Ak}(r_{AB}-c_{Ak})^2} + a_{Bk} e^{-b_{Bk}(r_{AB}-c_{Bk})^2} \right] \tag{2-36}$$

where k is between 2 and 4 depending on atom type. The quantities a, b and c are fitted parameters. Thus, compared to the MNDO method, AM1 has more optimized parameters for each kind of atom.

2.2.2.3 PM3

In 1989, Stewart re-parametrized AM1 and named the new method as PM3 (Parameterized Model 3) [103]. The main difference between PM3 and AM1 is as follows. The one-center two-electron integrals in equations (2-33) are now taken as optimized parameters instead of from atomic spectral data. The number of Gaussian terms for each atom in the core-repulsion function (equation (2-36)) is limited to only two. In addition, a different method was used to optimize parameters for PM3.

2.2.2.4 PM6

Stewart published the PM6 method in 2007 [104]. It has parameters for 70 elements. More than 9000 compounds were used in the parametrization, including both experimental and high level calculated data. The core–repulsion term in PM6 is defined as

$$V_{AB}^{\mathrm{PM6}} = C_A C_B (s_A s_A | s_B s_B) x_{AB} e^{-\alpha_{AB}(R_{AB}+0.0003R_{AB}^6)} + g_{AB} \tag{2-37}$$

$$g_{AB} = 10^{-8} (\frac{Z_A^{1/3} + Z_B^{1/3}}{R_{AB}})^{12} \tag{2-38}$$

where R_{AB} is the distance between atoms A and B in angstroms. x_{AB} and α_{AB} are two-atom parameters depending on atoms A and B. The function g_{AB} represents the repulsive interaction between the cores of A and B. Z_A and Z_B are the atomic numbers of atoms A and B. For the atom pairs NH and OH, the exponential in equation (2-38) is replaced by exp($\alpha_{AB} R_{AB}^2$) to provide a better representation of hydrogen bonding. Besides, an additional term is included in the core-repulsion term for the atom pair CC in order to improve the accuracy for carbon-carbon triple bonds.

PM6 uses Slater orbitals as the basis functions. In addition to s and p valence orbitals, d orbitals are required for transition elements. To achieve better performance, PM6 includes d orbitals for many main-group non-metals as well.

Other famous NDDO semi-empirical methods include PM6-D3H4 [105] and PM7 [106], which are not used in the present work, are not discussed in this chapter.

2.2.2.5 RM1

The RM1 method (Recife Model 1) has the same structure as AM1 [107]. But all parameters were re-evaluated using a new set of data from 1736 molecules compared to around 200 molecules used for AM1. Since RM1 parameters are only available for 10 elements, it is less widely used than AM1 and PM3.

2.2.2.6 DFTB

The self-consistent-charge density-functional tight-binding (SCC-DFTB) method is a semi-empirical QM method and similar to the semi-empirical MO methods [108]. The PBE functional is usually used as the exchange-correlation energy functional in SCC-DFTB. The method only treats the valence electrons explicitly and uses a minimal basis set of AOs to expand the Kohn–Sham orbitals. It neglects many integrals and makes approximations for many other integrals. The method contains single-atom parameters which are derived from atomic DFT calculations and interatomic parameters in repulsion-energy functions. The interatomic parameters are calculated from bond-stretching energies by the B3LYP functional with a double-zeta or triple-zeta basis set. Therefore, the parameters are found using DFT calculations instead of experimental data. The time for a DFTB calculation is similar to the time for other semi-empirical calculation such as AM1 or PM3, but the results are usually more accurate than these methods.

2.2.3 Density Functional Theory

One problem of the traditional wave function method is the large number of variables. The electronic wave function of a molecule with n electrons contains $4n$ variables, $3n$ spatial and n spin coordinates. Furthermore, the wave function lacks direct physical significance and is difficult to interpret. Therefore, people were trying to search some physical observables with less variables which determines the electronic energy and possibly other properties. Electron density is such a physical observable which has only 3 variables. Density functional theory (DFT) was developed by the idea of using electron density to calculate the properties of molecular systems.

2.2.3.1 The Hohenberg-Kohn Existence Theorem

The purely electronic Hamiltonian of an n-electron molecule, in atomic units, is

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{n}\nabla_i^2 + \sum_{i=1}^{n} v(\boldsymbol{r}_i) + \sum_{j}\sum_{i>j}\frac{1}{r_{ij}} \tag{2-39}$$

$$v(\boldsymbol{r}_i) = -\sum_{\alpha}\frac{Z_\alpha}{r_{i\alpha}} \tag{2-40}$$

where $v(\boldsymbol{r}_i)$ is the function of coordinates x_i, y_i and z_i of electron i. This potential is called external potential because it is produced by charges external to the electrons of the system. In 1964, Hohenberg and Kohn [109] proved that for a nondegenerate ground state of a molecule, the electron density determines the external potential, which in turn determines the Hamiltonian of the molecule. The intuitive proof of this theorem is quite simple. The number of electrons is defined by the integral of the electron density. The positions of the nuclei are defined by the cusps of the electron density and the corresponding nuclear charges are defined by the heights of the cusps. In principle, one can further solve the Schrödinger equation and obtain the ground-state wave function. Thus, the electronic energy and wave function of the ground state are uniquely determined by ground-state electron density of the molecule. In addition, the ground-state electronic energy E_0 is a functional of electron density ρ_0 and can be written as

$$E_0 = E[\rho_0] = T[\rho_0] + V_{Ne}[\rho_0] + V_{ee}[\rho_0] \tag{2-41}$$

where T, V_{Ne} and V_{ee} are all functional of electron density ρ_0 and refer to the kinetic energy of electrons, attraction to the nuclei and interactions between electrons respectively. The V_{Ne} term can be accurately calculated by

$$V_{Ne} = \left\langle \psi_0 \left| \sum_{i=1}^{n} v(\boldsymbol{r}_i) \right| \psi_0 \right\rangle = \int \rho_0(\boldsymbol{r})\, v(\boldsymbol{r})\, d\boldsymbol{r} \tag{2-42}$$

Thus, in equation (2-41), the V_{Ne} term is known, whereas the form of other two functionals are unknown.

2.2.3.2 The Hohenberg-Kohn Variational Theorem

The first Hohenberg-Kohn theorem is an existence theorem. It is unhelpful to predict the electron density of the system. Similar to the variation theorem described in section 2.1.3, Hohenberg and Kohn [109] proved that for each trial density function $\rho_{tr}(\boldsymbol{r})$ that satisfies

$$\int \rho_{tr}(\boldsymbol{r})\, d\boldsymbol{r} = n, \;\; \rho_{tr}(\boldsymbol{r}) \geq 0 \tag{2-43}$$

the following inequality holds

$$T[\rho_{tr}] + V_{ee}[\rho_{tr}] + \int \rho_{tr}\, v(\boldsymbol{r})\, d\boldsymbol{r} \geq E[\rho_0] \tag{2-44}$$

In another words, the ground-state electron density minimizes the electronic energy functional $E[\rho]$.

2.2.3.3 The Kohn-Sham SCF Method

Kohn and Sham [110] first considered a fictitious reference system of non-interacting electrons. The electron density of the fictitious system is the same as the electron density of the real system where the electrons do interact. The exact solution to the non-interacting system is a Slater-determinant wave function of Kohn-Sham spin orbitals

$$\psi_0 = |u_1^{\text{KS}}\, u_2^{\text{KS}}\, \ldots\, u_n^{\text{KS}}| \tag{2-45}$$

where Kohn-Sham spin orbital is the product of its spatial part and a spin function σ (either α or β)

$$u_i^{\text{KS}} = \theta_i^{\text{KS}}(\boldsymbol{r}_i)\sigma_i \tag{2-46}$$

Let $\Delta T[\rho]$ be defined by

$$\Delta T[\rho] = T[\rho] - T_s[\rho] \tag{2-47}$$

where $T_s[\rho]$ is the electronic kinetic energy of the reference system and ΔT is the difference in electronic kinetic energy between the real system and the reference system. Let

$$\Delta V_{ee}[\rho] = V_{ee}[\rho] - \frac{1}{2}\iint \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{r_{12}}\, d\boldsymbol{r}_1 d\boldsymbol{r}_2 \tag{2-48}$$

where the second term on the r.h.s of the equation is the classical expression of electrostatic repulsion energy when electrons are treated as a continuous distribution of charges with electron density ρ.

With the definitions (2-47) and (2-48), equation (2-41) becomes

$$E[\rho] = \int \rho(\boldsymbol{r}) v(\boldsymbol{r})\, d\boldsymbol{r} + T_s[\rho] + \frac{1}{2}\iint \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{r_{12}}\, d\boldsymbol{r}_1 d\boldsymbol{r}_2 + \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{2-49}$$

The exchange-correlation energy functionals $E_{xc}[\rho]$ is defined by

$$E_{xc}[\rho] = \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{2-50}$$

This term contains not only the quantum mechanical exchange and correlation interactions, but also the correction for the self-interaction energy and for the difference in kinetic energies between the real system and the fictitious non-interacting system. Now we have

$$E_0 = E[\,\rho\,] = \int \rho(\boldsymbol{r})v(\boldsymbol{r})\, d\boldsymbol{r} + T_s[\,\rho\,] + \frac{1}{2}\iint \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{r_{12}}\, d\boldsymbol{r}_1 d\boldsymbol{r}_2 + E_{xc}[\,\rho\,] \qquad \text{(2-51)}$$

The first three terms on the r.h.s of equation (2-51) are easy to calculate from electron density ρ. The last term E_{xc} is difficult to evaluate. The key to a successful KS DFT calculation is finding a good approximation to E_{xc}. Before we evaluate all terms in equation (2-51), the ground-state electron density need to be found. Recall that the fictitious reference system of non-interacting electrons has the same electron density as the ground state of the real system. Thus, the electron probability density of an n-electrons system whose wave function is a Slater determinant of the Kohn-Sham spin orbitals can be calculated by

$$\rho = \rho_s = \sum_{i=1}^{n} |\theta_i^{\text{KS}}|^2 \qquad \text{(2-52)}$$

The equation (2-51) can be further written as

$$\begin{aligned} E_0 = &-\sum_{\alpha} Z_\alpha \int \frac{\rho(\boldsymbol{r}_1)}{r_{1\alpha}} d\boldsymbol{r}_1 - \frac{1}{2}\sum_{i=1}^{n} \left\langle \theta_i^{\text{KS}}(1) \,|\, \nabla_1^2 \,|\, \theta_i^{\text{KS}}(1) \right\rangle \\ &+ \frac{1}{2}\iint \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{r_{12}}\, d\boldsymbol{r}_1 d\boldsymbol{r}_2 + E_{xc}[\,\rho\,] \end{aligned} \qquad \text{(2-53)}$$

With the equation (2-53), we can calculate the electronic energy from the Kohn-Sham spin orbitals if we know the expression of functional E_{xc}. The Hohenberg-Kohn variational theorem shows that one can obtain the electronic energy by varying electron density ρ so as to minimize the energy functional $E[\rho]$. Equivalently, rather than varying electron density ρ, we can vary the Kohn-Sham orbitals which determines ρ by equation (2-52). Just as one can prove that the orthonormal orbitals minimizing the Hartree-Fock expression for the electronic energy should satisfy the Fock equations (2-14), it can be proven [111] that the minimization condition for energy functional is equivalent to the set of orthonormal Kohn-Sham orbitals satisfying

$$\hat{h}^{\text{KS}}(1)\, \theta_i^{\text{KS}}(1) = \varepsilon_i^{\text{KS}}\, \theta_i^{\text{KS}}(1) \qquad \text{(2-54)}$$

where $\varepsilon_i^{\text{KS}}$ is the energy of the i-th Kohn-Sham orbital. The Kohn-Sham one-electron operator $\hat{h}^{\text{KS}}$ is defined as

$$\hat{h}^{\text{KS}}(1) = -\frac{1}{2}\nabla_1^2 - \sum_{\alpha} \frac{Z_\alpha}{r_{1\alpha}} + \int \frac{\rho(\boldsymbol{r}_2)}{r_{12}}\, d\boldsymbol{r}_2 + v_{xc} \qquad \text{(2-55)}$$

where v_{xc} is exchange-correlation potential, which is the functional derivative of the exchange-correlation energy:

$$v_{xc}(\boldsymbol{r}) = \frac{\delta E_{xc}[\,\rho(\boldsymbol{r})]}{\delta \rho(\boldsymbol{r})} \tag{2-56}$$

In practice, like the HF method, each Kohn-Sham spatial orbital is also written as a linear combination of a set of pre-defined basis functions (see section 2.2.4):

$$\theta_i^{\mathrm{KS}} = \sum_{\nu=1}^{b} c_{\nu i} \chi_\nu \tag{2-57}$$

In this case, one need to solve the equations

$$\sum_{\nu=1}^{b} c_{\nu i}(h_{\mu\nu}^{\mathrm{KS}} - \varepsilon_i^{\mathrm{KS}} S_{\mu\nu}) = 0, \qquad r = 1, 2, \ldots, b \tag{2-58}$$

$$h_{\mu\nu}^{\mathrm{KS}} = \left\langle \chi_\mu | \hat{h}^{\mathrm{KS}} | \chi_\nu \right\rangle \tag{2-59}$$

where $h_{\mu\nu}^{\mathrm{KS}}$ are elements of the Kohn-Sham matrix. KS DFT calculations are also performed by SCF procedure similar to the HF method mentioned above.

2.2.3.4 Local Density Approximation

The center topic of a DFT calculation is how to deal with the exchange-correlation functional. The functional E_{xc} is usually expressed as

$$E_{xc}[\,\rho\,] = \int \rho(\boldsymbol{r})\, \varepsilon_{xc}\,(\rho)\, d\boldsymbol{r} \tag{2-60}$$

where ε_{xc} is called energy density which depends on the electron density ρ. An early method to compute E_{xc} is called local density approximation (LDA), which assumes the value of energy density ε_{xc} at some point can be calculated exclusively from the value of the electron density ρ at that position. The next step is looking for a simple model system where we can compute ε_{xc} precisely. Jellium is such a hypothetical system, which is a model of interacting electrons in a space where positive charges are uniformly distributed. It is also known as homogeneous electron gas (HEG) or uniform electron gas (UEG). In this model, the physical significance of energy density is the exchange and correlation energy per electron in homogeneous electron gas with the electron density ρ. Energy density ε_{xc} can be further divided into exchange and correlation parts

$$\varepsilon_{xc}(\rho) = \varepsilon_x(\rho) + \varepsilon_c(\rho) \tag{2-61}$$

Based on the Jellium model, the exchange part of energy density can be calculated by [112, 113]

$$\varepsilon_x(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} (\rho(\boldsymbol{r}))^{\frac{1}{3}} \tag{2-62}$$

Therefore, the LDA exchange fuctional is written as

$$E_x^{\text{LDA}} = \int \rho\, \varepsilon_x \, d\boldsymbol{r} = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} \int [\rho(\boldsymbol{r})]^{\frac{4}{3}} \, d\boldsymbol{r} \tag{2-63}$$

However, even for the simple Jellium model, there is no analytical expression for the correlation energy density. Ceperley and Alder [114] used Quantum Monte Carlo method to compute total energy for homogeneous electron gases of several different electron densities with a very high numerical accuracy. By subtracting the exchange energy for each case, accurate correlation energies of these systems were determined. Vosko, Wilk and Nusair [115] fitted these values to obtain analytical interpretation formulas. The fitted VWN correlation energy density is written as

$$\begin{aligned}\varepsilon_c(r_s) = \frac{A}{2}\Bigg\{ &\ln \frac{r_s}{r_s + b\sqrt{r_s} + c} + \frac{2b}{\sqrt{4c - b^2}} \tan^{-1}\left(\frac{\sqrt{4c-b^2}}{2\sqrt{r_s}+b}\right) \\ &- \frac{b x_0}{x_0^2 + b x_0 + c}\left\{ \ln\left[\frac{(\sqrt{r_s} - x_0)^2}{r_s + b\sqrt{r_s} + c}\right] + \frac{2(b + 2x_0)}{\sqrt{4c-b^2}} \tan^{-1}\left(\frac{\sqrt{4c-b^2}}{2\sqrt{r_s}+b}\right)\right\}\Bigg\}\end{aligned} \tag{2-64}$$

where r_s is called effective radius, which is another expression of the electron density

$$r_s(\boldsymbol{r}) = \left(\frac{3}{4\pi\rho(\boldsymbol{r})}\right)^{\frac{1}{3}} \tag{2-65}$$

There is exactly one electron contained within the sphere defined by the effectively radius. A, x_0, b and c in equation (2-64) are empirical parameters. In this case, $A = 0.0621814$, $x_0 = -0.10498$, $b = 3.72744$ and $c = 12.9352$. The correlation functional with this set of parameters is usually called VWN5. There is another set of parameters published in the original paper [115], which is often called VWN3.

When both exchange and correlation energy densities are known, the exchange-correlation potential in equation (2-55) can be calculated by

$$v_{xc}^{\text{LDA}} = \frac{\delta E_{xc}^{\text{LDA}}}{\delta\rho} = \varepsilon_{xc}(\rho(\boldsymbol{r})) + \rho(\boldsymbol{r})\frac{\partial \varepsilon_{xc}(\rho)}{\partial\rho} \tag{2-66}$$

2.2.3.5 Local Spin-density Approximation

Local spin-density approximation (LSDA) [111] is a general form of LDA, which allows paired electrons to have different Kohn-Sham spatial orbitals. The spin densities as a function of position are typically expressed by the normalized spin polarization

$$\zeta(\boldsymbol{r}) = \frac{\rho_\alpha(\boldsymbol{r}) - \rho_\beta(\boldsymbol{r})}{\rho(\boldsymbol{r})} \tag{2-67}$$

where ρ_α and ρ_β are spin densities of α spin and β spin respectively. When ζ is 0, LSDA is reduced to LDA while all electrons have the same spin if ζ is 1. In LSDA, the energy density is not only the function of electron density ρ but also dependent on ζ. The exchange part of the LSDA energy density can be calculated by

$$\varepsilon_x[\rho(\boldsymbol{r}), \zeta] = -\frac{3}{8}\left(\frac{3\rho(\boldsymbol{r})}{\pi}\right)^{\frac{1}{3}} \left[(1+\zeta)^{\frac{4}{3}} + (1-\zeta)^{\frac{4}{3}}\right] \tag{2-68}$$

Then the exchange functional of LSDA is written as

$$E_x^{\text{LSDA}} = -\frac{3}{4}\left(\frac{6}{\pi}\right)^{\frac{1}{3}} \int \left[(\rho_\alpha)^{\frac{4}{3}} + (\rho_\beta)^{\frac{4}{3}}\right] d\boldsymbol{r} \tag{2-69}$$

Regarding the LSDA correlation, the mathematical form is very complicated. Equation (2-64) is just a special case for LDA. The general form of the VWN correlation [115] is

$$\varepsilon_c^{\text{VWN}}(r_s, \zeta) = \varepsilon_c(r_s, 0) + \varepsilon_a(r_s)\left[\frac{f(\zeta)}{f''(0)}\right](1-\zeta^4) + [\varepsilon_c(r_s, 1) - \varepsilon_c(r_s, 0)] f(\zeta)\zeta^4 \tag{2-70}$$

$$f(\zeta) = \frac{(1+\zeta)^{\frac{4}{3}} + (1-\zeta)^{\frac{4}{3}} - 2}{2(2^{\frac{1}{3}} - 1)} \tag{2-71}$$

where $\alpha(r_s)$ is the spin stiffness. $\varepsilon_c(r_s, 0)$, $\varepsilon_c(r_s, 1)$ and $\alpha(r_s)$ can all be calculated by equation (2-64). A, x_0, b and c for $\varepsilon_c(r_s, 1)$ are 0.0310907, -0.325, 7.06042 and 18.0578 respectively and parameters for $\alpha(r_s)$ are $-1/3\pi^{-2}$, -0.0047584, 1.13107 and 13.0045 respectively. We can see that when ζ is equal to 1, the equation (2-70) is also reduced to equation (2-64).

2.2.3.6 Generalized Gradient Approximation

LDA and LSDA functionals are developed based on the Jellium model, which is good for systems where electron density varies slowly over the space. Functionals of Generalized-gradient approximation (GGA) go beyond the LSDA functionals. A GGA functional is not only the function of the electron density at a position, but also the function of the gradient of the electron density at that position. Thus

$$E_{xc}^{\text{GGA}}\left[\rho_\alpha,\rho_\beta\right]=\int f\left(\rho_\alpha(\boldsymbol{r}),\rho_\beta(\boldsymbol{r}),\nabla\rho_\alpha(\boldsymbol{r}),\nabla\rho_\beta(\boldsymbol{r})\right)\,d\boldsymbol{r} \tag{2-72}$$

where $\nabla\rho_\alpha$ and $\nabla\rho_\beta$ stand for the gradients of α spin density and β spin density respectively. The LDA functional is usually called a local functional because it only involves the value of electron density at a point. The GGA functional is called semi-local because it involves the values of electron densities at a point and its infinitesimal neighborhood. Like other functionals, the GGA exchange-correlation functional can also be split into exchange and correlation parts

$$E_{xc}^{\text{GGA}}=E_x^{\text{GGA}}+E_c^{\text{GGA}} \tag{2-73}$$

One of the most popular GGA exchange functionals is Becke's 1988 functional, denoted B88 or B [116]. It adds a correction term into LSDA exchange functional (equation (2-69))

$$\begin{aligned}E_x^{\text{B88}}&=E_x^{\text{LSDA}}-b\sum_{\sigma=\alpha,\beta}\int\frac{(\rho_\sigma)^{\frac{4}{3}}x_\sigma^2}{1+6bx_\sigma\ln\left[x_\sigma+(x_\sigma^2+1)^{\frac{1}{2}}\right]}\,d\boldsymbol{r}\\&=E_x^{\text{LSDA}}+\Delta E_x^{\text{B88}}\end{aligned} \tag{2-74}$$

where

$$x_\sigma=\frac{|\nabla\rho_\sigma|}{(\rho_\sigma)^{\frac{4}{3}}} \tag{2-75}$$

Here, b is a fitted parameter, which is equal to 0.0042. ΔE_x is the B88 gradient correction term.

The popular GGA correlation functionals include the Perdew 1986 correlation (P86) functional [117], the Lee-Yang-Parr (LYP) functional [118] and the Perdew-Wang 1991 parameter-free correlation (PW91) functional [119]. Like B88 exchange functional, P86 correlation functional also adds a correction term to the LSDA correlation functional

$$\Delta E_c^{\text{P86}}=\frac{e^{-\Phi}\,C(r_s)\,|\nabla\rho|^2}{d\rho^{\frac{4}{3}}} \tag{2-76}$$

where

$$\Phi=1.745\,\frac{f\,C(0)\,|\nabla\rho|}{C(r_s)\rho^{\frac{7}{6}}} \tag{2-77}$$

$$C(r_s)=0.001667+\frac{0.002568+\alpha\,r_s+\beta\,r_s^2}{1+\gamma\,r_s+\delta\,r_s^2+10000\,\beta\,r_s^3} \tag{2-78}$$

$$d = \sqrt[3]{2}\sqrt{(\frac{1}{2}+\frac{1}{2}\zeta)^{\frac{5}{3}}+(\frac{1}{2}-\frac{1}{2}\zeta)^{\frac{5}{3}}} \tag{2-79}$$

Here, f, α, β, γ and δ are fitted parameters, which are 0.11, 0.023266, 7.389×10^{-6}, 8.723 and 0.472 respectively. Narrowly speaking, P86 refers only to the gradient correction term for the correlation functional and it can be combined with any LDA correlation functional. In most cases, P86 also stands for the VWN correlation functional plus the gradient correction term.

Contrasted to the P86 functional, LYP correlation functional computes the correlation energy *in toto* instead of correcting the LDA expression. The explicit form of the LYP correlation energy density is

$$\begin{aligned}
\varepsilon_c^{\mathrm{LYP}} &= -4a\frac{\rho_\alpha\rho_\beta}{\rho^2(1+d\rho^{-\frac{1}{3}})} - abw\Big\{\frac{\rho_\alpha\rho_\beta}{18}\Big[144(2^{\frac{2}{3}})\,C_F\,(\rho_\alpha^{\frac{8}{3}}+\rho_\beta^{\frac{8}{3}}) + (47-7\delta)|\nabla\rho|^2 \\
&\quad - (45-\delta)\left(|\nabla\rho_\alpha|^2+|\nabla\rho_\beta|^2\right) + 2\rho^{-1}(11-\delta)\left(\rho_\alpha|\nabla\rho_\alpha|^2+\rho_\beta|\nabla\rho_\beta|^2\right)\Big] \\
&\quad + \frac{2}{3}\rho^2\left(|\nabla\rho_\alpha|^2+|\nabla\rho_\beta|^2-|\nabla\rho|^2\right) - \left(\rho_\alpha^2|\nabla\rho_\beta|^2+\rho_\beta^2|\nabla\rho_\alpha|^2\right)\Big\} \\
C_F &= \frac{3}{10}3^{\frac{2}{3}}\left(\pi^2\right)^{\frac{2}{3}} \\
\omega &= \frac{e^{-cp^{-\frac{1}{3}}}}{\rho^{\frac{14}{3}}(1+d\rho^{-\frac{1}{3}})} \\
\delta &= cp^{-\frac{1}{3}} + \frac{d\rho^{-\frac{1}{3}}}{1+d\rho^{-\frac{1}{3}}}
\end{aligned} \tag{2-80}$$

where a, b, c and d are parameters fitted to the helium atom. They are 0.04918, 0.132, 0.2533 and 0.349 respectively.

The name of an exchange-correlation functional is, in most cases, simply the combination of its exchange and correlation parts. For example, BP86 functional is composed of B88 exchange functional and P86 correlation functional.

2.2.3.7 Hybrid Functionals

One big advantage of the HF method is that exchange interactions can be calculated exactly. In KS DFT, the exact exchange energy functional can be calculated by replacing the HF orbitals with the Kohn-Sham orbitals in the equation (2-13):

$$E_x^{\mathrm{exact}} = -\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\langle \theta_i^{\mathrm{KS}}(1)\,\theta_j^{\mathrm{KS}}(2)\left|\frac{1}{r_{12}}\right|\theta_j^{\mathrm{KS}}(1)\,\theta_i^{\mathrm{KS}}(2)\right\rangle \tag{2-81}$$

where the number of 1/4 comes from the factor that in equation (2-13) the summation is over the orbitals while in equation (2-81) the summation is over the electrons, which generates four times as many terms.

Since the exchange energy is the largest contributor to E_{xc}, it is a natural idea to use a functional which mixes the exact exchange functional together with the traditional exchange and correlation functionals. Such kind of functional is called hybrid functional. Becke [120] first developed a three-parameter hybrid functional B3PW91

$$E_{xc}^{\text{B3PW91}} = (1-a)E_x^{\text{LSDA}} + aE_x^{\text{exact}} + b\Delta E_x^{\text{B88}} + E_c^{\text{LSDA}} + c\Delta E_c^{\text{PW91}} \tag{2-82}$$

where a, b and c are optimized parameters, which are 0.20, 0.72 and 0.81 respectively. The name of the functional, B3PW91, stands for the scheme of three parameters and the use of B88 exchange functional and PW91 correlation functional.

One of the most widely used hybrid functional, B3LYP hybrid functional [121, 122], is defined similarly by

$$E_{xc}^{\text{B3LYP}} = 0.80E_x^{\text{LSDA}} + 0.20E_x^{\text{exact}} + 0.72\Delta E_x^{\text{B88}} + 0.19E_c^{\text{VWN}} + 0.81E_c^{\text{LYP}} \tag{2-83}$$

Since LYP functional computes the full correlation energy, the VWN correlation term is reduced to 19% in equation (2-83) instead of a 100% term in B3PW91 functional. The VWN correlation functional in equation (2-83) can be either VWN3 or VWN5. Different QM programs have different default setup for the VWN correlation part in B3LYP functional. For example, Gaussian uses VWN3 as default setup while Turbomole and Orca use VWN5 as default setup.

Generally speaking, the performance of the GGA functionals are better than that of the LDA functionals. Furthermore, hybrid functionals perform even better than the GGA functionals. Beyond hydrid functionals, there are double hybrid functionals such as B2PLYP [123] and PWPB95 [124]. Since these methods are not used in this dissertation, they are not discussed in this section.

2.2.3.8 Dispersion Correction

One major limitation of traditional DFT method is the poor performance in dealing with Van der Waals interactions between non-bonded atoms. Such interactions are very important in large biomolecules such as proteins which are the central topic in this dissertation. Currently, one of the most successful correction method is DFT-D3 developed by Grimme $et\ al.$ [125]. The number 3 refers to the third version of the DFT-D method. The idea of the approach is to simply add a correction term for dispersion E_{disp}

to the usual KS DFT energy:

$$E_{\text{DFT}-\text{D}} = D_{\text{DFT}} + E_{\text{disp}} \tag{2-84}$$

The simplest form of the correction term is

$$E_{\text{disp}} = -\sum_{\text{AB}} \frac{C_{\text{AB}}}{R_{\text{AB}}^{6}} f(R_{\text{AB}}) \tag{2-85}$$

where the summation is over all pairs of non-bonded atoms, C_{AB} is a constant which is calculated theoretically. R_{AB} is the distance between atoms A and B and f is a damping function which makes the corrected energy go to 0 when R_{AB} approaches to 0. In practice, more sophisticated forms of E_{disp} are used. DFT-D3 has three empirically determined parameters in E_{disp}. The values of the three parameters are dependent on which functional is used in calculation.

2.2.4 Basis Set

In sections 2.2.1 and 2.2.3 , HF and DFT methods are discussed. In principle, with the help of the variation theorem (equations (2-7 and 2-44)), one can try every possible trial function to obtain the ground-state electronic energy of a system. However, this is not feasible in practice. In most cases, people use the linear combination of a set of finite functions as a trial set for both *ab initio* and KS DFT calculations. A basis set is composed of these predefined functions (called basis functions). There are mainly two types of basis functions used in electronic structure calculations: Slater-type Orbitals (STO) and Gaussian-type Orbitals (GTO). The basis functions are also called Atomic Orbitals (AO), although they are usually not solutions to an atomic Schrödinger equation.

2.2.4.1 Slater and Gaussian Type Orbitals

The mathematical form of a normalized Slate-type orbital in atom-centered polar coordinate is

$$\varphi_{nlm}(r,\theta,\phi) = \frac{(2\zeta)^{n+1/2}}{[(2n)!]^{n+1/2}} r^{n-1} e^{-\zeta r} Y_l^m(\theta,\phi) \tag{2-86}$$

where ζ is called the orbital exponent, n, l and m are the principal quantum number, the angular momentum quantum number and the magnetic quantum number respectively. Y_l^m is the spherical harmonic functions. The major drawback of using STOs as basis functions is that there is no analytical expression for the three-center or four-center two-electron integrals. Therefore, these integrals must be evaluated numerically and this is very time consuming.

In order to avoid this problem, one can use Gaussian-type functions (GTFs) for atomic orbitals. The mathematical form of a normalized Gaussian-type orbital in atom-centered Cartesian coordinates is

$$\phi_{ijk}\left(x,y,z\right)=\left(\frac{2\alpha}{\pi}\right)^{3/4}\left[\frac{(8\alpha)^{i+j+k}\,i!\,j!\,k!}{(2i)!\,(2j)!\,(2k)!}\right]^{1/2}x^i\,y^j\,z^k e^{-\alpha(x^2+y^2+z^2)} \qquad \text{(2-87)}$$

where α is a positive orbital exponent which controls the width of the GTO and i, j and k are non-negative integers indicating the nature of the orbital. When $i+j+k=0$, the Gaussian-type function is called an s-type Gaussian. When $i+j+k=1$, it is a p-type Gaussian, which has three possibilities corresponding to p_x, p_y and p_z orbitals. When $i+j+k=2$, it is a d-type Gaussian. In Cartesian coordinates, there are six possible d-type Gaussians with prefactors of x^2, y^2, z^2, xy, xz and yz. By linear combinations, these six functions can be transformed to five spherical d-functions and an additional s-function $(x^2+y^2+z^2)$. The last linear combination of s-type function is omitted in some basis set. The GTO does not have the cusp at the nucleus and hence has a problem of representing proper behavior near the nucleus. Another problem of GTO is that it decays too rapidly far from the nucleus and thus the "tail" of the function is represented poorly. One way to solve these problems is using a linear combination of several Gaussians to accurately represent an AO:

$$\chi_r=\sum_u d_{ur}\phi_u \qquad \text{(2-88)}$$

where ϕ_u are normalized Gaussians centered on the same atom and having the same prefactors but with different orbital components. The contraction coefficients are held constant during the calculation. χ_r is called contracted Gaussian-type orbital (CGTO) and ϕ_u are called primitive Gaussians. The number u is often called the degree of contraction.

2.2.4.2 Classification of Basis Set

A minimal (or minimum) basis set refers to one basis function for each core and valence AOs of each atom. The basis functions can be either STOs or CGTOs. The next step to improve the basis set is increasing the number of basis functions for each atomic orbital. A double-zeta (DZ) basis set refers to two basis functions for each AO. Some DZ basis sets only double the number of basis functions for the valence orbitals. These basis sets are called valence double-zeta (VDZ) basis set or split-valence basis set (can be more than two basis functions for one valence orbital). Similarly, a triple-zeta (TZ) basis set uses three basis functions for each AO.

AOs are sometimes distorted in shape and charges are shifted upon molecule formation. To take this polarization effect into account, one can add basis functions whose angular momentum quantum numbers (l) are higher than the maximum l of the ground-state valence orbitals of an atom. Such a basis set is a polarized basis set. For example, a p-type function is added as a polarization function for a hydrogen atom and a d-type function is added as a polarization function for a carbon atom.

Some anions, compounds with lone pairs, hydrogen-bonded dimers and loose supermolecular complexes tend to have much more spatially diffuse electron density at large distances from the nucleus. If a basis set has no flexibility to allow a weakly bound electron to appear far from the remaining electron density, significant errors in energy or other properties can occur. To address this problem, some basis functions, called diffusion function, with small orbital components are added into basis set.

2.2.4.3 Several Contracted Basis Set

One of the most widely used basis sets are Pople style basis sets. The nomenclature of this type of basis sets is a direct guide to the contraction scheme. The first number represents the degree of contraction for the core functions. The numbers after hyphen stands for the numbers of the primitive GTOs used in the valence functions. If there are two numbers after hyphen, it is a valence double-zeta basis set. If there are three numbers after hyphen, it is a valence triple-zeta basis set. For example, 6-311G [126, 127] means each core orbital is contracted by 6 primitive GTOs. The valence orbital is split into three functions and they are contracted by 3, 1 and 1 GTO(s) respectively. Besides, diffusion functions are denoted by + or ++. The first + indicates a set of diffusion s- and p-functions on heavy atoms and the second + indicates a diffusion s-function is also added to hydrogen atom. Finally, polarization function is denoted by letters in parentheses or * after G. For example, 6-31G(d) [128–130], which is identical to 6-31G*, has one d-type polarization function added on heavy atoms. And 6-31G(d,p) [128–130], which is equivalent to 6-31G**, further adds one p-type polarization function on hydrogen atoms.

Another type of basis set was developed by Ahlrichs and coworkers [131, 132], which is also called Karlsruhe basis set. The latest version of this type of basis set has a prefix of “def2” [133, 134] which denotes the second generation default basis set in the QM program Turbomole [135]. The def2-SVP basis set [133] refers to the split valence polarized basis set. For the hydrogen atom, the valence orbital is composed of two functions which are contracted by 3 and 1 s-type GTO(s) respectively. One p-type function is added as a polarized function. Moreover, for the first row elements, the core s-orbital is contracted by 5 primitive GTOs. The valence s-orbital consists of two s-type GTO(s) and the valence p-orbital is composed of two functions which are contracted by 3 and 1 p-type GTO(s) respectively. One d-type function is added as polarized function. The def2-TZVP basis set [133] stands for the triple zeta valence polarized basis set. For the hydrogen atom, the valence orbital is made from three functions which are contracted by 3, 1 and 1 s-type GTO(s) respectively. One p-type function is added as a polarized function. In addition, for the first row elements, the core s-orbital is contracted by 6 primitive GTOs. The valence s-orbital consists of four functions which are contracted by 2, 1, 1 and 1 s-type GTO(s) respectively and the valence p-orbital is composed of three functions contracted by 4, 1 and 1 p-type GTO(s) respectively. Two d-type functions and one f-type function are added as polarized functions.

Correlation consistent basis sets were developed by Dunning and coworkers [136–140]. The basis sets were designed to recover the correlation energy for the valence electrons. The exponents and contraction coefficients were optimized not only for the HF method, but also for electron correlation methods. The nomenclature of this type of basis sets is fairly simple. For example, cc-PVQZ [138] refers to Correlation Consistent Polarized Valence Quadruple Zeta basis set.

2.3 Molecular Mechanics

In contrast to the Quantum mechanics methods described in previous sections, molecular mechanics uses classical mechanics to calculate properties of molecular systems. It is a useful method to study the system of biological macromolecules. The Born-Oppenheimer approximation still holds and the potential energy of the system is still a function of molecular geometries.

2.3.1 Force Field Methods

Force field approach is an ideal method to calculate potential function for a huge molecular system. In this method, potential energy of a system can be written as a parametric function of the nuclear coordinates, obtained from fitting the parameters to experimental or higher level computational data. For organic systems, such a force field is commonly composed of different contributions:

$$V(\boldsymbol{R}) = V_{str} + V_{bend} + V_{tor} + V_{vdw} + V_{el} + V_{cross} \tag{2-89}$$

Here V_{str} is the energy function for stretching a bond between two atoms, V_{bend} represents the energy required for bending an angle, V_{tor} is the torsion energy for rotation around a bond, V_{vdw} and V_{el} describe the non-bonding atom-atom interactions, and finally V_{cross} describes coupling between the first three terms. The building blocks in force field methods are atoms, this means that bonding information must be provided explicitly, rather than being the result of solving the electronic Schrödinger equation [141].

The potential energy function for bond stretching is given by

$$V_{str} = \sum_{\text{bonds}} k(r - r_{eq})^2 \tag{2-90}$$

where k is the force constant, r is the distance between two bonded atoms and r_{eq} is a parameter standing for the equilibrium bond length. Similarly, the potential energy function for angle bending is

$$V_{bend} = \sum_{\text{angles}} k(\theta - \theta_{eq})^2 \tag{2-91}$$

where k, θ and θ_{eq} are the force constant, actual bond angle and equilibrium bond angle respectively. The torsional potential energy can be written as

$$V_{tor} = \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \tag{2-92}$$

where V_n is the relative potential barrier, n is the periodicity, ϕ is the dihedral angle and γ is the phase shift.

The non-bonding interactions can be modeled by pair-wise Lennard-Jones and Coulomb potentials. The formulas of the potentials between two atoms are given by

$$V_{LJ}(r_{ij}) = \frac{C_{12}(i,j)}{r_{ij}^{12}} - \frac{C_6(i,j)}{r_{ij}^{6}} \tag{2-93}$$

$$V_C(r_{ij}) = \frac{1}{4\pi\varepsilon_0} \times \frac{q_i q_j}{\varepsilon_r r_{ij}} \tag{2-94}$$

where r_{ij} is the distance between two atoms. C_{12} and C_6 are constants depending on atom type. q_i and q_j are point charges and ε_r is the relative permittivity. While the Coulomb potential describes electrostatic interactions, the Lennard-Jones potential includes van der Waals and repulsive contributions.

However, one has to realize that there are some limits of force field method. One large drawback of this method is that it cannot treat systems involving formation or breaking of covalent bonds. This means that the force field representation of equation (2-89) can only be applied to systems where chemical reactions do not play a significant role. If this is not the case, the QM/MM method (see section 2.4) is a promising method to simulate the system.

2.3.2 Amber Force Field

Amber force field is the only force field used in this dissertation. It is a family of force fields originally implemented in Amber software suit [142, 143]. The ff99SB [144] and ff03.r1 [145, 146] Amber force fields were used to simulate ClpP and HCV systems respectively. Both force fields are modified versions of ff99 [147], which is a classical force field for general organic and biomolecular systems. Its partial charges were fitted at the HF/6-31G(d) level by using RESP charge fitting scheme [148]. Compared to the ff99 force field, the ff99SB force field [144] modified parameters of backbone dihedrals, which were fitted based on the gas-phase quantum data. In the ff03 force field, partial charges are derived from quantum calculations which use a continuum dielectric to imitate the polarization of solvent [145, 146]. The backbone torsions ϕ and ψ for proteins are modified and the preference for helical configurations is decreased. The ff03.r1 force field is a newer version of the ff03. It has new libraries for the N- and C-terminal amino acids. The ff14SB force field is a new type of Amber force fields which further modify the dihedral parameters for both backbone and side chains to improve the performance of reproducing experimentally indicating geometries [149].

The general amber force field (GAFF) [150] is usually used for organic molecules and it is designed to be compatible with the amber force field. The charge methods used in GAFF is either HF/6-31G(d) RESP charge or AM1-BCC [151, 152]. In this work, RESP charges for ligands were calculated at HF/6-31G(d) level by Gaussian09 [153].

2.3.3 Water Model and Ions Parameters

There are a lot of water models to describe water molecules as explicit solvent. TIP3P [154] and SPC/E [155] are two models used in this work. TIP3P [154] is a rigid 3-site water model. The bond distances and HOH bond angle are fixed. Partial charges are assigned to three atom sites. But only the site for the oxygen atom has the Lennard-Jones parameters.

SPC/E [155] is also a rigid 3-site model and has a different set of parameters in comparison to the TIP3P method. The HOH angle is set to an ideal tetrahedral shape (109.47°) instead of the experimentally observed angle of 104.5°. In addition, The SPC/E model also adds an average polarization energy, which is a constant correction of 1.25 kcal/mol to the potential energy function.

Joung/Cheatham monovalent ion parameters [156, 157] are used in this dissertation. These consistent parameters for alkali and halide ions were separately fitted to several experimental values for different water models.

2.4 QM/MM Method

2.4.1 QM/MM Energy

When the inner part of the system is treated by QM method and the outer part of the system is dealt with MM method, subtractive QM/MM energy expression is written as [158]

$$E^{\text{sub}}_{\text{QM/MM}} = E_{\text{MM}}(\text{total}) + E_{\text{QM}}(\text{I}) - E_{\text{MM}}(\text{I}) \tag{2-95}$$

where $E_{\text{QM/MM}}$ is the total electronic energy of the system, E_{MM}(total) is the energy of the entire system calculated by MM method, E_{QM}(I) is the energy of the inner part of the system at QM level and E_{MM}(I) is the energy of the inner part of the system calculated by MM method. Equation (2-95) also holds for the scheme of the link-atoms (see section 2.4.3.1). In this case, the last two terms are calculated on a capped inner system.

The advantage of the subtractive QM/MM scheme is that there are no explicit QM/MM coupling terms. The standard QM and MM calculations can be performed without any modifications. The drawback of a subtractive scheme is that a complete set of MM parameters is required for the inner subsystem, which can be difficult or even impossible to obtain. A more important limitation is that the QM/MM coupling is treated entirely at the MM level, especially for the electrostatic interactions (see section 2.4.2).

The energy expression for an additive QM/MM scheme is written as [158]

$$E^{\text{add}}_{\text{QM/MM}} = E_{\text{MM}}(\text{O}) + E_{\text{QM}}(\text{I}) + E_{\text{QM}-\text{MM}} \tag{2-96}$$

In contrast to the subtractive scheme, the MM calculation is only performed on the outer part of the system. In addition, the last term in equation (2-96) is an explicit QM/MM coupling term, which represents the interaction terms between the QM and the MM parts. QM calculation is performed on the capped inner subsystem as in the subtractive scheme. Currently, most QM/MM calculations use the additive scheme. In general, the coupling term includes bonded, electrostatic and van der Waals interactions between QM and MM atoms [158]

$$E_{\text{QM}-\text{MM}} = E^{\text{b}}_{\text{QM}-\text{MM}} + E^{\text{vdW}}_{\text{QM}-\text{MM}} + E^{\text{el}}_{\text{QM}-\text{MM}} \tag{2-97}$$

QM–MM bonded and van der Waals interactions are commonly treated at the MM level. Whereas the embedding scheme decides whether QM–MM electrostatic interactions are handled at the QM level or the MM level.

2.4.2 Embedding Mechanism

One of the most important technical detail of a QM/MM calculation is how to deal with the QM/MM electrostatic coupling. According to the different handling of the electrostatic interactions between QM and MM atoms, the QM/MM coupling schemes can be classified [159, 160] into mechanical embedding, electrostatic embedding and polarized embedding.

The most basis scheme is mechanical embedding, where the QM–MM electrostatic interaction is purely treated at the MM level. The charge model of the MM atoms (typically point charges) is also applied to the QM region. However, there are some major disadvantages [158] 1) The QM region is not directly polarized by the electrostatic environment of the MM region. 2) When the electron density of the QM part changes, for example bond forming or breaking, the point charges might be required to update. 3) The point charges in the QM part might not reproduce the true charge distribution of the QM region. Unlike the force field method, there is no overall balanced description of point charges in QM subsystem.

The electrostatic embedding can overcome the shortcomings of the mechanical embedding. In the case of electrostatic embedding, MM point charges are included as one-electron terms in the Hamiltonian of the QM region

$$\hat{H}^{el}_{\mathrm{QM-MM}} = -\sum_{i}^{N}\sum_{J}^{L}\frac{q_J}{|\boldsymbol{r}_i - \boldsymbol{R}_J|} + \sum_{\alpha}^{M}\sum_{J}^{L}\frac{q_J Q_\alpha}{|\boldsymbol{R}_\alpha - \boldsymbol{R}_J|} \tag{2-98}$$

where q_J are the point charges of MM atoms located at $\boldsymbol{R}_J$, Q_a are the charges of the QM nuclei at $\boldsymbol{R}_a$, and $\boldsymbol{r}_i$ denote electron positions. The N, L and M are the number of electrons, point charges and QM nuclei respectively. In an electrostatic embedding scheme, the electronic structure of the QM region will adapt to changes in the distribution of MM point charges and is polarized by the environment. The QM–MM electrostatic interaction is now handled at the QM level. Thus, it provides a more precise description than a mechanical embedding scheme. Special attention is needed at the QM–MM boundary, where MM point charges are placed in close vicinity to the QM electron density and might cause overpolarization. This problem is prominent when the boundary cut a covalent bond and it can be solved by shifting away the charges from MM boundary atoms (see section 2.4.3.1). Nowadays, electrostatic embedding is still the most popular embedding scheme, especially for computational biochemistry applications.

Polarized embedding is more complicated than other two schemes mentioned above. In this case, the MM point charge distribution is in turn polarized by QM region. In general, this requires a polarized force field for the MM part. However, there has been no well-established polarized force field yet. Thus, this scheme is currently not widely used.

2.4.3 Boundary Treatment

Sometimes, the cutting of the QM–MM boundary through a bond is unavoidable. There are several approaches to deal with covalent bonds cut by the QM–MM boundary. Two major methods, link atoms and frozen localized orbitals, will be introduced in this section.

2.4.3.1 Link Atoms

The link-atoms method is conceptually straightforward: An additional atom is placed between QM and MM boundary atoms (Q_1 and M_1) which is covalently bonded to Q_1 (Figure 2-1). In most cases, the link atom is a hydrogen atom. But any monovalent atom is also possible. QM calculations are then performed on the inner subsystem and the link atom(s). The bond between QM and MM boundary atoms (Q_1–M_1) is treated at the MM level. The introduction of an additional atom, which does not belong to the real system, creates several problems that should be addressed [158]. Firstly, each link atom introduces three additional structural degrees of freedom. These extra degrees of freedom can be eliminated by using constraints. The link atom is placed along Q_1–M_1, and the distance Q_1–L is derived from the bond distance of Q_1–M_1 by a scaling factor [161–168]. Therefore, exactly three degrees of freedom are removed. Currently most link-atoms schemes follow this approach. Secondly, the link atom is so close to the MM boundary atom M_1 that the point charge on the atom tend to overpolarize the electron density of QM region when electrostatic or polarized embedding is used. Several approaches were proposed to reduce the problem of overpolarization: 1) Deletion of the one-electron integrals related to the link atoms [159, 160, 167, 169–171]. 2) Deletion of point charges in the boundary region from the Hamiltonian [160, 161, 172–179]. 3) Shifting the point charges in the link region (Figure 2-1) [165, 179–183]. In order to preserve the charge and dipole of the MM region, point dipoles are added to compensate. This scheme cures the major deficiencies of charge-deletion schemes [179, 183, 184]. 4) Replacing the point charges close to the QM region by charge distributions [166, 184, 185]. Lastly, the chemical and electronic properties of link atoms are generally different from the replaced groups. Despite these problems, the link-atoms method is still widely used nowadays.

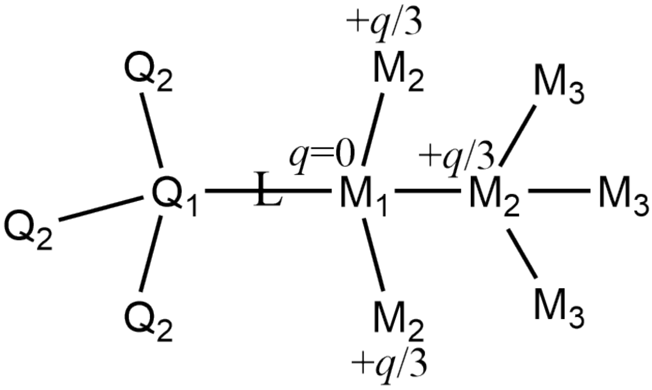


Figure 2-1. Labeling of atoms at the boundary region. Q_1, L and M_1 are the QM boundary atom, the link atom the MM boundary atom respectively. Q_2 are QM atoms bonded to Q_1 while M_2 are MM atoms bonded to M_1. M_3 refer to MM atoms seperated from M_1 by two bonds. The partial charges near M_1 and M_2 atoms illustrate a charge-shift scheme: The original partial charge q of M_1 is removed and evenly distributed onto the M_2 atoms. Additional pairs of charges are placed near the M_2 atoms to compensate the original M_1–M_2 dipoles (modified from ref. [158]).

2.4.3.2 Frozen Localized Orbitals

In the original paper of Warshel and Levitt [71], they already applied the frozen orbital method that uses frozen hybrid orbitals to saturate the free valences at the QM–MM boundary. Different schemes share the same idea of placing suitably oriented localized orbitals on the frontier atoms and keeping part of these orbitals frozen so that they are excluded in the SCF iterations. The two major schemes of frozen orbital method are local self-consistent field (LSCF) and generalized hybrid orbitals (GHO). The LSCF method was developed by Rivail and co-workers [186–190]. First a QM calculation on a model system which contains the Q_1–M_1 bond is performed. Then a strictly localized bond orbital for this bond is constructed. This orbital only has contributions from the boundary atoms. In the QM/MM calculation, the orbital does not participate in the SCF iterations and therefore not mix with other orbitals. It is oriented along the direction of Q_1–M_1 bond and can be considered as a lone pair on Q_1 which points towards M_1. The GHO method was developed by Gao and co-workers [191–196], which also constructs localized hybrid orbitals and keeps some of them frozen. However, the set of localized hybrid orbitals is placed on M_1 instead of Q_1. Thus M_1 becomes a boundary atom. The orbital pointing towards Q_1 is included in the SCF iterations, while the remaining orbitals on M_1 are kept frozen and are not mixed with the other orbitals.

2.5 Energy Minimization

When the potential function of a molecular system, which can be calculated by methods mentioned above, is known, one can perform the energy minimization to bring the system towards the nearest energy minimum. There are many approaches to find a local minimum of the system on potential energy surface.

2.5.1 Steepest Descent

The idea of steepest descent method is quite simple. A multivariable function decreases fastest in the direction of negative gradient. The new positions are given by

$$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \gamma \boldsymbol{g}_i \tag{2-99}$$

where γ is step width parameter and $\boldsymbol{g}_i$ is the gradient of the i-th step. Note that steepest descent algorithm is robust and easy to implement. But it is not the most efficient way to search local minimum because the convergence rate slows down when the minimum is near.

2.5.2 Conjugate Gradient

In contrast to the steepest descent method, each search step in conjugate gradient method [197] depends on not only the current gradient but also the previous search direction. Thus, the new positions are determined by

$$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \gamma \boldsymbol{d}_i \tag{2-100}$$

where $\boldsymbol{d}_i$ is the search direction of the i-th step, which is a mixture of negative gradient and search direction of the previous step.

$$\boldsymbol{d}_i = -\boldsymbol{g}_i + \beta_i \boldsymbol{d}_{i-1} \tag{2-101}$$

The quantity of β can be calculated by several ways naming after their developers: Fletcher-Reeves (FR) [198], Polak-Ribiere (PR) [199] and Hestenes-Stiefel (HS) [197].

The formulas for β are

$$\begin{aligned}\beta_i^{\mathrm{FR}} &= \frac{|\boldsymbol{g}_i|^2}{|\boldsymbol{g}_{i-1}|^2}\\ \beta_i^{\mathrm{PR}} &= \frac{\boldsymbol{g}_i(\boldsymbol{g}_i-\boldsymbol{g}_{i-1})}{|\boldsymbol{g}_{i-1}|^2}\\ \beta_i^{\mathrm{HS}} &= \frac{\boldsymbol{g}_i(\boldsymbol{g}_i-\boldsymbol{g}_{i-1})}{\boldsymbol{d}_{i-1}(\boldsymbol{g}_i-\boldsymbol{g}_{i-1})}\end{aligned} \tag{2-102}$$

2.5.3 L-BFGS

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [200–203], naming after its developers, is a quasi-Newton method for solving optimization problems. In quasi-Newton methods, the searching direction is the product of the inverse of the Hessian matrix and the negative gradient, which is a better estimation compared to the methods described above. In the BFGS method, the searching direction is given by

$$\boldsymbol{d}_i = -H_i\boldsymbol{g}_i \tag{2-103}$$

where H_i is an estimate of the inverse of the Hessian matrix. The recursive formula for the matrix is

$$H_{i+1} = \left(I-\rho_i\boldsymbol{s}_i\boldsymbol{y}_i^{\mathrm{T}}\right)H_i\left(I-\rho_i\boldsymbol{y}_i\boldsymbol{s}_i^{\mathrm{T}}\right)+\rho_i\boldsymbol{s}_i\boldsymbol{s}_i^{\mathrm{T}} \tag{2-104}$$

where

$$\rho_i = \frac{1}{\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{s}_i} \tag{2-105}$$

$$\boldsymbol{s}_i = \boldsymbol{r}_{i+1}-\boldsymbol{r}_i \tag{2-106}$$

$$\boldsymbol{y}_i = \boldsymbol{g}_{i+1}-\boldsymbol{g}_i \tag{2-107}$$

Limited-memory BFGS (L-BFGS) [204] is a variant of the BFGS and it is a default algorithm implemented in dl_ find [205]. In the L-BFGS, the searching direction is computed by two recursive steps to avoid calculations of the matrix. Thus, two sequences of $\boldsymbol{s}_i$ and $\boldsymbol{g}_i$ for the last several steps are stored and only a limited amount of computer memory is required. Due to its linear memory requirement, the L-BFGS method performs particularly well for optimization problems with a large amount of variables

2.6 Molecular Dynamics

2.6.1 Basic Concepts of Molecular Dynamics

Molecular Dynamics (MD) simulation is a technique for computing the equilibrium and evolution properties of a classical many-body system. Here the word classical implies that the motion of the particles obeys the rules of classical mechanics and the quantum effects are neglected [206]. MD calculates the trajectory of the particles within the system by numerically solving Newton's equations of motion.

To measure an observable quantity in MD simulation, this observable must be a function of positions and momenta of the particles in the system. For example, the temperature of a many-body system is related to the average kinetic energy per degree of freedom:

$$\frac{1}{2}k_B T = \left\langle \frac{1}{2} m_i\, v_{\alpha,i}^2 \right\rangle \tag{2-108}$$

where k_B is Boltzmann's constant and $v_{\alpha,i}$ is the component of the velocity of a given particle.

A typical MD calculation includes: 1) Initialization of the system, 2) Computing forces on all atoms, 3) Integration of Newton's equation of motion, 4) Computing the averages of measured quantities.

2.6.1.1 Initialization

Before the running of a MD simulation, initial positions and velocities of all particles in the system should be assigned. In general, velocities are generated randomly by Maxwell-Boltzmann distribution at a given absolute temperature T [207]

$$p(v_{\alpha,i}) = \sqrt{\frac{m_i}{2\pi k_B T}}\, \exp\left(-\frac{m_i v_{\alpha,i}^2}{2k_B T}\right) \tag{2-109}$$

To achieve this, normally distributed random numbers are generated. Then the result is multiplied by $(k_B T/m_i)^{1/2}$ which is the standard deviation of the velocity distribution. Since the resulting average kinetic energy will not correspond exactly to the required temperature T, an instantaneous temperature $T(t)$ at time t can be defined as follows:

$$k_B T(t) = \sum_{i=1}^{N} \sum_{\alpha} \frac{m_i v_{\alpha,i}^2(t)}{N_f} \tag{2-110}$$

where i is the index for particles and N_f is the degree of freedom ($= 3N - 3$ for the system with fixed total momentum). Now a correction can be made: the motion of center-of-mass is removed and then all velocities are scaled with a factor of $(T/T(t))^{1/2}$ so that the initial total energy corresponds exactly to T.

2.6.1.2 Computing Forces

The calculation of the force is usually the most time-consuming part for a MD simulation. According to classical mechanics, the force on one atom can be expressed as the derivative of the potential energy with respect to the position of the atom:

$$\boldsymbol{F}_i = -\frac{\partial V(\boldsymbol{R})}{\boldsymbol{r}_i} \tag{2-111}$$

where $\boldsymbol{R} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n)$. When the explicit form of the potential function $V(\boldsymbol{B})$ is known, the force on each atom can be calculated based on equation (2-111). This means determining the potential energy surface of the system is the key to calculating the forces. All computational techniques mentioned above (see sections 2.2 and 2.3) can be used to obtain the function of potential energy.

2.6.1.3 Integration of Equations of Motion

After calculating the forces, the time evolution of a system can be calculated based on classical Newton's equations of motion:

$$\dot{\boldsymbol{r}}_i = \boldsymbol{v}_i \tag{2-112}$$

$$\dot{\boldsymbol{v}}_i = \frac{\boldsymbol{F}_i}{m_i} \tag{2-113}$$

There are many methods to integrate equations (2-112) and (2-113) numerically. A classical integrator is leap-frog [208]. It is based on a discrete time step Δt, and uses the position at time t and velocities at time $t - 1/2\Delta t$. Positions and velocities are updated by using the following equations:

$$\boldsymbol{v}(t + \frac{1}{2}\Delta t) = \boldsymbol{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\boldsymbol{F}(t) \tag{2-114}$$

$$\boldsymbol{r}(t + \Delta t) = \boldsymbol{r}(t) + \Delta t \boldsymbol{v}(t + \frac{1}{2}\Delta t) \tag{2-115}$$

The updated positions and velocities then serve as starting point for the next integration step. This way a sequence of snapshots of the system is created, which represents the trajectory of the system in phase space. The accuracy of the leap-frog method is of the order Δt^4 in positions and this algorithm is time reversible [206]. Note that the equations of motion will be changed if temperature coupling or pressure coupling is applied (see details in sections 2.6.2 and 2.6.3).

2.6.1.4 Average Calculation and Error Estimation

After completion of the simulation, the averages of measured quantities are calculated and the statistical error can be estimated according to the standard deviation of all data points sampled:

$$\sigma_{\bar{A}} = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}(A_i - \bar{A})^2} \tag{2-116}$$

where N is the sample size and $\bar{A}$ is the mean value of a measured quantity. This estimation is based on the assumption that each data point is independent.

2.6.2 Temperature Coupling

The molecular dynamics simulation according to equations (2-114) and (2-115) leads to the microcanonical ensemble which is also called NVE (constant number of particles, constant volume and constant energy) ensemble. However, many quantities of interest are obtained in other important ensembles such as NVT (constant number of particles, constant volume and constant temperature) and NPT (constant number of particles, constant pressure and constant temperature) ensembles. The former is called canonical ensemble and the latter isothermal-isobaric ensemble. Two commonly used thermostats, Berendsen and Nosé-Hoover thermostats are discussed here.

2.6.2.1 Berendsen Thermostat

The method of weakly coupling a system to an external bath was proposed by Berendsen *et al.* [209]. The velocities of all particles in the system are scaled every step by a factor λ

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)} \tag{2-117}$$

where T_0 is the reference temperature, T the instantaneous temperature, Δt the time step of MD simulation and τ_T the time constant of the thermostat. Furthermore, it can be proven that the deviation of the instantaneous temperature from the reference temperature is corrected based on the formula

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau_T} \tag{2-118}$$

Equation (2-118) indicates that the temperature difference decays exponentially with a time constant τ_T. Although this is a quite efficient method to couple a system to the target temperature, it is noticeable, however, that the Berendsen thermostat does not generate an exact canonical ensemble because the distribution of velocities is not reproduced correctly in the Berendsen thermostat.

2.6.2.2 Nosé-Hoover Thermostat

The Nosé-Hoover approach was first proposed by Nosé [210] and then modified by Hoover [211]. Compared to Berendsen weak-coupling method, a Nosé-Hoover thermostat can generate the correct canonical ensemble. The following equations of motion are used in the thermostat:

$$\dot{\boldsymbol{r}}_i = \frac{\boldsymbol{p}_i}{m_i} \tag{2-119}$$

$$\dot{\boldsymbol{p}}_i = \boldsymbol{F}_i - \frac{p_\xi}{Q}\boldsymbol{p}_i \tag{2-120}$$

$$\dot{\xi} = \frac{p_\xi}{Q} \tag{2-121}$$

$$\dot{p}_\xi = \sum_i \frac{\boldsymbol{p}_i^2}{m_i} - N_f k_B T \tag{2-122}$$

Here ξ is thermodynamic friction coefficient and p_ξ is associated momentum. Q is a constant called mass parameter and N_f is total number of degrees of freedom. Note that instead of the total energy which is conserved in conventional MD simulation, the conserved quantity for Nosé-Hoover thermostat is

$$H_{\text{Nosé-Hoover}} = H(r,p) + \frac{\xi^2}{2Q} + N_f k_B T \xi \tag{2-123}$$

Although the Nosé-Hoover thermostat generates true NVT ensemble, it causes an oscillatory relaxation. This leads to a relaxation time being several times longer compared to the Berendsen method.

2.6.3 Pressure Coupling

Most experiments are conducted at constant pressure rather than constant volume. Thus in some simulations it is desired to control the pressure rather than the volume. In the same spirit as the temperature coupling discussed before, a system can also be coupled to an external bath with constant pressure.

2.6.3.1 Berendsen Barostat

The difference between Berendsen thermostat and barostat is that the temperature is scaled in the thermostat while coordinates and box vectors are scaled by the barostat

[209]. The deviation of pressure from the reference pressure decays exponentially like the deviation of temperature in Berendsen thermostat:

$$\frac{d\boldsymbol{P}}{dt} = \frac{\boldsymbol{P}_0 - \boldsymbol{P}}{\tau_p} \tag{2-124}$$

where $\boldsymbol{P}_0$ is the reference pressure, $\boldsymbol{P}$ the instantaneous pressure, Δt the time step of the MD simulation and τ_p time constant of the barostat. The scaling matrix $\boldsymbol{\mu}$ which is used to scale box vectors and coordinates every step can be expressed as

$$\boldsymbol{\mu} = \boldsymbol{E} - \frac{\beta \Delta t}{3\tau_p}(\boldsymbol{P}_0 - \boldsymbol{P}) \tag{2-125}$$

where $\boldsymbol{E}$ is identity matrix and β is the estimated compressibility of the system. For water the value of compressibility is 4.5×10^{-10} Pa^{-1} at 1 atm and 300K.

2.6.3.2 Parrinello-Rahman Barostat

Similar to the previously mentioned Nosé-Hoover temperature coupling method, Parrinello-Rahman [212, 213] approach can generate a true NPT ensemble. The matrix equation of motion of the box matrix $\boldsymbol{B}$ is

$$\ddot{\boldsymbol{B}} = V\boldsymbol{W}^{-1}(\boldsymbol{B}^{\mathrm{T}})^{-1}(\boldsymbol{P} - \boldsymbol{P}_0) \tag{2-126}$$

Here, $\boldsymbol{B}^{\mathrm{T}}$ denotes the transpose of box matrix $\boldsymbol{B}$ and corresponds to the matrix $\boldsymbol{h}$ in the original paper [212, 213]. V is the volume of the box and $\boldsymbol{W}$ is a matrix that determines the strength of the coupling. In addition, equations of motion of particles for Parrinello-Rahman method are given by

$$\ddot{\boldsymbol{r}}_i = \frac{\boldsymbol{F}_i}{m_i} - \boldsymbol{M}\dot{\boldsymbol{r}}_i \tag{2-127}$$

$$\boldsymbol{M} = \boldsymbol{B}^{-1}(\boldsymbol{B}\dot{\boldsymbol{B}}^{\mathrm{T}})(\boldsymbol{B}^{\mathrm{T}})^{-1} \tag{2-128}$$

Just as Nosé-Hoover thermostat, the relaxation time with Parrinello-Rahman method is several times longer compared to the Berendsen method. Besides, note that both pressure coupling methods can be combined with the temperature coupling techniques discussed previously.

2.6.4 Stochastic Dynamics

In a system which is tightly coupled to a heat bath, the equations of motion of stochastic dynamics can be expressed by the Langevin equation [214]:

$$\ddot{\boldsymbol{r}}_i = -\xi_i \dot{\boldsymbol{r}}_i + \frac{\boldsymbol{F}_i}{m_i} + \frac{\eta(t)}{m_i} \tag{2-129}$$

where ξ_i is the friction constant and $\eta(t)$ is noise term. The correlation function of $\eta(t)$ fulfills:

$$\langle \eta_i(t)\, \eta_j(t') \rangle = 2\, m_i\, \xi_i\, k_B\, T\, \delta(t' - t)\, \delta_{ij} \tag{2-130}$$

Here the form of delta functions in time and particles implies that the random force is uncorrelated in time or between different particles. The stochastic dynamics can be regarded as molecular dynamics with stochastic temperature coupling. Compared to Berendsen and Nosé-Hoover temperature coupling, the sequence of magnitudes of velocities for different particles can be changed by a noise term. On the contrary, in the Berendsen thermostat, velocities of all particles are scaled by the same factor, which means a particle with large velocity still has relatively large velocity after scaling. In other words, Berendsen and Nosé-Hoover methods are global temperature coupling techniques while stochastic dynamics is a local temperature coupling technique. Stochastic dynamics generates correct canonical ensemble, but it alters the dynamics of the particles. Therefore, it is only applied to sample an ensemble.

2.7 QM/MM Free Energy Perturbation

QM/MM Free Energy Perturbation (QM/MM FEP) method was developed by Zhang and co-workers [215] and implemented in Chemshell [216, 217]. A typical QM/MM-FEP calculation contains two major steps. The first step is to obtain an energy profile of the reaction by a series of structure optimizations. A reaction coordinate $\xi(\boldsymbol{r}_{qm})$ which depends only on QM positions is defined. The reaction is then split into discrete windows and each window is characterized by a value ξ_i. For each window i, the entire system including both QM and MM parts is optimized by constraining the reaction coordinate to ξ_i. This results in a set of optimized geometries and a profile of the QM/MM energy of the reaction at 0K. The second step is calculating the energy of the perturbation (ΔE_{pert}) and sampling of ΔE_{pert}. The energy of perturbation between windows i and i + 1 is given by

$$\Delta E_{\text{pert}}^{i \to i+1} = E_{\text{qm/mm}}(\boldsymbol{r}_{\text{qm}}^{i+1}, \boldsymbol{r}_{\text{mm}}^{i}) - E_{\text{qm/mm}}(\boldsymbol{r}_{\text{qm}}^{i}, \boldsymbol{r}_{\text{mm}}^{i}) \tag{2-131}$$

The first term on the r.h.s is called perturbed energy, which refers to the QM/MM energy calculated with the QM positions of windows $i + 1$ and the MM positions of windows i. The second term is called unperturbed energy, which is the QM/MM energy calculated with all atoms at their positions of windows i. Then free energy difference between windows i and $i + 1$ is calculated by

$$\Delta A^{i \to i+1} \approx \Delta E_{\text{qm}}^{i \to i+1} + \Delta A_{\text{qm/mm}}^{i \to i+1} \tag{2-132}$$

The first term on the r.h.s is the difference of QM energies between two windows. The second term is the free energy change due to the QM/MM interactions and MM part. It is obtained from sampling at window i

$$\Delta A_{\text{qm/mm}}^{i \to i+1} = -\frac{1}{\beta} \ln \langle \exp(-\beta \Delta E_{\text{pert}}^{i \to i+1}) \rangle_{mm,i} \tag{2-133}$$

where $\langle\rangle$ denotes the ensemble average and $1/\beta$ is $k_B T$.

Regarding the perturbation of link atoms, there are four possibilities to deal with the positions of the link atoms (Figure 2-2) [216]: 1) Only the QM boundary atom is moved to its position in window $i + 1$. 2) The QM boundary atom is moved to its position in window $i + 1$ and the link atom is placed between the position of the QM boundary atom in window $i + 1$ and the position of the MM boundary atom in window i. 3) Both QM boundary and link atoms are moved to their positions in window $i + 1$. 4) All three atoms are moved to their positions in window $i + 1$. The original paper [216] said that the method 4 is the most consistent and promising choice. However, one can still change the setup of perturbation method for link atoms in Chemshell if it is desired [217].

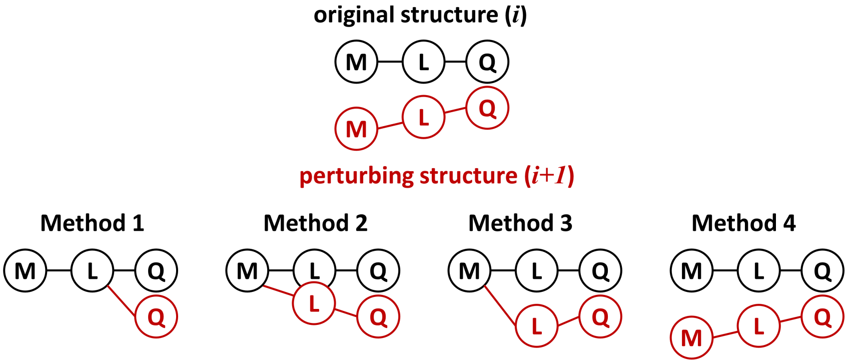


Figure 2-2. Four methods of perturbing link atoms. M, L, and Q refer to the MM boundary atom, the link atom, and the QM boundary atom respectively (modified from ref. [216]).

In practice, in order to reduce the computational cost, the QM region is usually approximated by a set of ESP charges or RESP charges [148] on QM atoms. This is conducted by a fitting process before the FEP sampling at each window. In addition, the QM part is kept frozen during the sampling. The boundary MM atoms are also fixed in order to ensure that the link atoms are also frozen.

2.8 Enzyme Kinetics Model

In chemical kinetics, a reaction rate constant plays an important role to connect experimental and theoretical studies. On the one hand, the rate constant can be measured by experiments. On the other hand, it is related to calculated activation free energy ($\Delta G^{\neq}$) by classical transition state theory:

$$k = \frac{k_B T}{h} \exp\left(-\frac{\Delta G^{\neq}}{RT}\right) \tag{2-134}$$

By equation (2-134), we can compare our calculation results with experimental values. Therefore, in the following sections, several kinetics models will be briefly discussed.

2.8.1 Michaelis-Menten Kinetics

Michaelis–Menten kinetics [218], which is named after Leonor Michaelis and Maud Menten, is one of the most famous models of enzyme kinetics in biochemistry. The mechanism of a typical enzyme reaction is

$$\mathrm{E + S} \underset{k_{\text{-1}}}{\overset{k_1}{\rightleftharpoons}} \mathrm{ES} \xrightarrow{k_{\text{cat}}} \mathrm{E + P} \tag{2-135}$$

The enzyme first binds to its substrate reversibly to form an enzyme-substrate (ES) complex in a relatively fast step. Then the ES complex breaks down in a slower second step to release the reaction product P and the free enzyme. Because the second reaction is the rate-limit step ($k_{cat} >> k_1$) of the overall reaction, the reaction velocity V is proportional to the concentration of the ES complex

$$V = k_{cat}[\mathrm{ES}] \tag{2-136}$$

As the conservation law of the enzyme, the concentration of unbound enzyme can be represented by

$$[\mathrm{E}] = [\mathrm{E_t}] - [\mathrm{ES}] \tag{2-137}$$

Then we use steady-state approximation, which means the rate of ES generation is equal to the rate of ES breakdown. This gives

$$k_1[\mathrm{E}][\mathrm{S}] = k_{-1}[\mathrm{ES}] + k_{cat}[\mathrm{ES}] \tag{2-138}$$

Combining equations (2-137) and (2-138), the concentration of ES complex can be expressed by

$$[\mathrm{ES}] = \frac{[\mathrm{E_t}][\mathrm{S}]}{K_\mathrm{M} + [\mathrm{S}]} \tag{2-139}$$

$$K_\mathrm{M} = \frac{k_{-1} + k_{cat}}{k_1} \tag{2-140}$$

where K_M is called Michaelis constant. Substituting equation (2-139) into (2-135) yields

$$V = \frac{k_{cat}[\mathrm{E_t}][\mathrm{S}]}{K_\mathrm{M} + [\mathrm{S}]} \tag{2-141}$$

This equation can be further written as

$$V = \frac{V_\mathrm{max}[\mathrm{S}]}{K_\mathrm{M} + [\mathrm{S}]} \tag{2-142}$$

where V_max is defined by $k_{cat}[\mathrm{E_t}]$. The equation (2-142) is called Michaelis–Menten equation, which is a classical rate equation for one-substrate enzyme catalyzed reactions.

2.8.2 Enzyme Kinetics with Reversible Inhibitors

An enzyme inhibitor is a molecule which binds to the enzyme and reduce its activity. In the presence of enzyme inhibitor, the enzymatic reaction is slower or even stopped. If the binding of the inhibitor to the enzyme is reversible, the inhibitor is called a reversible inhibitor; otherwise, it is an irreversible inhibitor. According to the effect of changing the concentration of the substrate on the reaction rate, the reversible inhibitors can be classified into three categories: competitive inhibitors, uncompetitive inhibitors and mixed inhibitors [219]. The kinetics schemes for three types of inhibition are shown in Figure 2-3.

A competitive inhibitor binds to the active site of the enzyme and competes with the substrate. In this case, the rate equation becomes

$$V = \frac{V_\mathrm{max}[\mathrm{S}]}{\alpha K_\mathrm{M} + [\mathrm{S}]} \tag{2-143}$$

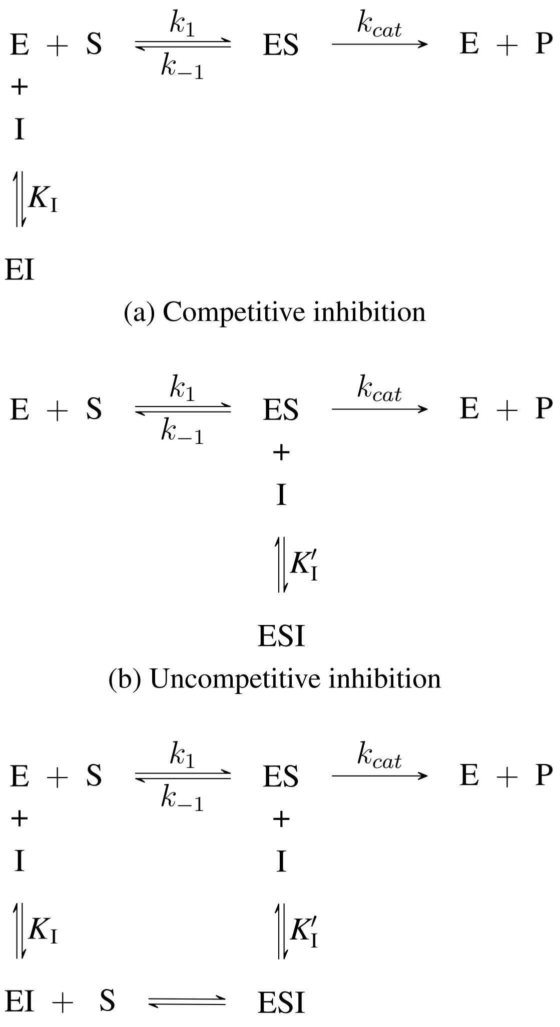
where

$$\begin{aligned} \alpha &= 1 + \frac{[\mathrm{I}]}{K_\mathrm{I}} \\ K_\mathrm{I} &= \frac{[\mathrm{E}][\mathrm{I}]}{[\mathrm{EI}]} \end{aligned} \tag{2-144}$$

The equation (2-143) has a similar form to the Michaelis–Menten equation. The variable of αK_M, which can be determined by experiment, is often called apparent K_M.

In uncompetitive inhibition, an inhibitor only binds to the enzyme-substrate complex. The rate equation is now altered to

$$V = \frac{V_\mathrm{max}[\mathrm{S}]}{K_\mathrm{M} + \alpha'[\mathrm{S}]} \tag{2-145}$$



(a) Competitive inhibition

(b) Uncompetitive inhibition

(c) Mixed inhibition

Figure 2-3. Three types of reversible inhibition. a) competitive inhibition, b) uncompetitive inhibition, c) mixed inhibition (modified from ref. [219]).

where

$$\begin{aligned}\alpha' &= 1 + \frac{[\mathrm{I}]}{K'_{\mathrm{I}}} \\ K'_{\mathrm{I}} &= \frac{[\mathrm{ES}][\mathrm{I}]}{[\mathrm{ESI}]}\end{aligned} \tag{2-146}$$

The maximum reaction rate is reduced to V_{max}/α' and the apparent K_{M} is also decreased by a factor of α'.

A mixed inhibitor can bind to either enzyme or enzyme-substrate. The rate equation for mixed inhibition is

$$V = \frac{V_{\mathrm{max}}[\mathrm{S}]}{\alpha K_{\mathrm{M}} + \alpha'[\mathrm{S}]} \tag{2-147}$$

where α and α' are defined as above.

2.8.3 Slow-binding Enzyme Inhibitors

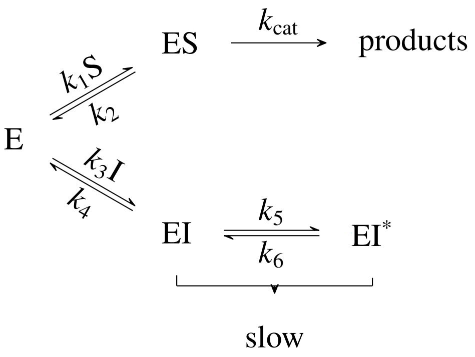


Figure 2-4. The kinetics mechanism of enzyme-catalyzed reaction with a slow-binding enzyme inhibitor (modified from ref. [220]).

If an enzyme is inhibited by a compound in a time-dependent manner, the inhibitor is called a slow-binding enzyme inhibitor [220]. The establishment of equilibrium between enzyme (E), inhibitor and enzyme-inhibitor complex (EI*) occurs slowly. Figure 2-4 illustrates the kinetics mechanism of enzyme-catalyzed reaction with a slow-binding enzyme inhibitor. In this Figure, ES represents the enzyme-substrate complex and EI an initial collision complex with the inhibitor. In this case, enzyme kinetics cannot be described by classical Michaelis-Menten kinetics. The progress curve for a given concentration of inhibitor as a function of time is described by

$$P = v_s t + (v_0 - v_s)(1 - e^{-kt})/k \tag{2-148}$$

where P is the quantity of the product, v_0 the initial rate, v_s the final steady-state reaction rate and k the apparent first-order rate constant for establishing equilibrium between EI and EI*. The detail derivation of the equation please refer to the original book [220]. The initial rate v_0 as a function of the concentration of inhibitor is given by

$$v_0 = \frac{V_{\text{max}}[\text{S}]}{K_{\text{M}}(1 + [\text{I}]/K_{\text{I}}) + [\text{S}]} \tag{2-149}$$

where K_{I} is the dissociation constant for the EI complex which is equal to k_3/k_4. The equation (2-149) has the same form as the rate equation (2-143) for competitive inhibitors. Similarly, the final steady-state rate v_s is also a function of [I], which is given by

$$v_s = \frac{V_{\text{max}}[\text{S}]}{K_{\text{M}}(1 + [\text{I}]/K_{\text{I}}^*) + [\text{S}]} \tag{2-150}$$

where K_{I}^* is overall inhibition constant, which is defined as

$$K_{\text{I}}^* = \frac{[\text{E}][\text{I}]}{[\text{EI}] + [\text{EI*}]} = \frac{K_{\text{I}} k_6}{k_5 + k_6} \tag{2-151}$$

The apparent first-order rate constant k for the interconversion between EI and EI* is a function of [I] and is expressed as

$$k = k_6 + k_5 \times \frac{[\text{I}]/K_{\text{I}}}{1 + [\text{S}]/K_{\text{M}} + [\text{I}]/K_{\text{I}}} \tag{2-152}$$

Chapter 3

QM/MM-based prediction of covalently-bound ligands binding to ClpP

3.1 The generally accepted Reaction Mechanism

3.1.1 Acylation step

The generally accepted reaction mechanism of the acylation reaction of SaClpP with three different kinds of ligands or substrates is shown in Figure 3-1. The first step is nucleophilic attack of S98 to the ligand forming a tetrahedral intermediate. In the meanwhile, proton moves from S98 to H123. Then proton transfers from H123 to the ligand. For β-lactones, proton transfer will cause the open of the four-membered ring. For the phenyl ester and the fluorescent substrate, proton transfer will lead to the departure of the leaving group.

3.1.2 Deacylation step

The generally accepted reaction mechanism of the deacylation reaction is shown in Figure 3-2. All four ligands have the same reaction mechanism because the deacylation reaction is simply a hydrolysis reaction of an ester. The water molecule is the nucleophile for the deacylation reaction. Thus the first step is nucleophilic attack of water to the ester (modified S98) forming a tetrahedral intermediate. In the meanwhile, the proton is bonded to H123. Then proton transfers from H123 to S98 with the release of a carboxylic acid and the regeneration of S98.

3.2 Deriving experimental rate constants

3.2.1 Acylation step

Comparing calculation results with experimental values can help us judge the quality of our computational model. Our experimental collaborators observed that the inhibitory

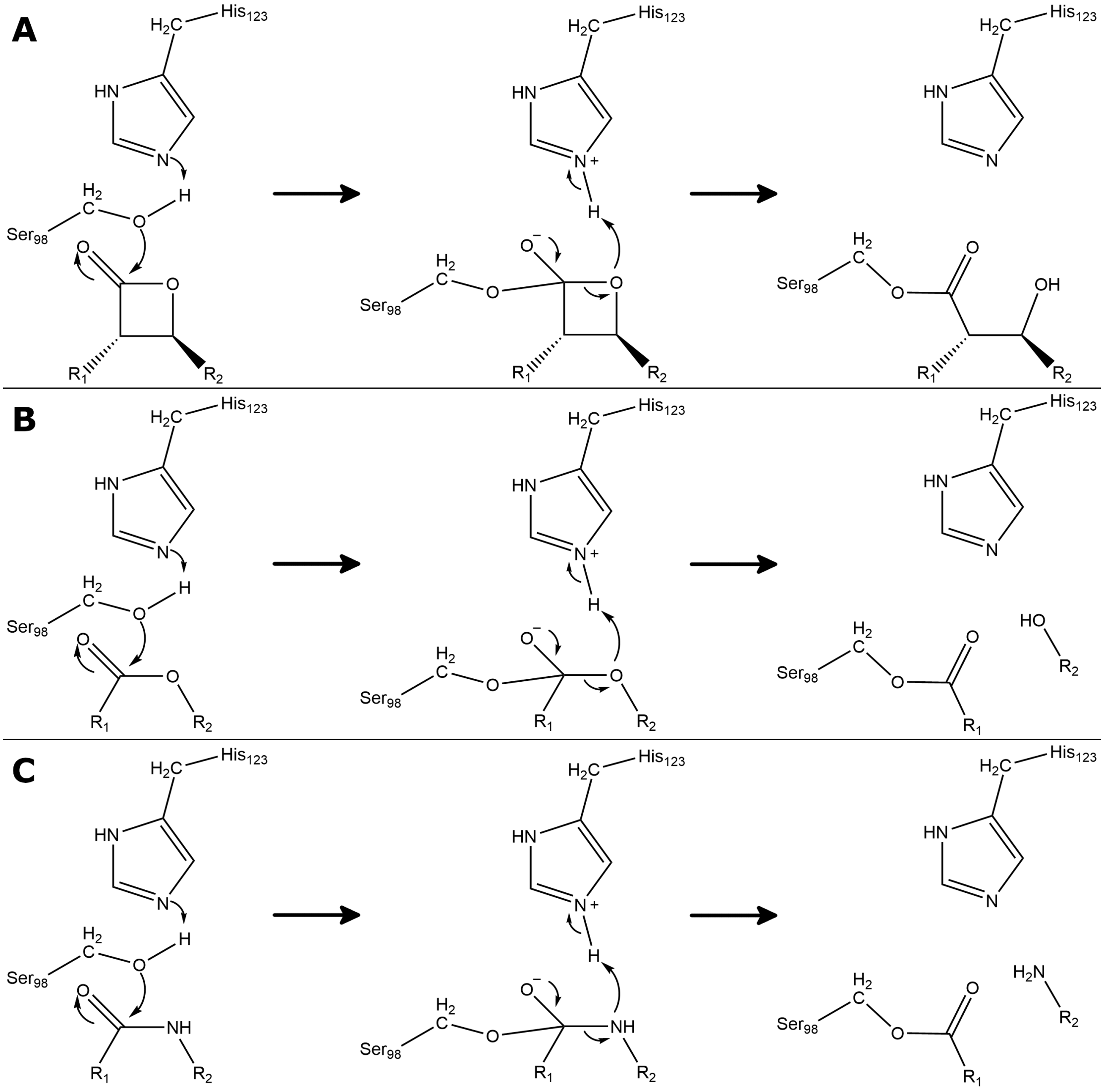


Figure 3-1. The generally accepted mechanism of the acylation reaction for SaClpP with (A) β-lactones, (B) the phenyl ester and (C) the fluorescent substrate.

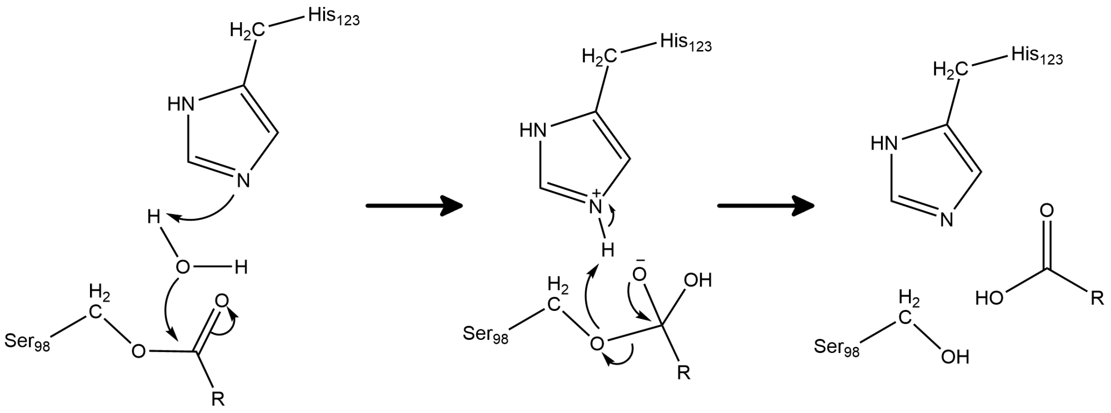


Figure 3-2. The generally accepted mechanism of the deacylation reaction for acyl-enzyme complexes of SaClpP.

reaction does not follow classical Michaelis-Menten kinetics [31, 221]. They only measured k_{obs}/[I] instead of rate constants that are related to activation free energies. Therefore, our calculation results cannot be directly compared with experimental values. However, by further inspecting the original experimental data, we noticed that the systems of SaClpP with ligands follow the kinetics mechanism of slow-binding enzyme inhibitors [220]. Firstly, experimental progress curves [31, 221] show a similar pattern of an inhibition with a slow-binding inhibitor [220]. Although it is a reversible reaction for the conversion between EI and EI* in the sketch of the mechanism (Figure 2-4), the mechanism is still applicable to covalently-bound inhibitors. The reason is that "when the rate at which an inhibitor dissociates from an enzyme-inhibitor complex becomes very slow, no kinetic distinction can be made between inactivation, because of covalent attachment of the inhibitor, and slow-binding inhibition" [220]. Mathematically, this simply implies that k_6 is a very small value compared to k_5. Furthermore, John Morrison also gave a rough indication of k_{obs}/[I] to determine whether it is a slow-binding inhibitor: "A calculated value of $< 10^6$ M^{-1}s^{-1} for the bimolecular rate constant associated with the

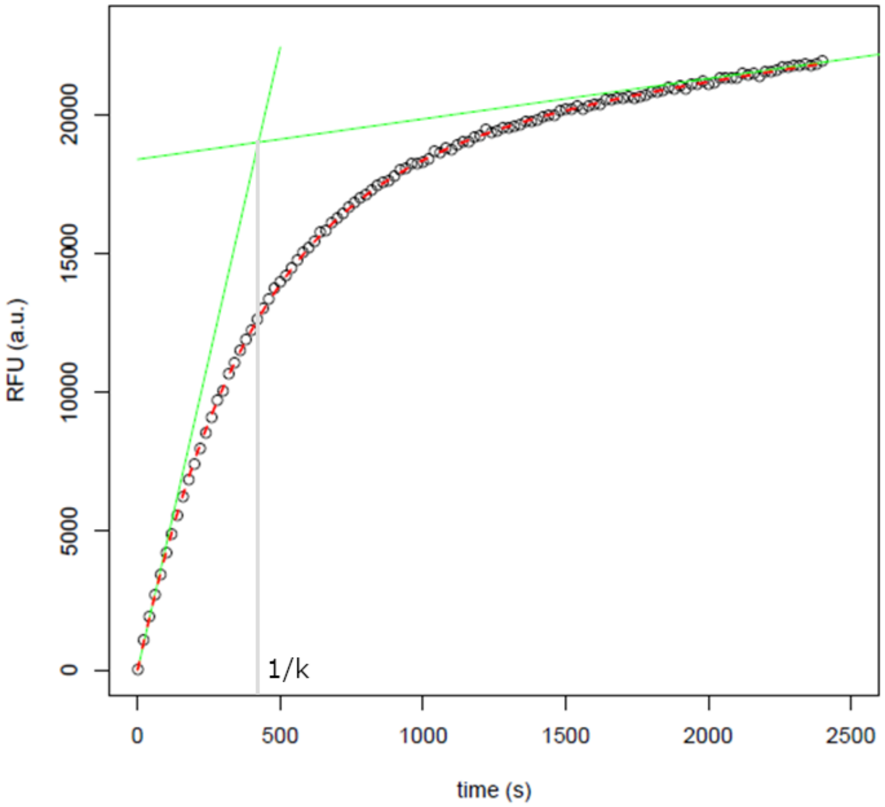


Figure 3-3. An example of estimating k by non-linear fitting a progress curve. k is the reciprocal of time at the intersection point of the tangent with the asymptote. The progress curve corresponds to the reaction of the fluorescent substrate (200 μM) with SaClpP (1 μM) in the presence of 35 μM Lig24 at 32 °C [221].

interaction of enzyme and inhibitor (k_{obs}/[I]) can be taken as a preliminary indication that the inhibition involves slowly equilibrating enzyme-inhibitor species" [220]. Since k_{obs}/[I] for two β-lactones and the phenyl ester are smaller than this criteria, we thought the mechanism applies to our systems.

The rate constants for the acylation reaction of SaClpP with different ligands were derived by the kinetics model of slow-binding inhibitor shown in section 2.8.3. The procedures are listed as follows:

1. For each progress curve with a different concentration of inhibitor, draw a tangent of the curve at the beginning and an asymptote of the curve to estimate v_0 and v_s (Figure 3-3).
2. Estimate the k value from the intersection point of the tangent with the asymptote where $t = 1/k$.
3. Perform a non-linear fitting to obtain precise k by equation (2-148).
4. Plot all k values against [I] (Figure 3-4).
5. Perform a non-linear fitting to calculate the rate constant by equation (2-152).

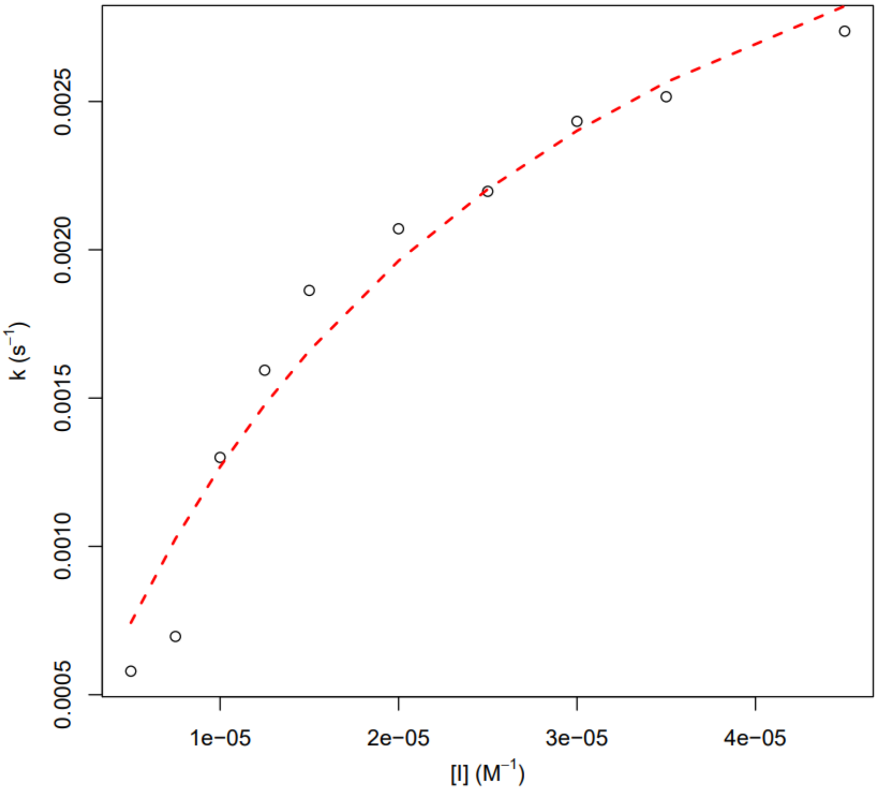


Figure 3-4. A plot of k against [I] for Lig24. The red dotted line is a fitted line to derive the rate constant.

The derived rate constants for the acylation reaction are listed in Table 3.1. Experimental activation free energies are calculated by equation (2-134). As can be seen from

Table 3.1. Derived experimental rate constants and free energy barriers for acylation reactions of SaClpP with different ligands. (*: Data from ref. [26])

Ligand	k_{obs}/[I]	k_5	$\Delta G^{\neq}$
	[$\text{s}^{-1}\text{M}^{-1}$]	[s^{-1}]	[kcal mol^{-1}]
Lig24	77	4.34×10^{-3}	20.8
Lig25	2.18	6.63×10^{-3}	20.6
ML90	39	3.41×10^{-3}	21.0
Sub	N/A	0.57*	18.5*

Table 3.1, the experimental activation free energies are almost the same for all three ligands and our computational model is difficult to differentiate them. Regarding the reaction of SaClpP with the fluorescent substrate Suc-Leu-Tyr-AMC, our experimental collaborators show that the reaction velocity depends linearly on the concentration of the substrate in the range of 50 to 2200 μM with no observable saturation [221]. However, Zhang and Ye [24, 26] reported that the reaction still follows Michaelis-Menten kinetics. We, therefore, put the rate constant from their latest paper in Table 3.1 for comparison. The reason for the difference in experiments between two groups is still unclear.

3.2.2 Deacylation step

In contrast to the acylation reaction, one can use mass spectrometry to measure the half-life of the covalently bound complex to derive the kinetics constant and the corresponding activation free energy for the deacylation reaction. Experimental half-lives [31, 222], corresponding rate constants, and activation free energies of deacylation reactions for three ligands are listed in Table 3.2. However, no covalent adducts of the acyl-enzyme intermediate of SaClpP bound to the fluorescent substrate could be detected [222]. The substrate was present in 1000-fold excess over the protease during the measurement, still there is no covalent adduct. Therefore, we were not be able to determine $t_{1/2}$ of the substrate acyl-enzyme intermediate. It is just too short-lived to be measured by mass spectrometry, which also implies that the activation free energy of the deacylation step is lower than that of the acylation step for the fluorescent substrate.

3.3 Systems preparation and calculation setup

3.3.1 Acylation step

The initial structure of the SaClpP protein was prepared based on the PDB file 3V5E [25]. Nine chains were cut out and only chains A, B, G, K, and L were kept to reduce

Table 3.2. Experimental half-lives, rate constants, and free energy barriers for deacylation reactions of SaClpP with different ligands

Ligand	$t_{1/2}$	k	$\Delta G^{\neq}$
	[h]	[s^{-1}]	[kcal/mol]
Lig24	5.5	3.50×10^{-5}	23.7
Lig25	8	2.41×10^{-5}	23.9
ML90	9.5	2.03×10^{-5}	24.0
Sub	N/A	N/A	N/A

the size of the system. The ligands were docked by the program Dynadock [223] to get the structures of the complexes. The docking for the fluorescent substrate used knowledge of the crystal structure of *E. Coli* ClpP with Z-LY-CMK (PDB ID: 2FZS) [224]. The ff99SB Amber force field [144] was used for protein, whereas the ligands were parametrized with the general amber force field (GAFF) [150]. The RESP charges [148] of the ligands were determined at the HF/6-31G* level by Gaussian09 [153]. The structures of the complexes were prepared for minimization with the tleap utility [225] from AmberTools15 [142] and all calculations were conducted in a neutralized, rectangular TIP3P [154] water box extending at least 15 Å from any protein atom

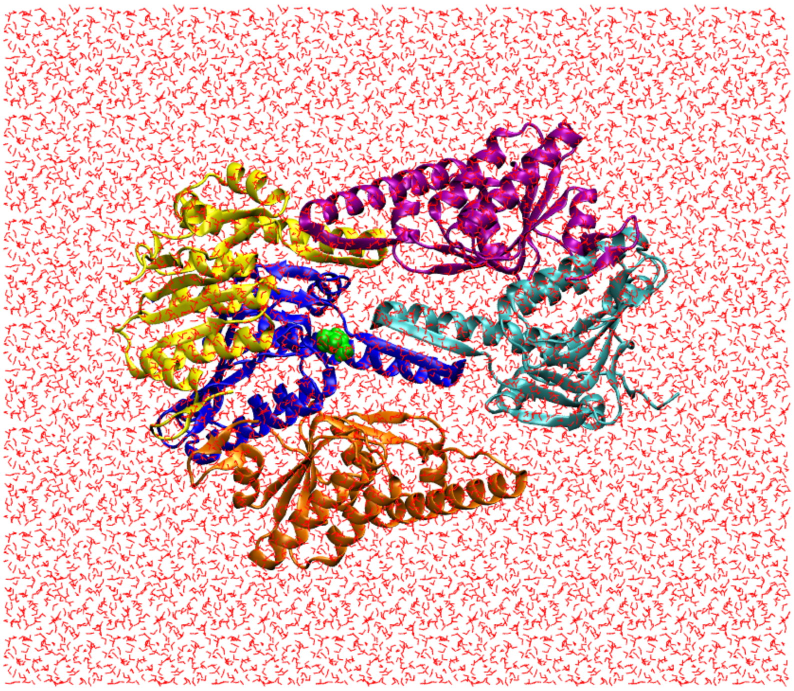


Figure 3-5. Illustration of the simulation box (Lig25) prepared for the MD simulations. The ligand is shown in green, water molecules are in red and five chains are in different colors.

at each side of the box (Figure 3-5). Energy minimizations were performed with pmemd.cuda or pmemd.MPI from the Amber16 software package [143]. For every complex, two subsequent minimizations were conducted. Firstly, 100100 steps of restrained minimization (100 steps with the steepest descent algorithm and 100000 steps with the conjugate gradient method) were performed with the protein atoms restrained using a 50 kcal·mol^{-1}·Å^{-2} force constant. Then 100100 steps of energy minimization (100 steps with the steepest descent and 100000 steps with the conjugate gradient method) were conducted without restraints. Both minimizations were considered as converged if the root-mean-square of the Cartesian components of the energy gradient was less than 0.002 kcal·mol^{-1}·Å^{-1}. The non-bonded interaction cutoff was set to 8.0 Å for both energy optimizations.

Before conducting production runs, all systems were heated up by stepwise increasing the temperature over 660 ps while at the same time incrementally decreasing the number of restraint atoms as well as the force acting on them. The detailed heat-up protocol is listed in Table 3.3. At each heat-up step, the initial velocities of atoms were randomly assigned from a Maxwell-Boltzmann distribution at the given target temperature. MD simulations were performed with 1 fs time step. Non-bonded interactions were computed applying a cutoff of 10 Å. The Particle Mesh Ewald method [226] was used to calculated long-range electrostatic interaction. The SHAKE algorithm [227] was applied

Table 3.3. The protocol of heat-up procedures for MD simulations of SaClpP with ligands

Equilibration step	Temperature [K]	Time [ps]	Force constant [kcal·mol^{-1}·Å^{-2}]	Restraint Atoms
1	0	10	2.39	all atoms
2	5	50	2.39	all atoms
3	10	50	2.39	all atoms
4	20	50	2.39	all atoms
5	50	50	2.39	all backbone heavy atoms and ester group of the ligand
6	100	50	2.39	all backbone heavy atoms and ester group of the ligand
7	200	50	2.39	all backbone heavy atoms and ester group of the ligand
8	200	50	0.24	all backbone heavy atoms and ester group of the ligand
9	200	100	0.24	protein backbone heavy atoms
10	300	100	0.24	protein backbone heavy atoms
11	300	100	0	no atoms

to constrain all bonds involving hydrogen atoms. The temperature was kept constant using the Berendsen thermostat [209] with a time constant of 0.1 ps to ensure a constant temperature. The Berendsen barostat [209] was applied with a compressibility of 45 × 10^{-6} bar^{-1} and a pressure relaxation time of 1 ps to keep a constant target pressure of 1 bar. All MD simulations were performed by pmemd.MPI or pmemd.cuda programs from the Amber16 software package [143]. After equilibration molecular dynamics were performed for 30 ns under periodic boundary conditions in the NPT ensemble at 300 K. The initial structure for the reaction simulation was chosen based on the following criterion. There must be hydrogen bonds between H123 and D172 as well as between S98 and H123. The oxygen in the carbonyl group has to be in the oxyanion hole. Then, the value of the first reaction coordinate (equation (3-1)), which is negative in the reactant state, was calculated for each filtered structure. The structure with the maximum of the calculated value (namely the minimum in terms of the absolute value) was chosen as the initial structure.

After selecting the initial structure as described above, the system was first optimized at the MM level while keeping the QM region frozen. The QM region consists of side chains of the catalytic triad (S98, H123, and D172) and the ligand. All residues and water molecules further than 35 Å away from the oxygen of the S98 hydroxyl group were cut out using an in-house python script (Figure 3-6). The remaining termini were capped by either ACE or NME residues. If the residues placed directly before and after the deleted residue were inside the sphere, the deleted residue was substituted by a glycine. The outmost 5 Å layer was fully fixed for the calculations.

All QM/MM calculations were driven by ChemShell software package, [182, 217], which provides interfaces to different QM programs and dl_poly [228]. Turbomole [135] was used as a QM program to perform all DFT QM/MM calculations. Orca 4.0 [229, 230] was used for semi-empirical QM/MM calculations. Geometry optimizations were performed with the dl_find module [205] of ChemShell. The maximum number of optimization cycles was set to 1000. The RI approximation was used for density functional theory (DFT) calculations. The criterion of SCF convergence was 10^{-6} Hartree. The link-atom scheme was used to treat the QM-MM boundary and electrostatic embedding was used to deal with electrostatic QM-MM interactions. In addition, point charges of MM frontier atoms were shifted away and point dipoles were added to compensate [158]. Electrostatic interactions between MM atoms were evaluated for all atom pairs. Some semi-empirical QM/MM calculations were conducted using Amber16 [143]. The setup for Amber QM/MM calculations were almost identical to ChemShell QM/MM calculations. One difference was that the maximum number of optimization

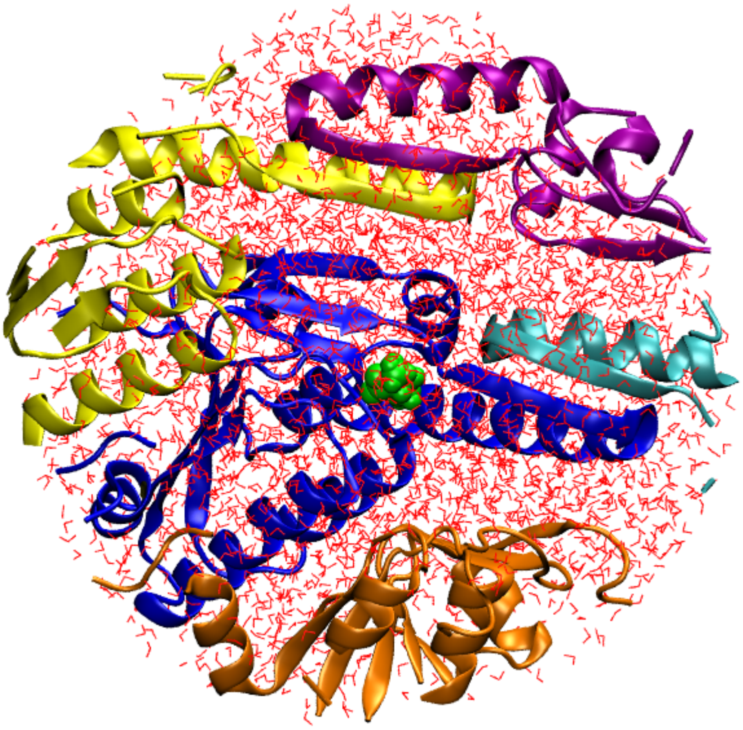


Figure 3-6. Illustration of the sphere model (Lig25) for the QM/MM calculations. The ligand is shown in green, water molecules are in red and five chains are in different colors.

cycles was set to 10000 for Amber QM/MM calculations (more details in section 3.4.2.1).

Two reaction coordinates were used for the study (Figure 3-7). The first reaction coordinate (nucleophilic attack reaction coordinate) is the traditional 1D reaction coordinate which is used in many previous studies [8, 72–78, 80–85]. It is defined as

$$\xi 1 = d(\mathrm{O_{Ser}} - \mathrm{H_{Ser}}) - d(\mathrm{N_{His}} - \mathrm{H_{Ser}}) - d(\mathrm{O_{Ser}} - \mathrm{C_{Lig}}) \tag{3-1}$$

for all ligands (Figure 3-7A-C). The second reaction coordinate (proton transfer reaction coordinate) is defined as

$$\xi 2 = d(\mathrm{C_{Lig}} - \mathrm{O_{Lig}}) - d(\mathrm{O_{Lig}} - \mathrm{H_{Ser}}) \tag{3-2}$$

for β-lactones and the phenyl ester (Figure 3-7A,B) and

$$\xi 2 = d(\mathrm{C_{Lig}} - \mathrm{N_{Lig}}) - d(\mathrm{N_{Lig}} - \mathrm{H_{Ser}}) \tag{3-3}$$

for the fluorescent substrate (Figure 3-7C). In this study, most potential energy surfaces were calculated in two dimensions. After preparing the model for QM/MM calculations, an initial energy minimization was performed while keeping the QM region frozen.

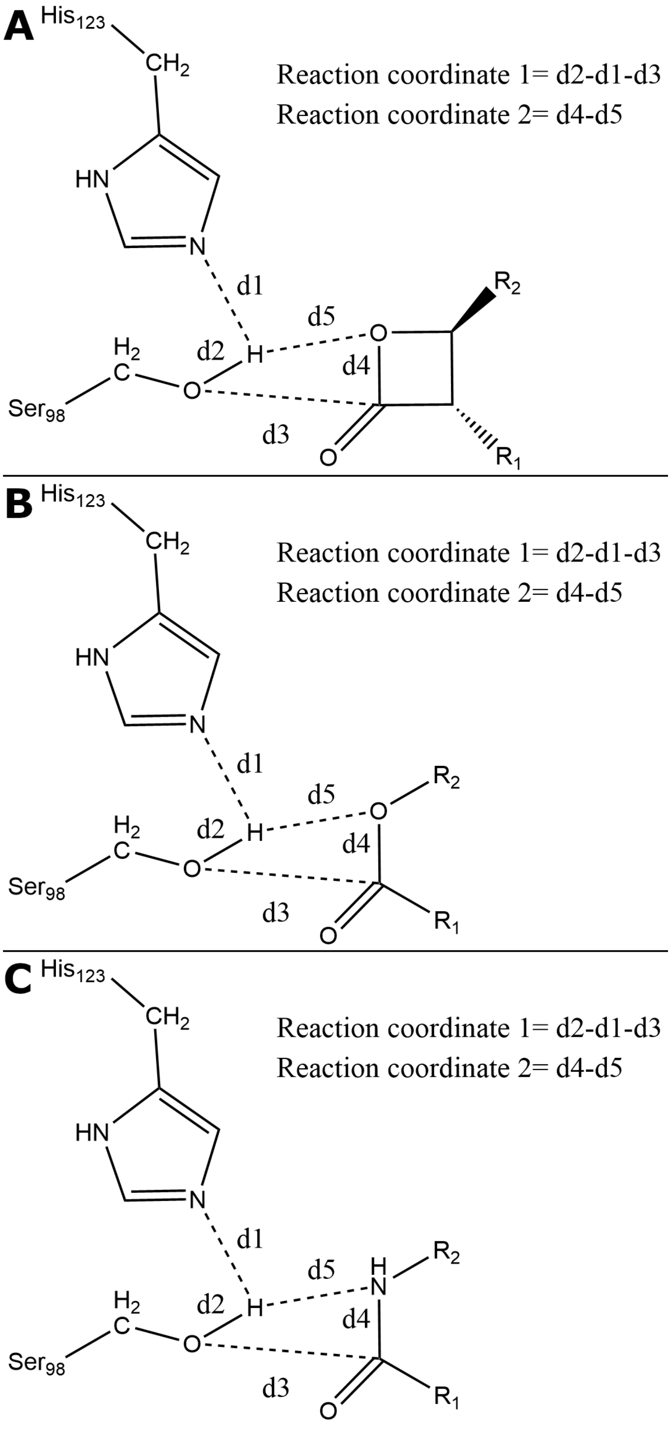

A
His123
CH2
Reaction coordinate 1= d2-d1-d3
Reaction coordinate 2= d4-d5
HN
N
d1
d5
d2
d4
d3
H
O
R2
R1
Ser98
B
His123
CH2
Reaction coordinate 1= d2-d1-d3
Reaction coordinate 2= d4-d5
HN
N
d1
d5
d2
d4
d3
H
O
R2
R1
Ser98
C
His123
CH2
Reaction coordinate 1= d2-d1-d3
Reaction coordinate 2= d4-d5
HN
N
d1
d5
d2
d4
d3
H
N
O
R2
R1
Ser98

Figure 3-7. The setup of reaction coordinates for the acylation reaction of SaClpP with (A) β-lactone, (B) phenyl ester and (C) fluorescent substrate.

For 1D calculations, a minimum energy reaction profile was determined by two steps. The first step was performing restrained geometry optimizations by gradually increasing $\xi 1$ until a local minimum was reached. Then a new set of restrained geometry optimizations was conducted by increasing $\xi 2$ until another local minimum was reached. At each point, the reaction coordinate was restrained to a specific value by a harmonic potential with a force constant of 3.0 Hartree/Bohr^2. The interval between two neighboring points is 0.1 Å.

A 2D potential energy surface was determined by restrained geometry optimizations at grid points. The calculation started from the grid point at the bottom left corner. The first series of geometry optimizations were performed by gradually increasing $\xi 2$ (blue arrow towards up in Figure 3-8). Then, several parallel calculations were conducted by fixing $\xi 2$ and increasing $\xi 1$ (blue arrows towards right in Figure 3-8) to obtain the bottom part of the surface. The top left part of the surface (white area in Figure 3-8) was not calculated due to very high potential energies in this area. In this case, the proton of S98 is close to the ligand and the C-O bond (or the peptide bond for the fluorescent substrate) is partially broken whereas the ester bond is not formed. The top right part of the potential energy surface was obtained by several parallel geometry optimizations along the reaction coordinate 2 (red arrows in Figure 3-8). The interval between grid points along the reaction coordinate 1 is 0.1 Å and along the reaction coordinate 2 is 0.2 Å. At each grid point, both reaction coordinates were restrained by a harmonic potential with a force constant of 3.0 Hartree/Bohr^2. The geometry was optimized at the BP86/def2-SVP level [116, 117, 133] Afterwards, single point calculations were performed for all optimized structures at the B3LYP/def2-TZVP level [121, 122, 133] to obtain the final potential energy surface. D3 dispersion correction [125] was applied to all geometry optimization and single point calculations.

QM/MM free-energy perturbation calculations were performed to estimate free energy barriers of the acylation step. The method calculates the free energy difference between two states when an energy profile of a reaction with optimized structures is known. The reaction is then split into discrete windows. In our case, these windows are grid points on calculated 2D potential energy surface. For each state of interest, a path connecting the reactant state and that state was constructed to perform free energy perturbation calculations (e.g., red path in Figure 3-23). The qmmmfep module implemented in ChemShell was used for this purpose [217]. We slightly modified the module to use RESP charges for QM atoms instead of default ESP charges. The 200 closest MM atoms were used as reference points when fitting the RESP charges. The fitting of QM charges was based on DFT calculations at the B3LYP/def2-TZVP level. The QM part, MM frontier atoms and outmost 5 Å layer were kept frozen. The method of link atom perturbation was set to 4, implying that QM boundary atoms, link atoms and MM boundary atoms were all perturbed [216]. Free water molecules were kept internally rigid by using the SHAKE algorithm [227]. In each window, the system was

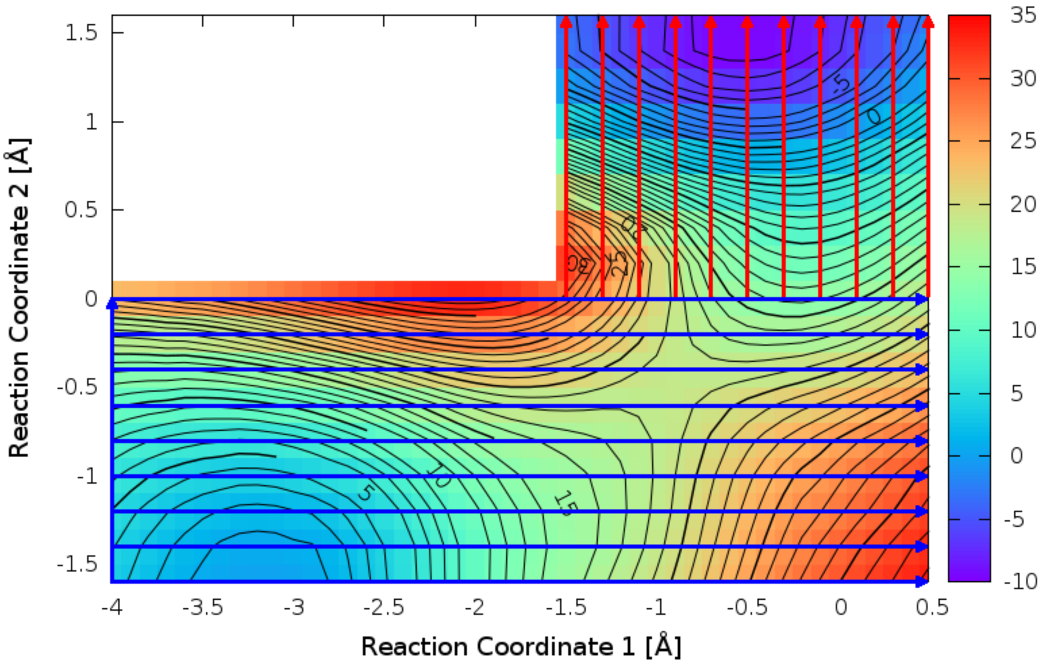
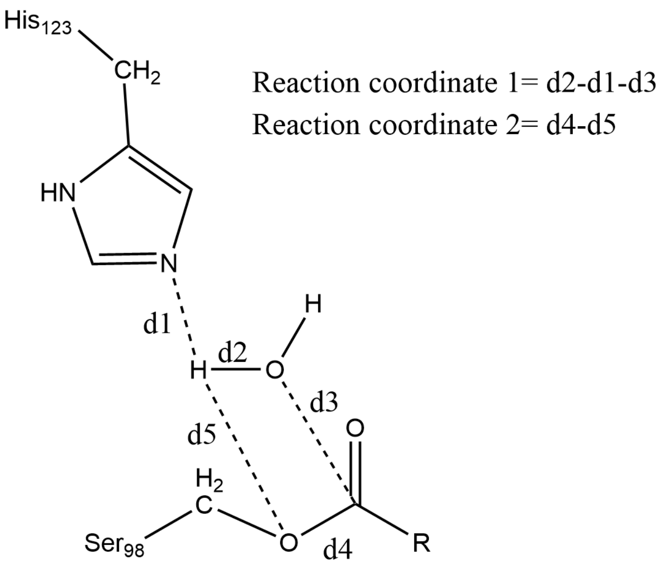


Figure 3-8. Illustration of scan directions to obtain a 2D PES. The gray area is not calculated due to high potential energies at grid points in this area. Blue arrows are scan direction for the bottom part of the surface and red arrows are for the upper right part.

heated up and equilibrated at 300 K for 10 ps in the NVT ensemble with 1 fs time step. The Nosé-Hoover chain thermostat [210, 211] was used with a chain length of 4 and time constant (τ_T) of 0.02 ps. The FEP production run was performed for forward perturbation with a time step of 1 fs for 10 to 15 ps depending on the convergence check [216, 231]. All hydrogen atoms were assigned the mass of deuterium to ensure energy conservation with a time step of 1 fs [216].

3.3.2 Deacylation step

Similar to the calculations of the acylation step, molecular dynamics simulations were performed before the QM/MM calculations. Since the protein-substrate complex was cut for QM/MM simulations of the acylation step, the whole system of the acyl-enzyme complex had to be rebuilt to perform MD simulations. The initial structures for MD simulations of the acyl-enzyme complex were taken from the original structures before the preparation for QM/MM simulations. The positions of active sites and surrounding atoms (unfrozen atoms during acylation simulations) were replaced by the positions of product states of acylation reactions calculated at the B3LYP/def2-TZVP level. For the phenyl ester ML90 and the fluorescent substrate, the leaving group in previous simulations was removed from the system. RESP charges of the modified S98 were derived again by Gaussian 09 [153]. The setup of MD simulations for the deacylation step was the same as that for the acylation step, except that the Langevin thermostat [214] was used instead of the Berendsen thermostat. After equilibration molecular dynamics were performed for 200 ns under periodic boundary conditions in the NPT ensemble at 300 K. A two-step clustering scheme was used to select the starting structure for subsequent QM/MM calculations. All clusterings were performed by cpptraj program from the Amber16 software package [143]. For each snapshot from the MD trajectory, the closest water molecule near the ester group and H123 was kept and all other water molecules were removed. The first clustering was performed based on all backbone atoms by a hierarchical agglomerative approach [232]. The largest cluster was further clustered based on the oxygen atom of the water molecule. Then representative structures of all sub-clusters were checked until the largest sub-cluster with a reasonable position of the water molecule was found, namely there is a hydrogen bond between the water molecule and H123 and the water molecule has to be close enough to the ligand to start nucleophilic attack. In this sub-cluster,the value of the first reaction coordinate (equation (3-4)) was calculated for each structure. The structure with the maximum of the calculated value was selected as the initial structure.



Reaction coordinate 1= d2-d1-d3
Reaction coordinate 2= d4-d5

Figure 3-9. The setup of reaction coordinates for the deacylation reaction of SaClpP with ligands.

QM/MM calculations for the deacylation step were performed with four different setups. The reason for conducting calculations with different setups will be discussed in detail in the results part (see section 3.4.3). The differences between them are as follows:

Setup 1 The normal calculation which is similar to the calculation for the acylation step

Setup 2 After performing a short MD simulation at the transition state from calculations with setup 1, conducting QM/MM calculations again from the last frame of the MD simulation to obtain a new PES

Setup 3	Similar to the setup 2, but including more water molecules into the QM region
Setup 4	1D QM/MM FEP calculations based on the PES from calculations with setup 3

Calculations with setup 1 are similar to the acylation step. Potential energy surfaces were calculated in two dimensions and two reaction coordinates were defined as (Figure 3-9):

$$\xi 1 = d(\mathrm{O_{Wat}} - \mathrm{H_{Wat}}) - d(\mathrm{N_{His}} - \mathrm{H_{Wat}}) - d(\mathrm{O_{Wat}} - \mathrm{C_{Lig}}) \tag{3-4}$$

$$\xi 2 = d(\mathrm{C_{Lig}} - \mathrm{O_{Ser}}) - d(\mathrm{O_{Ser}} - \mathrm{H_{Wat}}) \tag{3-5}$$

for all ligands. The QM region consists of side chains of the catalytic triad (modified S98, H123, and D172) and the reactive water molecule. The rest setup of QM/MM calculations for the deacylation step was exactly the same as that for the acylation step.

Calculations with setup 2 added a short MD simulation before QM/MM calculations to obtain a new PES. The initial structure of the MD simulation was the transition state on the potential energy surface calculated with setup 1 at the B3LYP level. During the MD simulation, only MM water molecules within 20 Å of the reactive center were allowed to move. All other atoms including the QM part were kept frozen. ESP charges for the QM atoms were first derived and a short 20 ps of MD simulation were performed at 300 K with 1 fs time step. Then an energy minimization was performed at the last frame by fixing all frozen atoms. After that, restrained geometry optimizations were performed with the normal setup to obtain a new 2D potential energy surface.

Calculations with setup 3 were very similar to setup 2. The only difference was the number of QM water molecules. All water molecules within 6 Å of the reactive water molecule were also treated at the QM level. In order to prevent proton hopping, weak distance restraints of 0.03 Hartree/Bohr^2 were applied to all OH bonds of these QM water molecules. Energy minimizations were started from the optimized structure after the short MD simulation in setup 2 with a larger QM region.

Calculations with setup 4 were 1D QM/MM free-energy perturbation calculations [215, 216] based on the 2D potential energy surface derived from calculations with setup 3. The aim was to further estimate activation free energies and free energy differences of deacylation reactions. The setup of QM/MM FEP calculations for the deacylation step is the same as that of the acylation step.

3.4 Calculation results

3.4.1 Binding of ligands to ClpP

Initial structures for QM/MM simulations are shown in Figure 3-10 to illustrate the binding site of the protease. The binding sites of SaClpP for all four ligands share some common features. Two hydrogen bonds connect the residues of the catalytic triad. One is between S98 and H123 and another one is between H123 and D172. The oxygen atom of the carbonyl group forms two hydrogen bonds to the backbone of G69 and M99 and is stabilized by the oxyanion hole. In the case of Lig24 and Lig25, aliphatic chains of β-lactones sit in the binding pocket (Figure 3-10A, B). For the phenyl ester, the phenol group is in the binding pocket (Figure 3-10C) while for the fluorescent substate, the side chain of the tyrosine residue is in the pocket (Figure 3-10D).

3.4.2 Acylation step

3.4.2.1 1D semi-empirical calculations

Five different semi-empirical QM methods were tested to simulate the nucleophilic attack step of the acylation reaction for four ligands by Amber [143]. Among these methods, AM1 and PM3 calculations were also performed by ChemShell program [182, 217]. Calculated reaction profiles are shown in Figure 3-11. In general, reaction profiles calculated by Chemshell and Amber have a similar trend. But the difference in calculated energy barriers between two programs is up to 6.9 kcal mol^{-1}. The difference is mainly coming from two aspects: first, generalized Born (GB) model is automatically applied to the entire system in all Amber QM/MM calculations and can not be turned off [143, 233]. Second, Amber QM/MM calculations do not support the charge shift scheme which is a default setup in ChemShell.

PM3 calculated reaction profiles show very large energy barriers in all cases (36.9–51.0 kcal mol^{-1}). These values are significantly higher than the experimental values (Table 3.1). On the other hand, although PM3 and PM6 belong to the same family of semi-empirical QM methods, PM6 calculated energy barriers (1.8–9.5 kcal mol^{-1}) are remarkably smaller than the experimental values. In addition, RM1 calculated energy barriers (13.5–17.4 kcal mol^{-1}) are slightly lower than the experimental values for β-lactones and the phenyl ester. But all RM1 reaction profiles show a cusp at the peak of the reaction profile. Besides, an energy drop was observed for reaction profiles calculated by the DFTB method for β-lactones (Figure 3-11A, B). This is because the proton automatically transfers to the ligand. On the contrary, DFTB reaction profile for

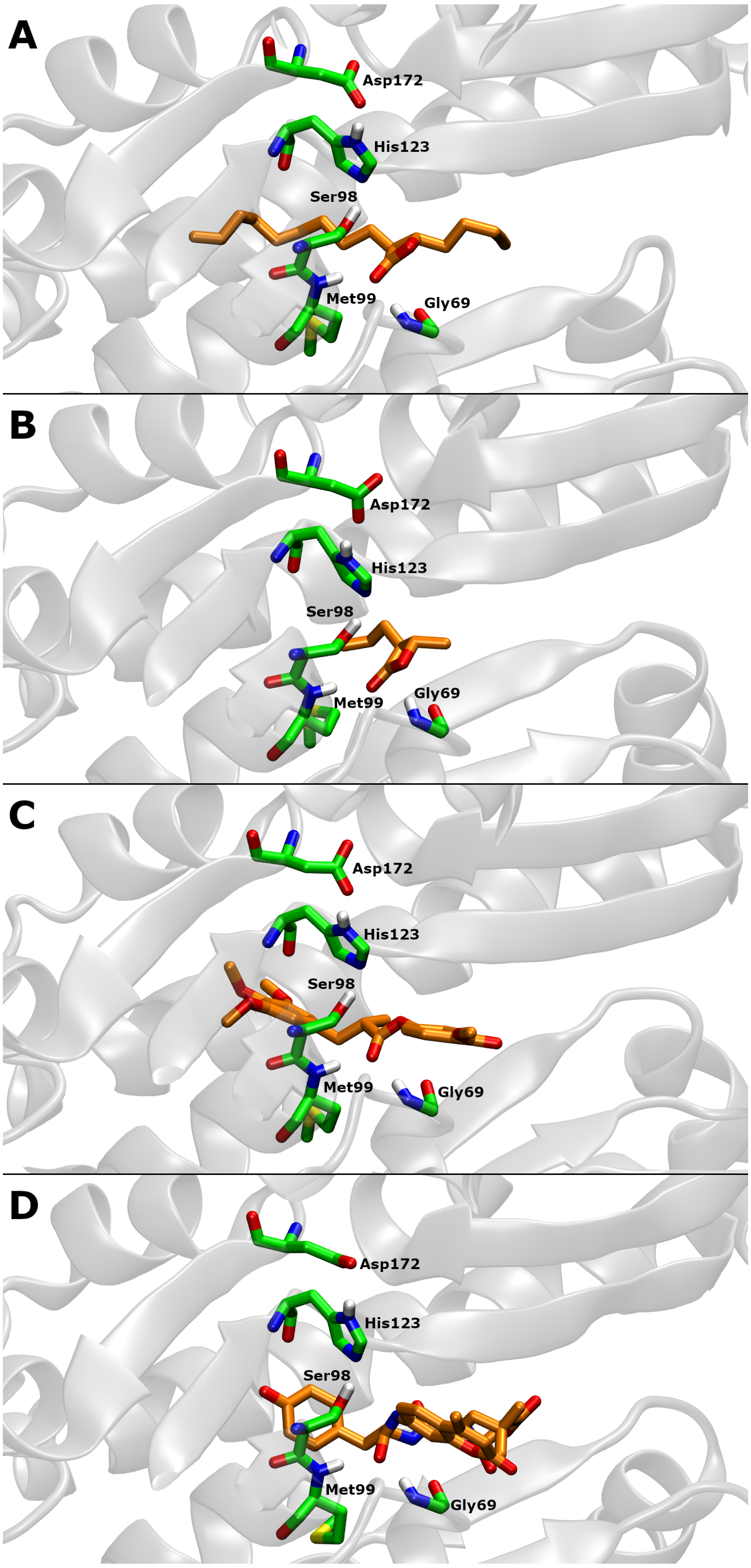


Figure 3-10. Illustration of the binding site of SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) the fluorescent substrate. Structures were extracted from MD simulations as initial structures for QM/MM calculations.

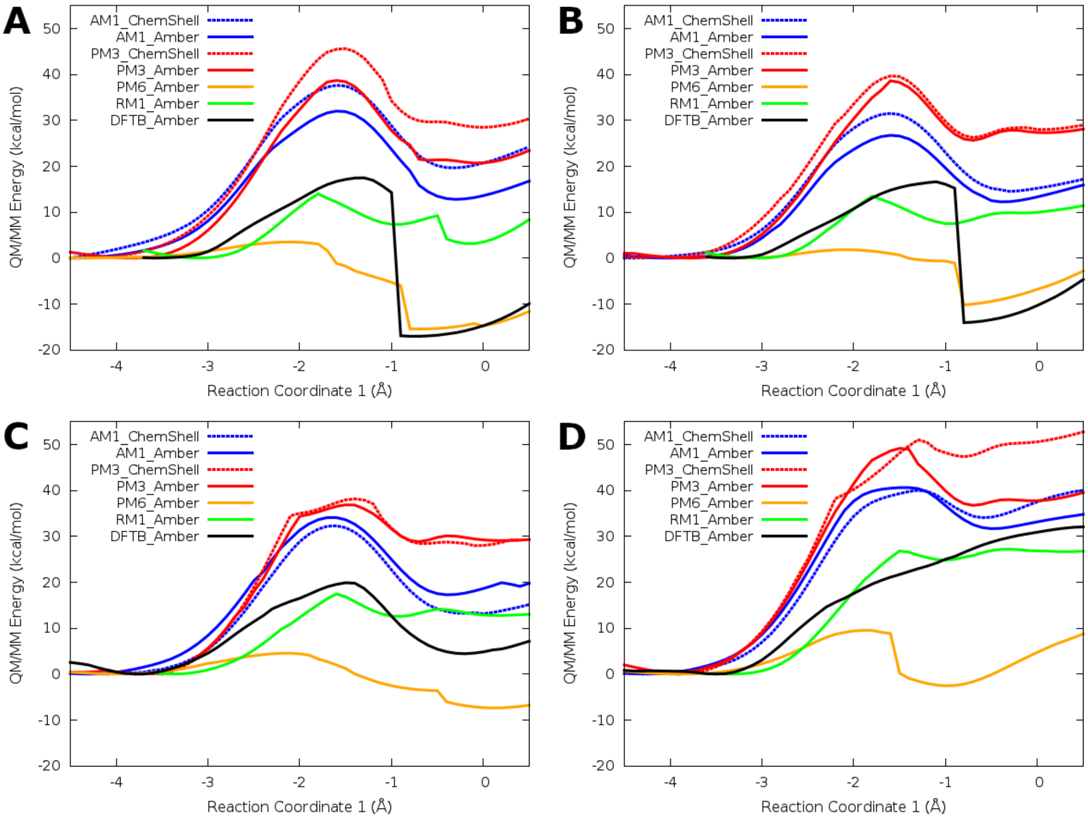


Figure 3-11. Reaction profiles of nucleophilic attack step of the acylation reaction calculated by different semi-empirical QM/MM methods for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

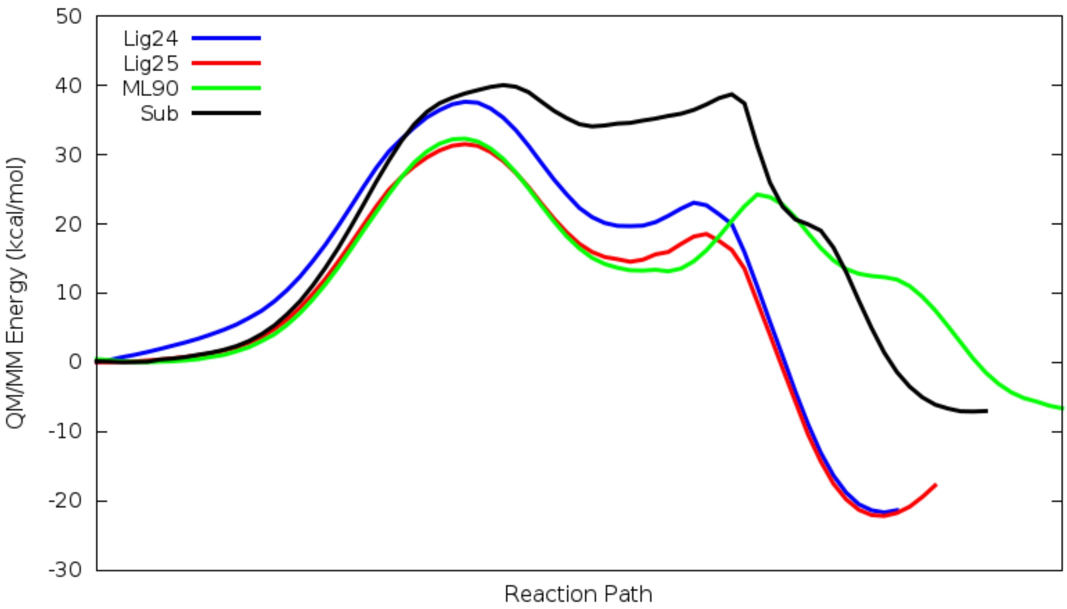


Figure 3-12. Reaction profiles of the acylation reaction calculated by AM1 QM/MM method for SaClpP with ligands.

the fluorescent substrate is a monotonically increasing curve and no transition state can be identified (Figure 3-11D). In the end, smooth reaction profiles for all four ligands were obtained by the AM1 method. Therefore, AM1 is the most solid semi-empirical QM method to simulate the acylation reaction although calculated energy barriers are higher than the experimental values.

Reaction profiles of the entire acylation reaction were further calculated by AM1 QM/MM method for all four ligands (Figure 3-12). Calculated activation energies and energy differences for the acylation step are also listed in the first part of Table 3.4. Our AM1 QM/MM calculation results reveal that the acylation reaction has two separate transition states and one tetrahedral intermediate for all ligands. The first peak of the reaction profile corresponds to the transition state of the nucleophilic attack step and the second one corresponds to the transition state of the proton transfer step. The local minimum in the middle of the profile is corresponding to the tetrahedral intermediate. The reaction profile for the fluorescent substrate has two transition states with similar values of energy barriers (40.0 and 38.7 kcal mol^{-1}). While for the other three ligands, the second transition states have relatively lower energy barriers (18.5–24.2 kcal mol^{-1}). The calculation results also show that the nucleophilic attack step is the rate-limiting step and calculated activation energies are at least 10 kcal mol^{-1} higher than the experimental values. In addition, acylation is an exothermic reaction for all four ligands based on AM1 calculations (ΔE is between -22.2 and -6.7 kcal mol^{-1}).

3.4.2.2 1D DFT calculation

1D DFT reaction profiles of the nucleophilic attack step were calculated for two β-lactones at the BP86/def2-SVP level (Figure 3-13). As we can see from the figure, there is a huge energy drop of roughly 30 kcal mol^{-1} for both ligands. After checking the structures, we found out that the drop is due to proton transfer to the ligand before we simulating the proton transfer step. To avoid this problem, a second reaction coordinate, which is defined by equation (3-2) or equation (3-3), was used to control the process of proton transfer from H123 to the ligand. Therefore, in the following part of this study, all potential energy surfaces were calculated in two-dimensional.

3.4.2.3 2D AM1 calculations

As 2D QM/MM calculations were required to obtain a DFT potential energy surface, 2D AM1 QM/MM calculations were also performed to compare with both 1D AM1 and 2D DFT QM/MM calculations. Four 2D AM1 potential energy surfaces are shown in Figure 3-14 and calculated energy differences and energy barriers are listed in Table

Table 3.4. Summary of QM/MM calculation results for the acylation step

QM method	Ligand	Experimental	$\Delta E^{\neq 1}$ [kcal mol^{-1}]	ΔE^{int} [kcal mol^{-1}]	$\Delta E^{\neq 2}$ [kcal mol^{-1}]	ΔE [kcal mol^{-1}]
AM1-1D	Lig24	20.8	37.6	19.7	23.1	-21.7
	Lig25	20.6	31.5	14.5	18.5	-22.2
	ML90	21.0	32.3	13.1	24.2	-11.7
	Sub	N/A	40.0	34.0	38.7	-7.1
AM1-2D	Lig24	20.8	36.1	18.2	21.5	-22.9
	Lig25	20.6	31.7	15.0	18.9	-22.2
	ML90	21.0	32.3	12.9	22.9	-13.3
	Sub	N/A	39.9	33.9	39.9	-5.8
BP86/ def2-SVP	Lig24	20.8	15.9	N/A	N/A	-12.1
	Lig25	20.6	18.5	N/A	N/A	-9.7
	ML90	21.0	12.2	N/A	N/A	-7.9
	Sub	N/A	13.9	13.5	17.8	-3.3
B3LYP/ def2-TZVP	Lig24	20.8	20.9	N/A	N/A	-12.0
	Lig25	20.6	21.8	N/A	N/A	-10.0
	ML90	21.0	20.3	N/A	N/A	-5.4
	Sub	N/A	22.9	20.8	25.1	-1.3
QM/MM-FEP	Lig24	20.8	14.4	N/A	N/A	-14.5
	Lig25	20.6	20.7	N/A	N/A	-12.3
	ML90	21.0	16.9	N/A	N/A	-8.6
	Sub	N/A	21.8	19.9	25.9	6.3

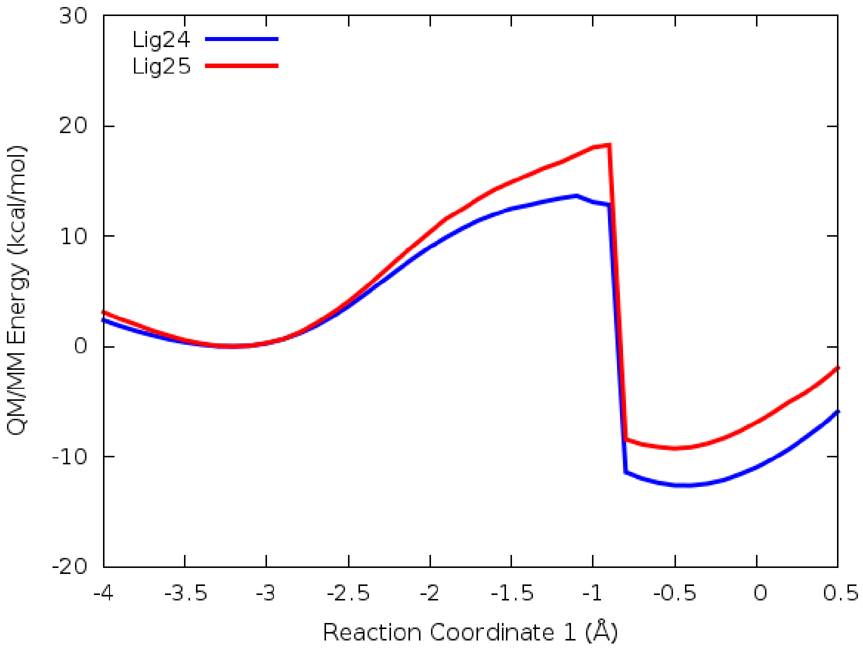


Figure 3-13. Reaction profiles of nucleophilic attack step of the acylation reaction calculated by AM1 QM/MM method for SaClpP with Lig24 and Lig25.

3.4. In general, the results of 2D and 1D AM1 QM/MM calculations are consistent with each other. Most calculated values are almost identical for 1D and 2D calculations and the maximum difference in calculated value between 1D and 2D calculations is only 1.6 kcal mol^{-1}. The difference is partially resulted from the assembling of two calculations (the nucleophilic attack step and the proton transfer step) together to get a 1D reaction profile. The intermediate was restrained by $\xi 1$ for simulating the nucleophilic attack step and during the calculation of the proton transfer step, it was restrained by $\xi 2$. When we assembled two partial calculations together, these two systems with different restraints were treated as the same. One big advantage of 2D calculations is to avoid this type of error because two different restraints were applied for all grid points. These contour pictures of 2D potential energy surfaces share some common features. The local minimum in the bottom left corner is the reactant state. The saddle point at the bottom is the transition state of the nucleophilic attack step. In the bottom right part, there is a local minimum, which corresponds to the tetrahedral intermediate. On the right of the surface, the second saddle point is corresponding to the transition state of the proton transfer step. In the end, the local minimum in the top right part is the product state.

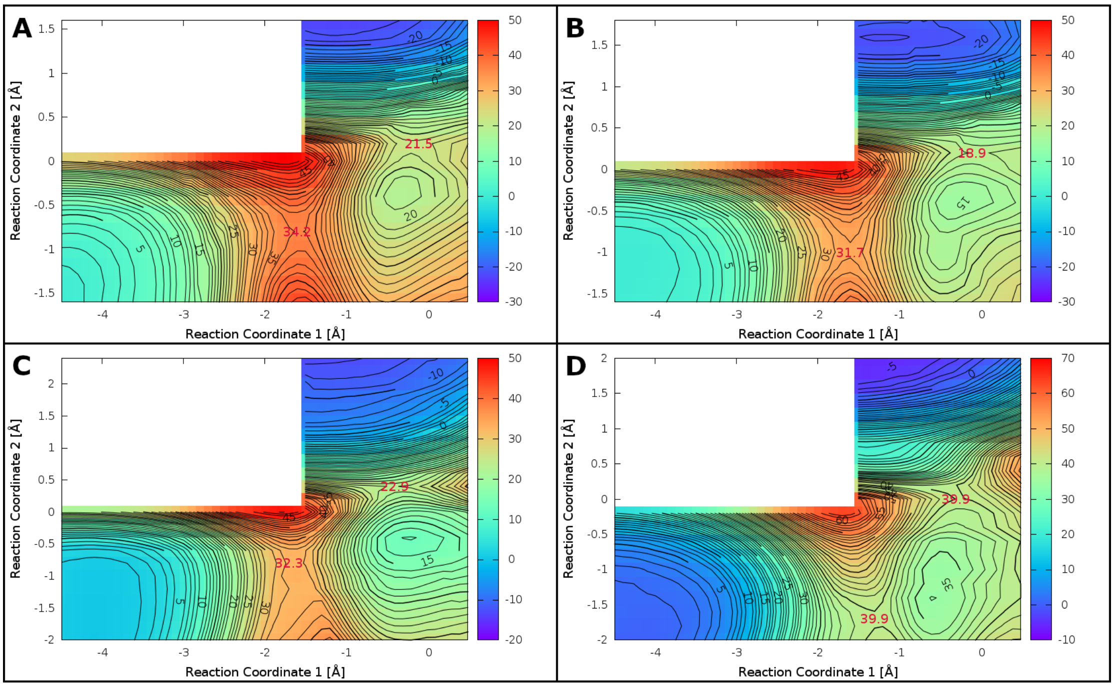


Figure 3-14. Potential energy surfaces of the acylation reaction calculated by AM1 QM/MM method for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

3.4.2.4 2D DFT calculations

For DFT QM/MM calculations, energy minimizations were performed at the BP86/def2-SVP level and single point energies were further calculated at the B3LYP/def2-TZVP level on the optimized structures. The idea of performing geometry optimization at a relatively low level of theory and then calculating single point energies on the optimized structure at a high level of theory is a common strategy in computational chemistry. This strategy has a good balance between the computational cost and the calculation accuracy. Calculated potential energy surfaces at both BP86 and B3LYP levels for all ligands are shown in Figure 3-15 and Figure 3-16. The top left corner for all two-dimensional potential energy surfaces was not calculated due to unrealistic high energies in this area. For each ligand, the BP86 PES and the B3LYP PES have a similar pattern. Only the positions of the reaction state, the transition state, the intermediate and the product state shift slightly after single point calculations. Our DFT QM/MM calculations reveal that there is only one transition state of acylation for β-lactones and the phenyl ester (Figure 3-15A-C and Figure 3-16A-C) in comparison with two transition states for the fluorescent substrate (Figure 3-15D and Figure 3-16D). The reason for having only one transition state for β-lactones is probably because that the opening of the four-membered ring with the release of strain compensates the energy required for proton transfer. For the phenyl ester, the leaving group is a phenoxide ion which is a better leaving group compared to amide for the fluorescent substrate. This might explain why the acylation reaction for the phenyl ester also takes a one-step mechanism. The calculation results of the fluorescent substrate show a two-step mechanism, which agrees with other theoretical studies [73, 75–78, 80, 82–84].

The main difference between BP86 and B3LYP calculations is that B3LYP calculated values are higher than BP86 calculated values in all cases (Table 3.4). In general, the B3LYP functional performs better than the BP86 functional and def2-TZVP is a larger basis set than the def2-SVP basis set. Thus we believe that results calculated at the B3LYP/def2-TZVP level are more reliable and we will use these results to compare with the experimental values and analyze potential energy surfaces, the reaction mechanism, and important characteristic structures. Overall, activation energies calculated at the B3LYP level are in good agreement with the experimental values (Table 3.4). Although phenyl ester ML90 has a different structure compared to the β-lactones, the calculated energy barrier is very close to the values for β-lactones. The energy barrier of the acylation reaction for the fluorescent substrate is 25.1 kcal mol^{-1}, which is 4–5 kcal mol^{-1} higher than the energy barriers for the other three ligands. This is probably because the peptide bond is relatively difficult to cleavage compared to the ester bond. Similar to

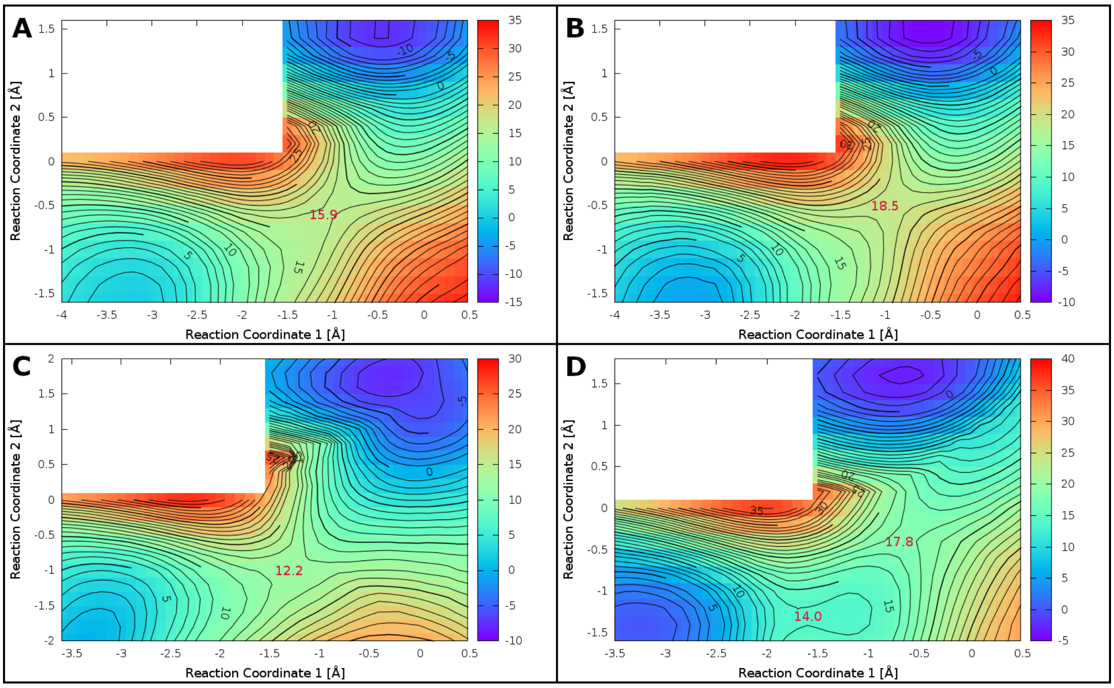


Figure 3-15. Potential energy surfaces of the acylation reaction calculated by DFT QM/MM method at the BP86/def2-SVP level for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

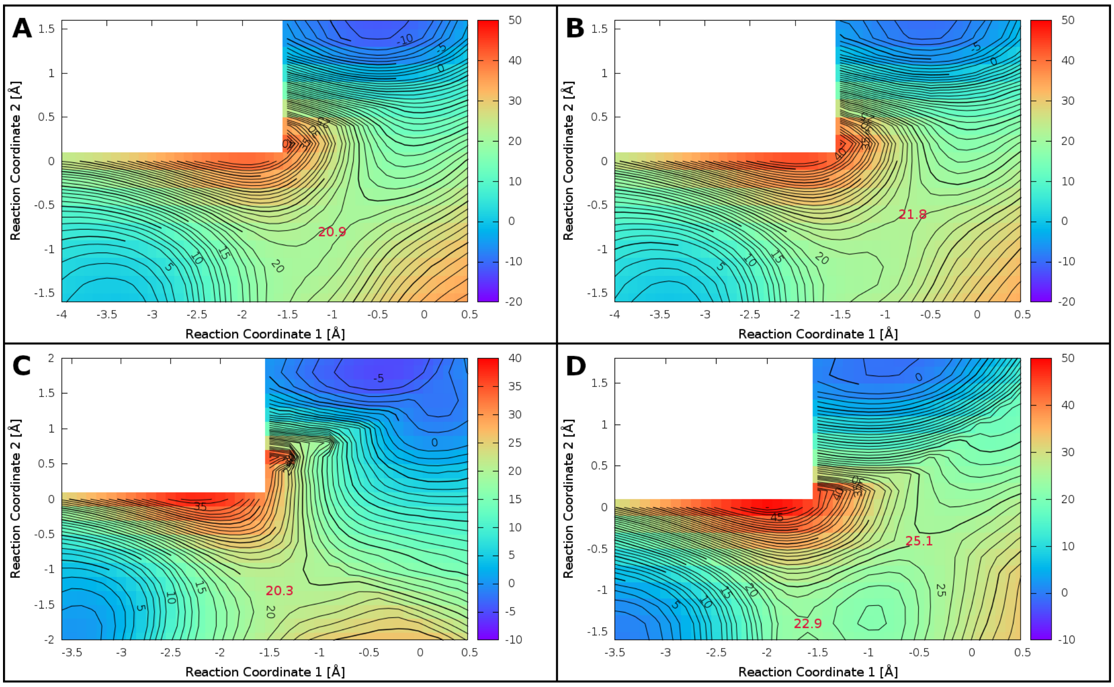


Figure 3-16. Potential energy surfaces of the acylation reaction calculated by DFT QM/MM method at the B3LYP/def2-TZVP level for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

AM1 results, DFT calculations also indicate that acylation reaction is an exothermic reaction.

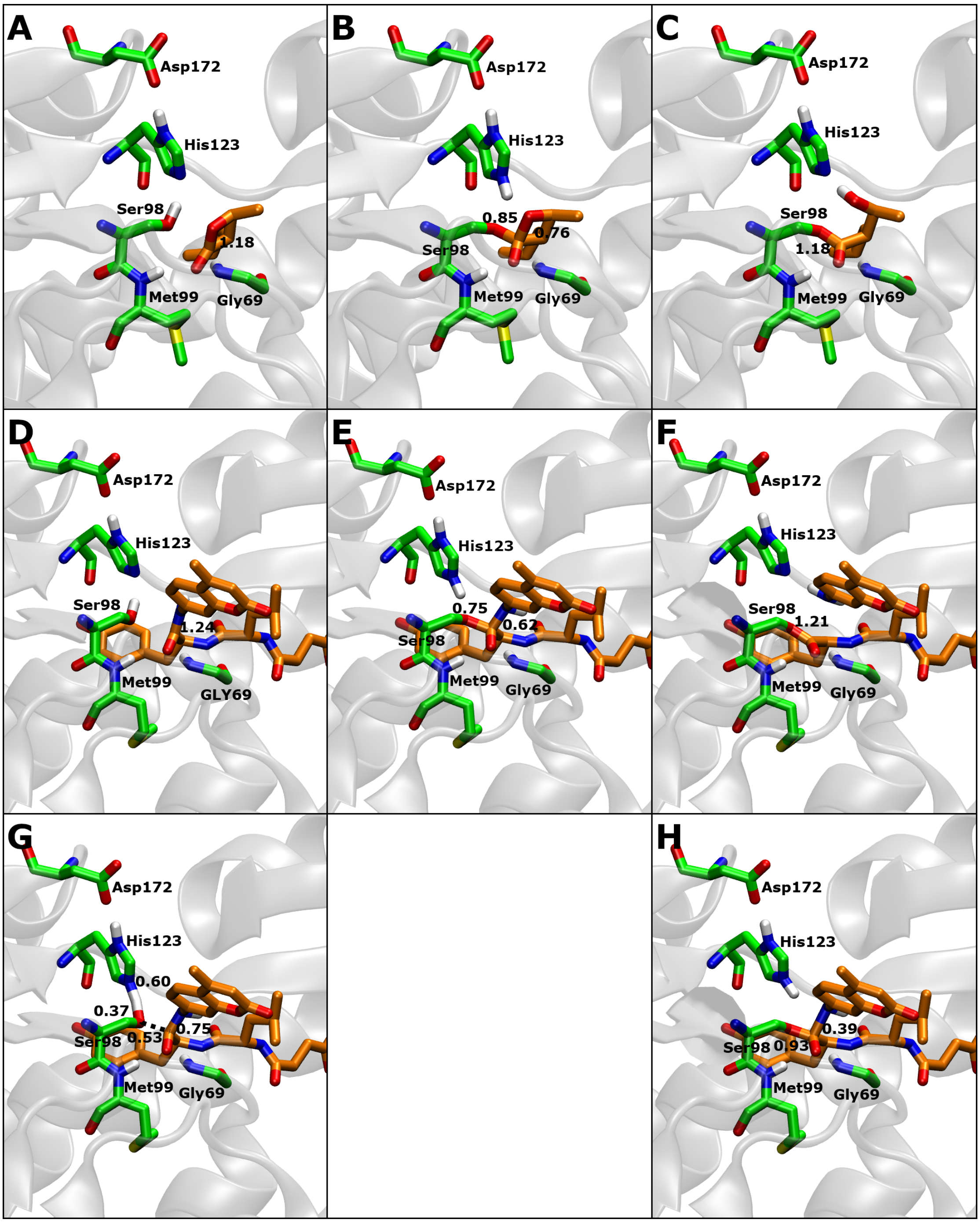


Figure 3-17. Illustration of characteristic structures and some important bond orders for the acylation reaction of SaClpP. Carbon atoms of protein are shown in green and carbon atoms of ligand are shown in orange. (A) Reactant state, (B) transition state and (C) product state of the acylation reaction for Lig25. (D) Reactant state, (E) tetrahedral intermediate, (F) product state, (G) the first transition state and (F) the second transition state of the acylation reaction for the fluorescent substrate.

Next, we will discuss the reaction mechanism and structural features of the acylation reaction based on DFT calculations. Some characteristic optimized structures calculated at the B3LYP level for the acylation step are shown in Figure 3-17. Mayer's bond orders (BOs) were also calculated from the density matrix elements [234, 235] using an in-house python script to analyze the progress of the acylation reaction. We use Lig25 and the fluorescent substrate as examples to illustrate reaction mechanisms and characteristic structures of the acylation reaction. For β-lactones, the reaction starts from the reactant state (Figure 3-17A), which is very similar to the initial structure (Figure 3-10B) extracted from the MD simulation. The bond order of the peptide bond at the reactant state is 1.18 showing a partial double bond character, and there is no bond (BO = 0.02) between S98 and the ligand. The nucleophilic attack of S98 to the carbonyl carbon of the ligand is concerted with the proton transfer from S98 to H123. H123 serves as a general base to accept a proton from S98. The negatively charged transition state is stabilized by the backbone NH groups of G69 and M99 (Figure 3-17B). At the tetrahedral transition state, a bond between H123 and HG of S98 is formed (BO = 0.80) and D172 stabilizes the positively charged H123. The bond order of the ester bond decreases to 0.76 showing that the bond is partially broken, while a bond between S98 and the ligand is partially formed (BO = 0.85). After that, the ester bond of the lactone is fully broken (BO = 0.04) with the opening of the four-membered ring (Figure 3-17C). On the other hand, the fluorescent substrate shows a similar reaction path compared to β-lactones. The major difference is that there is a tetrahedral intermediate in the acylation reaction (Figure 3-16E). At this state, the proton is bonded to H123 (BO = 0.77). The peptide bond is partially broken (BO = 0.62) compared to the reactant state (Figure 3-16D, E). There are two transition states of the reaction. At the first transition state (Figure 3-16G), S98 is approaching the substrate and the proton is roughly in the middle of S98 and H123. The bond orders of the proton with the oxygen atom of S98 and nitrogen atom of H123 are 0.37 and 0.60, respectively. The bond order of the peptide bond decreases to 0.75 showing that the π conjugation of the bond is disrupted, while a bond between S98 and the fluorescent substrate is partially formed (BO = 0.53). At the second transition state (Figure 3-16H), the ester bond is almost formed (BO = 0.93) and the peptide bond (BO = 0.39) is nearly broken. Although it is the transition state of the proton transfer step, the proton is still bonded to H123 (BO = 0.85). After this transition state, the scissile bond is fully broken (BO = 0.02) to form a covalently bound complex with the release of the leaving group (Figure 3-16F). ML90 shows a slightly different reaction mechanism compared to the other three ligands. Although ML90 takes a one-step reaction mechanism for the acylation reaction (Figure 3-16C), the reaction releases a phenoxide ion as a leaving group instead of a ring opening for β-lactones. In addition, all transition states for β-lactones and the phenyl ester and the intermediate for the fluorescent substrate show fully protonated H123. This clearly indicates that proton transfer to the ligand happens after the nucleophilic attack of serine for all four ligands.

3.4.2.5 2D DFTB calculations

Since 1D DFTB reaction profiles are similar to DFT reaction profiles for two β-lactones, 2D DFTB potential energy surfaces were also calculated. For β-lactones, 2D DFTB potential energy surfaces are, in general, similar to BP86 and B3LYP potential energy surfaces (Figures 3-15A, B, 3-16A, B, and 3-18A, B). Furthermore, DFTB calculated energy barriers are between BP86 values and B3LYP values. The PES for ML90 (Figure 3-18C) shows a second transition state with an energy barrier of 5.4 kcal mol^{-1} at (-0.2 Å, 0.8 Å). As for the fluorescent substrate, although there are also two transition states on the DFTB PES (Figure 3-18D), the positions of these two transition states and corresponding energy barriers differ from the results of DFT calculations. The 2D DFTB calculations indicate that the DFTB method cannot correctly describe all ClpP systems.

3.4.2.6 QM/MM FEP calculations

QM/MM FEP calculations for the acylation step were also performed to compare with free energy calculations for the deacylation step (see details in section 3.5.5). Calculated free energies and free energy differences are listed in Table 3.4. For Lig25 and the fluorescent substrate, calculated free energy barriers are very close to energy barriers calculated at the B3LYP level. While for Lig24 and ML90, calculated free energy

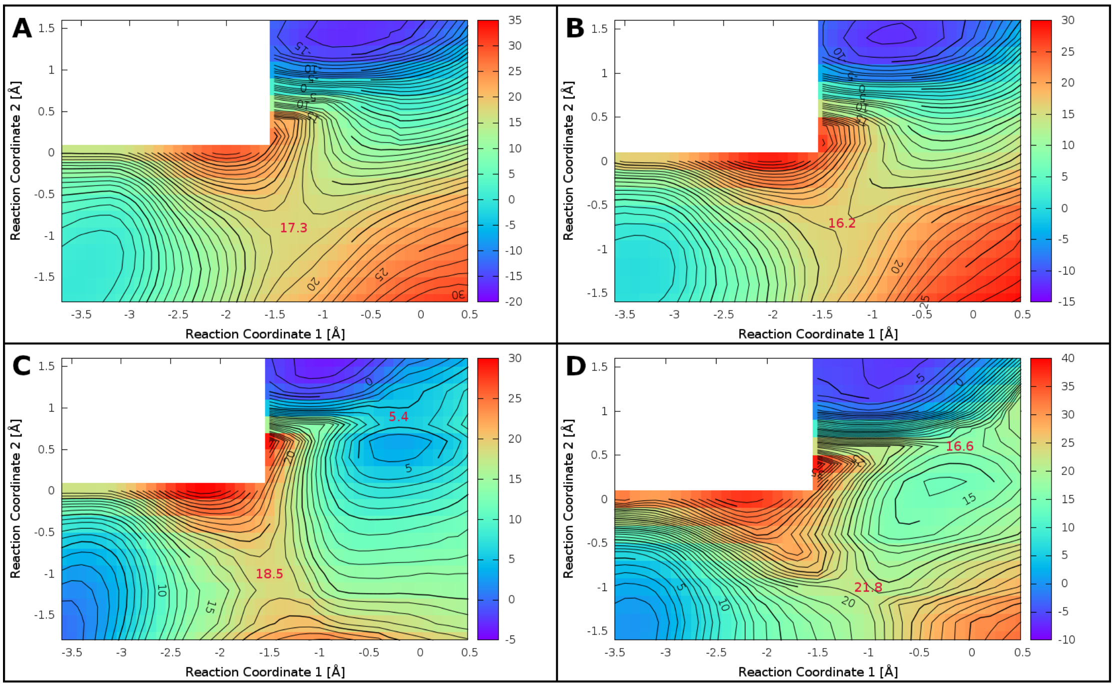


Figure 3-18. Potential energy surfaces of the acylation reaction calculated by DFTB QM/MM method for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

barriers are 6.5 kcal mol-1 and 3.4 kcal mol-1 lower than corresponding energy barriers, respectively. The similar effect was also observed in calculations of the deacylation step (see section 3.4.3.4). Although our calculated free energy barrier agree with the experimental value for Lig25 very well, our calculations slightly underestimate free energy barriers for Lig24 and ML90. One possible reason is the missing of the dynamics of the QM part in our calculations. Lig24 has a long aliphatic side chain and ML90 has a large trimethoxyphenyl group in the binding pocket. Compared to Lig25, these two ligands are relatively big. Not including dynamics effect of the ligand in our calculations might results in the underestimation of our calculated free energy barriers for Lig24 and ML90. On the other hand, although the fluorescent substrate also has a relatively big tyrosine residue in the binding pocket, it is a peptide like substrate. Thus, above mentioned explanation may not applied to the fluorescent substrate.

3.4.3 Deacylation step

Deacylation reactions of SaClpP were simulated with four different setups (see section 3.3.2). By iterative calculations and comparisons with experimental values, we continuously improved our simulation protocol for the deacylation step. Our calculation results with different setups will be summarized in the following sections.

3.4.3.1 Setup 1

Deacylation calculations with setup 1 used the same calculation protocol as acylation simulations. Calculated energy barriers and energy differences of the deacylation reaction with setup 1 for all ligands are summarized in Table 3.5. 2D potential energy surfaces calculated at the B3LYP/def2-TZVP level are shown in Figure 3-19. Similar to the simulations of the acylation reaction, B3LYP calculated values are always higher than

Table 3.5. Calculated energy barriers and energy differences for the deacylation step of SaClpP with Setup 1

QM method	Ligand	Experimental	$\Delta E^{\neq 1}$ [kcal mol-1]	ΔE^{int} [kcal mol-1]	$\Delta E^{\neq 2}$ [kcal mol-1]	ΔE [kcal mol-1]
BP86/ def2-SVP	Lig24	23.7	28.4	N/A	N/A	5.3
	Lig25	23.9	42.5	N/A	N/A	16.0
	ML90	24.0	27.4	N/A	N/A	5.3
	Sub	N/A	22.7	N/A	N/A	8.6
B3LYP/ def2-TZVP	Lig24	23.7	35.5	N/A	N/A	6.8
	Lig25	23.9	48.5	N/A	N/A	17.8
	ML90	24.0	36.7	N/A	N/A	8.9
	Sub	N/A	27.5	22.8	24.0	12.5

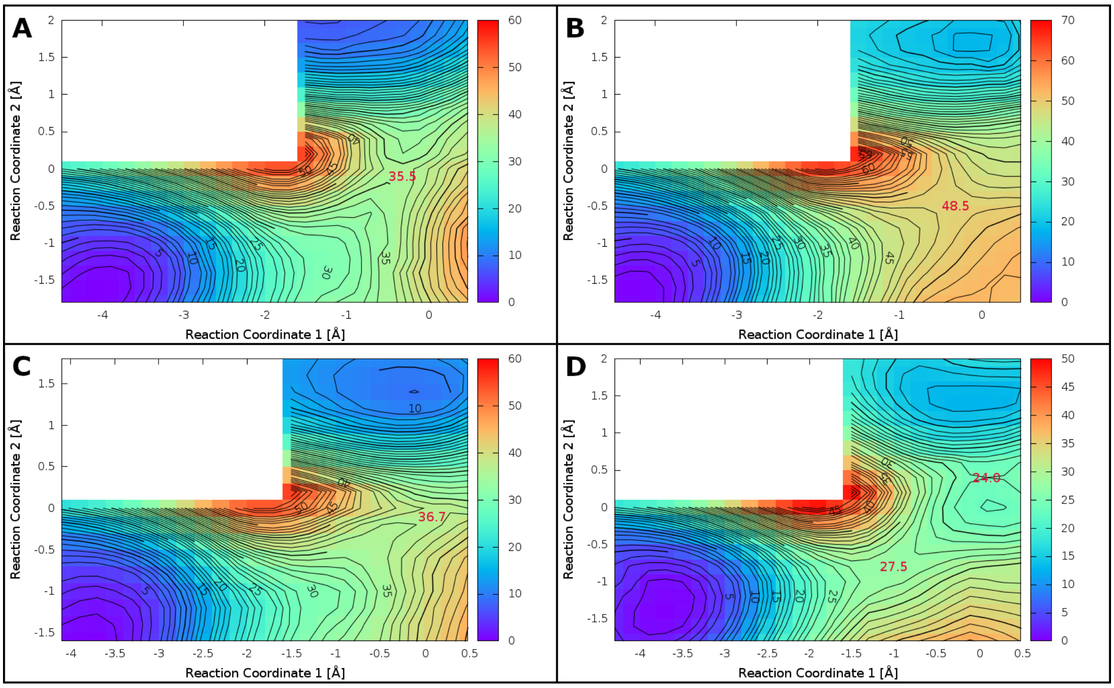


Figure 3-19. Potential energy surfaces of deacylation reaction calculated by DFT QM/MM method at the B3LYP/def2-TZVP level with Setup 1 for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

BP86 calculated values for both energy barrier and ΔE. The calculation results with setup 1 indicate that the deacylation is a one-step reaction for β-lactones and the phenyl ester. Nevertheless, two DFT functionals give different reaction mechanisms of the deacylation reaction for the fluorescent substrate. According to the results calculated at the BP86 level, there is only one transition state for the fluorescent substrate like the other three ligands. On the other hand, B3LYP calculations show that there is an intermediate and a second transition state of the deacylation reaction. But the energy difference between the intermediate and the second transition state is only 1.2 kcal mol^{-1}, which means this intermediate is very shallow on the potential energy surface. Although this setup works perfectly for the acylation step, it obviously does not work very well for the deacylation step. As can be seen from Table 3.5, B3LYP calculated energy barriers are at least 10 kcal mol^{-1} higher than the experimental values. Especially the calculated value for Lig25 is 48.5 kcal mol^{-1}, which roughly doubles the experimental value. The reason for the overestimation of the energy barrier is probably due to some missing factors in our simulations.

3.4.3.2 Setup 2

After carefully checking optimized structures of both acylation and deacylation simulations, we thought the missing of water dynamics in simulations might lead to

Table 3.6. Calculated energy barriers and energy differences for the deacylation step of SaClpP with Setup 2

QM method	Ligand	Experimental	$\Delta E^{\neq 1}$ [kcal mol^{-1}]	ΔE^{int} [kcal mol^{-1}]	$\Delta E^{\neq 2}$ [kcal mol^{-1}]	ΔE [kcal mol^{-1}]
BP86/ def2-SVP	Lig24	23.7	28.5	N/A	N/A	7.4
	Lig25	23.9	28.4	N/A	N/A	10.9
	ML90	24.0	22.9	N/A	N/A	7.6
	Sub	N.A.	23.4	N/A	N/A	9.9
B3LYP/ def2-TZVP	Lig24	23.7	36.9	N/A	N/A	9.8
	Lig25	23.9	34.8	N/A	N/A	14.2
	ML90	24.0	30.0	29.9	30.4	12.2
	Sub	N/A	27.1	23.4	24.7	12.2

overestimation of energy barriers. In order to verify this assumption, we simulated the deacylation step again. A 20ps of molecular dynamics was performed for each ligand at the transition state (the first transition state for the fluorescent substrate) from the potential energy surface with setup1 (Figure 3-19). After performing this short MD, new potential energy surfaces (Figure 3-20) were obtained by restrained energy minimizations starting from the end structures of the short MD simulations. Calculated energy barriers and energy differences of the simulations with setup 2 are listed in Table 3.6. It can be noticed that calculated energy barriers are reduced for Lig25 and ML90.

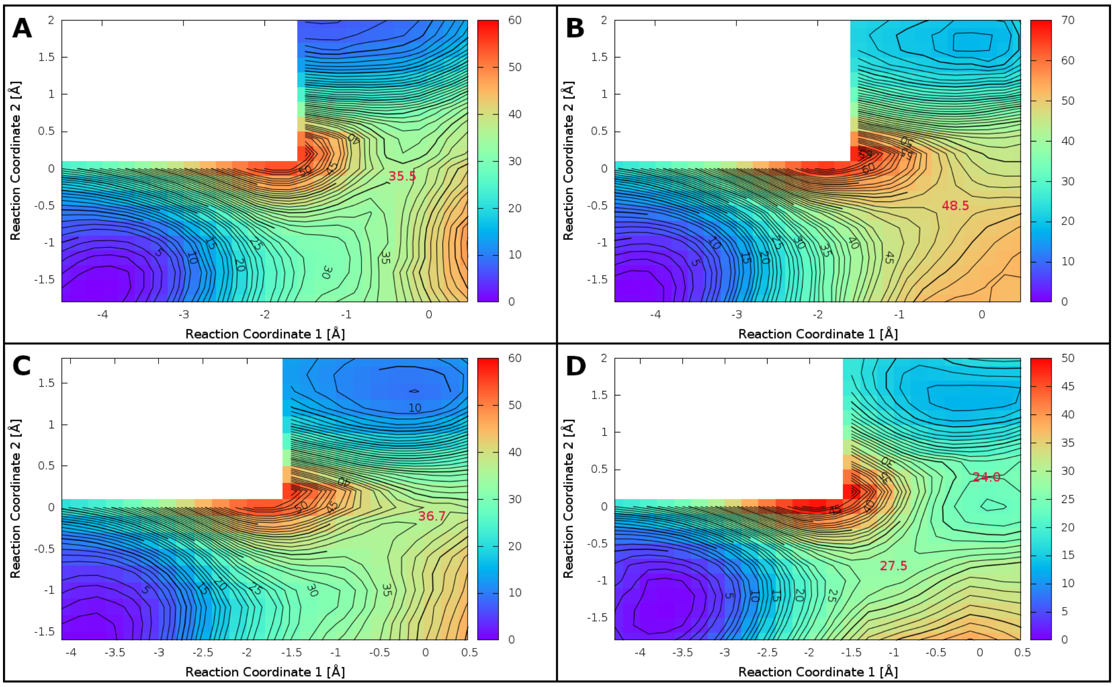


Figure 3-20. Potential energy surfaces of deacylation reaction calculated by DFT QM/MM method at the B3LYP/def2-TZVP level with Setup 2 for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

Especially the energy barrier of the deacylation reaction for Lig25 calculated at the B3LYP level is reduced by 13.7 kcal mol^{-1}. Our calculation results clearly indicate that the positions of water molecules will affect the calculated energy barrier. The network of water molecules changes a lot after the short MD simulation while atoms of the catalytic triad only move slightly (Figure 3-21). This explains that the decrease of the energy barrier after the short MD simulation is mainly due to the adaptation of surrounding water molecules to the change of the reaction center. For Lig24 and the fluorescent substrate, however, the decrease of energy barriers was not observed. This is probably because in these cases, surrounding water molecules were already in good positions to stabilize the reaction center. With this setup, BP86 calculations still imply that the deacylation step has a one-step reaction mechanism for ML90. However, the B3LYP calculations reveal a two-step mechanism for the ligand. Similar to the calculations with setup 1, the intermediates are very shallow on the potential energy surfaces. The energy differences between intermediates and transition states with the lower energy are 0.1 and 1.3 kcal mol^{-1} for ML90 and the fluorescent substrate respectively. The calculations with setup 2 prove that the potential energies of systems around the transition state are very sensitive to the positions of surrounding water molecules around the reactive water molecule.

3.4.3.3 Setup 3

Another factor that might influence calculation results is the treatment of the QM–MM boundary. In our current model, non-bonded interactions between the reactive water molecule and its surrounding water molecules are dealt by the QM–MM boundary. It is not a big problem for the acylation simulation due to no water molecules directly involved in the reaction. However, for the deacylation reaction, the structure and the

Table 3.7. Calculated energy barriers and energy differences for the deacylation step of SaClpP with Setup 3

QM method	Ligand	Experimental	$\Delta E^{\neq 1}$ [kcal mol^{-1}]	ΔE^{int} [kcal mol^{-1}]	$\Delta E^{\neq 2}$ [kcal mol^{-1}]	ΔE [kcal mol^{-1}]
BP86/ def2-SVP	Lig24	23.7	22.8	22.0	22.6	11.3
	Lig25	23.9	26.2	N/A	N/A	13.6
	ML90	24.0	19.7	N/A	N/A	5.0
	Sub	N/A	21.4	19.3	19.9	13.2
B3LYP/ def2-TZVP	Lig24	23.7	32.2	31.8	33.0	12.0
	Lig25	23.9	31.0	N/A	N/A	14.1
	ML90	24.0	27.2	N/A	N/A	6.8
	Sub	N/A	27.7	25.9	27.5	13.9

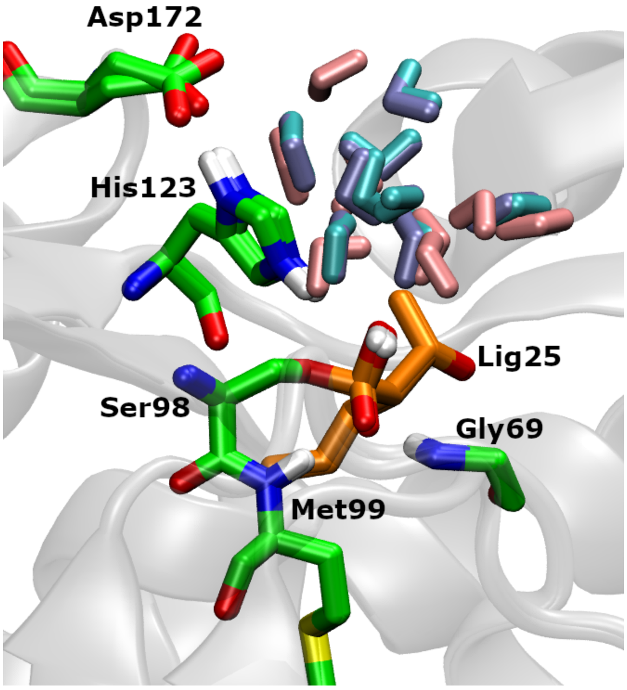


Figure 3-21. Superimposed structures of transition states calculated at B3LYP level for Lig25. Surrounding water molecules are shown in pink (Setup 1), cyan (Setup 2) and iceblue (Setup 3) respectively.

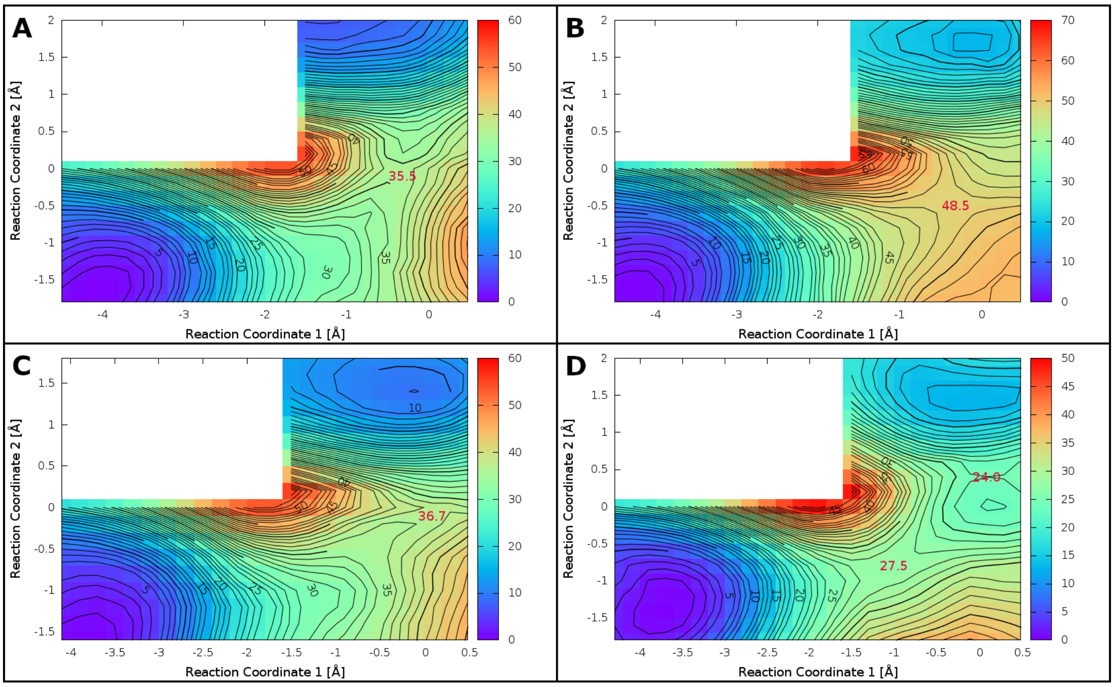


Figure 3-22. Potential energy surfaces of deacylation reaction calculated by DFT QM/MM method at the B3LYP/def2-TZVP level with Setup 3 for SaClpP with (A) Lig24, (B) Lig25, (C) ML90 and (D) fluorescent substrate.

charge distribution of the reactive water molecule changes a lot from the reactant state to the transition state. It may cause extra errors in calculating the energy barrier. This problem can be partially solved by increasing the number of QM water molecules since the QM–MM boundary only changes slightly from the reactant state to the transition state in this case. QM/MM calculations with more QM water molecules were carried out. New potential energy surfaces were obtained (Figure 3-22). In general, these surfaces share a similar pattern: a reactant state is in the bottom left corner; (a) transition state(s) and an intermediate stay on the right and a product state is in the part of the top right. It is also noticeable that after adoption of this new setup, positions of QM water molecules only move slightly (Figure 3-21). Calculated energy barriers and energy differences with setup 3 are summarized in Table 3.7. Except for the energy barrier calculated at the B3LYP level for the fluorescent substrate, all other calculated energy barriers with more QM water molecules are lower than the values calculated with setup 2. This means the strategy of including more water molecules into the QM region can improve our calculation results. With this setup, BP86 and B3LYP give consistent results for all ligands with respect to the reaction mechanism. Compared to previous calculations, this time Lig24 shows a two-step mechanism while ML90 follows a one-step mechanism. Energy differences between the intermediate and the transition state with lower energy for Lig24 and the fluorescent substrate are 0.4 kcal mol^{-1} and 1.6 kcal mol^{-1} respectively. Since many other factors might affect calculated energies, it seems there is no one simple factor which determines whether the deacylation reaction follows a one-step mechanism or a two-step mechanism.

3.4.3.4 Setup 4

By partially taking water dynamics into account, the calculations with setup 3 improved a lot in comparison with the calculations with setup 1. However, calculated energy barriers are still higher than the experimental values by 3.2 to 9.3 kcal mol^{-1}. This means further including dynamics of the system into our simulations is necessary for accurately modelling the deacylation step. Since Bohr-Oppenheimer QM/MM MD is still unaffordable under our conditions and semi-empirical methods cannot provide qualitatively correct results for all systems in the simulations of the acylation step (see detailed discussion in section 3.5.1), QM/MM FEP is a promising method to improve our calculations. Due to the restriction of computational resources, we only performed one-dimensional calculations, which means we have to choose a path connecting the reactant state to the state of interest. Theoretically, different paths should give the same results. We performed a test simulation which shows that the free energy difference between two paths is less than 0.5 kcal mol^{-1} (Figure 3-23). It is noticeable that the curve of free energy differences is not the reaction profile. We can see that both red curves have a similar trend and the curve of free energy differences calculated by QM/MM FEP is below the curve of static QM/MM minimized energies calculated at the B3LYP level with setup 3. This feature holds true for test calculations as well (blue curves in Figure 3-23). When QM/MM FEP calculations are not along the path of obtaining the potential energy surface, the trend of the free energy curve is slightly different compared to its corresponding energy curve (blue lines in Figure 3-23). In this study, we always use the same path as obtaining the potential energy surface to perform QM/MM FEP calculations.

Detailed calculated free energy barriers and free energy differences of deacylation reactions are shown in Figure 3-24. Calculated free energy barriers for all ligands are in the same range (23.5–27.0 kcal mol^{-1}). This is because that the key reaction of the deacylation step is the same for all four ligands, namely the breakage of the ester bond. Our calculation results agree with experimental data very well. The differences between calculated free energy barriers and experimental values for Lig24, Lig25, and ML90 are -0.2, 3.1, and 0.8 kcal mol^{-1} respectively (Table 3.8). The final difference between the experimental value and the calculated value is probably due to the frozen QM region in QM/MM FEP calculations. The dynamics of the QM part are still missing in our calculations.

Free energy barriers of the deacylation step for β-lactones and the phenyl ester are higher than those of the acylation step (Table 3.8). This means deacylation is the rate-determining step in the entire hydrolysis process for these ligands. Thus, the acyl-enzyme for these ligands can be accumulated and their deacylation process can be measured

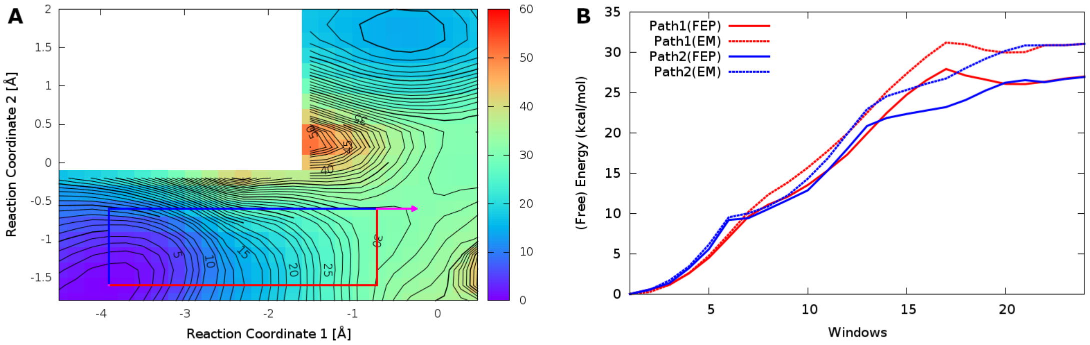


Figure 3-23. (A) Two paths connecting the reactant state and the transition state on the potential energy surface of the deacylation reaction for SaClpP with Lig25 calculated at the B3LYP/def2-TZVP level. (B) Free energy differences (solid lines) calculated by QM/MM FEP and energy differences (dashed lines) calculated by single point QM/MM calculations along two different paths. The scan path is shown in red and another path is shown in blue.

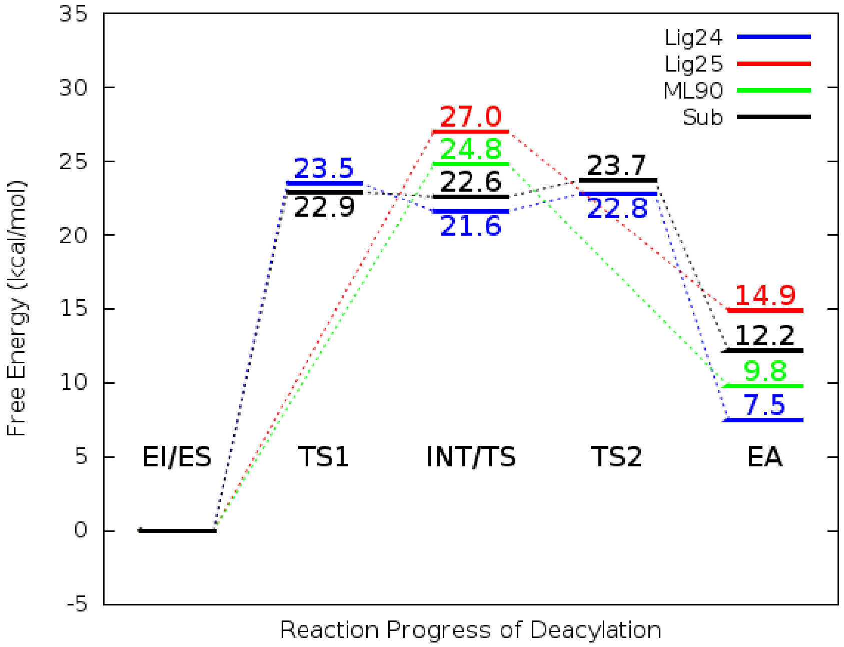


Figure 3-24. Free energy profiles calculated by QM/MM-FEP for the deacylation step of SaClpP (unit: kcal mol^{-1})

by mass spectrometry. The relatively low energy barriers of the acylation reaction for β-lactones can be attributed to the high strain of the four-membered ring. For the phenyl ester, the low energy barrier of the acylation reaction might be resulted from the fact that the phenoxide ion is a good leaving group that promotes the reaction. On the contrary, the calculated free energy barrier of the deacylation reaction for the fluorescent substrate (23.7 kcal mol^{-1}) is lower than the energy barrier of the acylation reaction (25.1 kcal mol^{-1}), which means the acyl-enzyme complex cannot be accumulated. This result matches the experimental finding that no covalently bound complex can be detected by mass spectrometry and can be explained by the ability of the leaving group. For the acylation reaction, the leaving group of the fluorescent substrate is the amide ion which is worse than the alkoxide ion for the deacylation step. In addition, our calculation results also show that free energy differences for all four ligands are positive, which means the deacylation process is endothermic. Actually, this might not be true because our simulations end up with the formation of the carboxylic acid and the further dissociation of the carboxylic acid is not included in our model.

Some characteristic structures calculated by at the B3LYP level for the deacylation step are shown in Figure 3-25. Mayer's bond orders (BOs) were also calculated [234, 235]. In general, the reaction mechanism of the deacylation step is quite similar to the acylation step. The main difference is the nucleophile. The water molecule is a nucleophile for the deacylation reaction instead of S98 for the acylation reaction. Taking

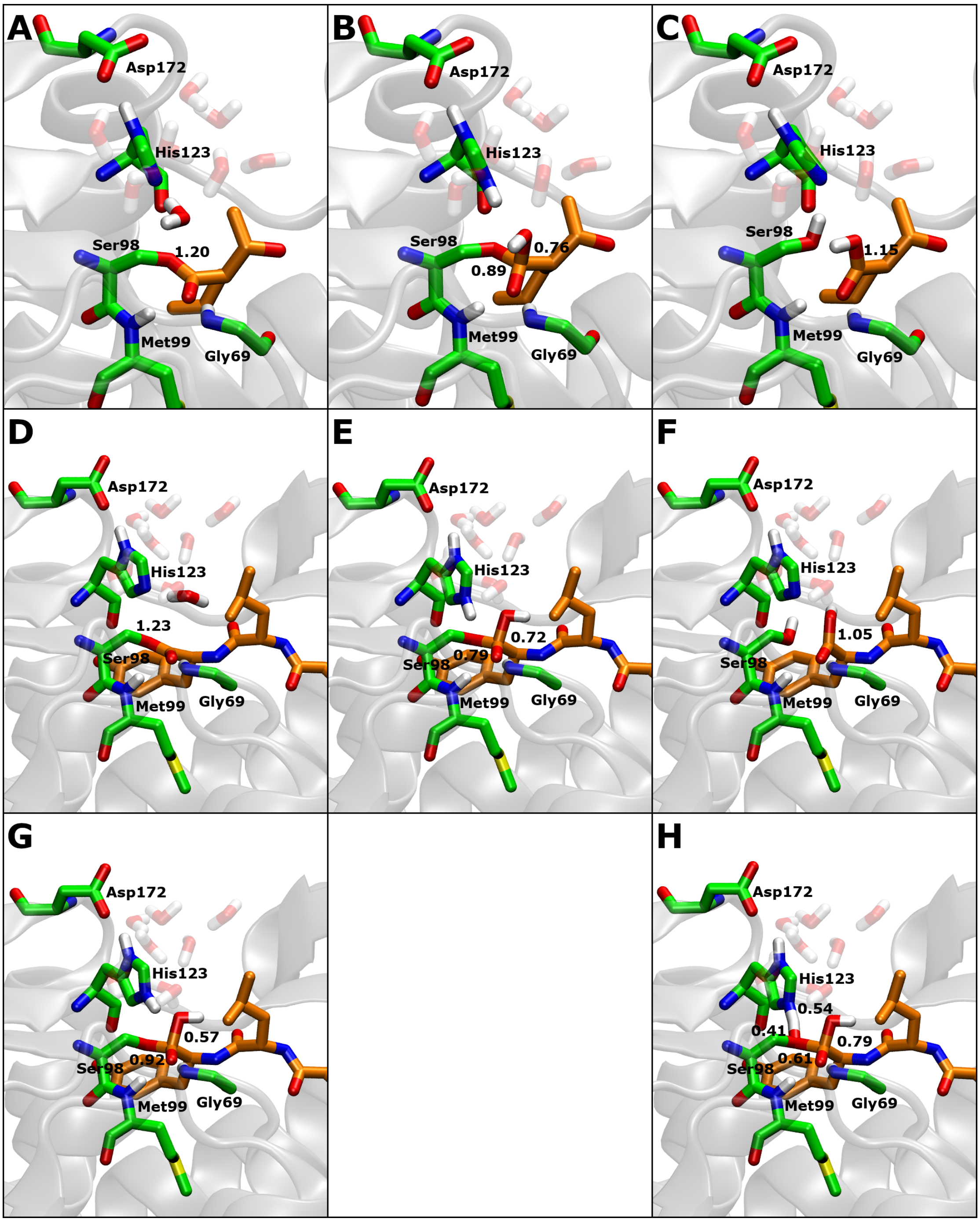


Figure 3-25. Illustration of characteristic structures and some important bond orders for the deacylation reaction of SaClpP. Carbon atoms of protein are shown in green and carbon atoms of ligand are shown in orange. (A) Reactant state, (B) transition state and (C) product state of the deacylation reaction for Lig25. (D) Reactant state, (E) tetrahedral intermediate, (F) product state, (G) the first transition state and (F) the second transition state of the deacylation reaction for the fluorescent substrate.

Lig25 as an example, the nucleophilic attack of the water molecule to the carbon atom of the ester group is also concerted with the proton transfer from the water molecule to H123. The negatively charged transition state is again stabilized by the backbone NH groups of G69 and M99 (Figure 3-25B). H123 serves as a general base to accept a proton from the water molecule and D172 stabilizes protonated H123 as well. At the transition state for Lig25 (Figure 3-25B), the ester bond is partially broken (BO = 0.89). In the end, the ester bond is broken (BO = 0.07) and the ligand is fully hydrolyzed to a carboxylic acid (Figure 3-25C). For the fluorescent substrate, the deacylation step also involves the breakage of the ester bond. The final hydrolyzed product for the fluorescent substrate is also a carboxylic acid (Figure 3-25F). Compared to Lig25, the major difference is that there is an tetrahedal intermediate. At this state, the ester bond starts breaking (BO = 0.79). Meanwhile, another C-O bond is basically formed (BO = 0.72). (Figure 3-25E). At the first transition state (Figure 3-25G), the proton is bonded to H123 (BO = 0.84). Whereas at the second transition state, the proton is roughly in the middle of H123 and S98 (Figure 3-25H). Besides, similar to the acylation step, proton transfer from H123 to S98 happens after the nucleophilic attack of the water molecule for all four ligands.

3.5 Discussion

3.5.1 Comparison of DFT QM/MM calculations with semi-empirical QM/MM calculations

Since AM1 is the most solid semi-empirical QM method for our ClpP systems. We start from caparison of B3LYP QM/MM calculations with AM1 calculations. There are several major differences between AM1 calculations and B3LYP calculations for the acylation reaction. Firstly, AM1 calculated energy barriers are at least 10 kcal mol^{-1} higher than the experimental values. Secondly, two-dimensional potential energy surfaces calculated at the AM1 level always have two transition states and one intermediate for all ligands. In addition, positions of transition states and characteristic features of potential energy surfaces around transition states are quite different compared to B3LYP potential energy surfaces (Figures 3-14 and 3-16). Since the DFT method, in general, performs better than semi-empirical QM methods and B3LYP calculated values are closer to experimental values. The acylation reaction is more likely to be conducted through the mechanism suggested by B3LYP calculations. This implies that the AM1 method cannot provide even qualitatively correct results for the acylation reaction.

On the other hand, DFTB potential energy surfaces for β-lactones are very similar to B3LYP potential energy surfaces (Figures 3-18A, B and 3-16A, B). But there are still significant differences between DFT and DFTB potential energy surfaces for the phenyl ester and the fluorescent substrate (Figures 3-18C, D and 3-16C, D). In addition, DFTB calculated energy barriers still differ from values calculated at the B3LYP/def2-TZVP level. Since tcl-ChemShell doesn't support the interface to the QM program that can carry out DFTB calculations, we have to use Amber to conduct DFTB QM/MM calculations. Additionally, Amber QM/MM calculations have limitations of not supporting the charge shift scheme and compulsory GB calculations. Based on these reasons, the DFTB method is not a good choice for our projects despite its advantage in computational cost.

Although semi-empirical QM methods have the advantage of calculation efficiency and were used in many previous theoretical studies about both acylation and deacylation reactions of serine proteases, it is not that useful if semi-empirical calculation results are qualitatively incorrect. One potential development direction of semi-empirical QM methods is to re-parameterize a semi-empirical method for a specific system by fitting the semi-empirical PES to the PES of higher level QM method. Then people can perform re-parameterized semi-empirical QM/MM MD to obtain more accurate results.

3.5.2 Explanation of the behaviour of 1D simulations

After we obtained 1D reaction profiles and 2D potential energy surfaces calculated at both AM1 and BP86 levels, we can easily understand the results of 1D simulations. The 1D simulation of the nucleophilic attack step can be interpreted like this, for any given $\xi 1$, the 1D calculation is just looking for the structure with the lowest energy along the reaction coordinate 2 on the 2D PES. But crossing the barrier is not allowed. The 1D simulation of the proton transfer step can be understood similarly. Now we use Lig25 as an example to illustrate this relationship between 1D and 2D simulations. As can be seen from the 2D BP86 PES for Lig25 (Figure 3-15B), when $\xi 1$ value is smaller than -0.8 Å, there is an energy barrier to stop the system towards the upper part of the surface where optimized structures have relatively low energies. However, when the reaction coordinate 1 is larger than -0.8 Å, the system will directly descend to a state with a lower energy due to no energy barrier any more. And that's why a huge energy drop is observed for 1D reaction profiles calculated at the BP86 level for β-lactones (Figure 3-13). On the contrary, there is an intermediate on the 2D AM1 potential energy surface for Lig25. Before the acylation reaction reaching the intermediate at (-0.3 Å, -0.4 Å), the system is forbidden to open the four-membered ring due to a small energy barrier. That is why a smooth reaction profile for Lig25 calculated at the AM1 level was obtained (Figure 3-12).

3.5.3 Significance of 2D potential energy surfaces

In this chapter, we mainly study both acylation and deacylation reactions of SaClpP by calculating two-dimensional potential energy surfaces. To the best of our knowledge, it is the first time to calculate two-dimensional potential energy surfaces of both acylation and deacylation reactions for a serine protease by these two reaction coordinates. The significance of a two-dimensional potential energy surface includes: firstly, a 2D PES can avoid an error of assembling the nucleophilic attack step and the proton transfer step together to get a 1D reaction profile. We assumed that the product state of the nucleophilic attack step has the same energy as the reactant state of the proton transfer step. However, these two states are not exactly the same since applied restraints on these two systems are different. Secondly, it can also avoid the energy drop observed in the 1D reaction profile calculated at the DFT level (Figure 3-12). Furthermore, a 2D potential energy surface contains more information about entire reaction compared to the 1D reaction profile. Therefore, a two-dimensional potential energy surface is a good tool to study detailed reaction mechanisms of both acylation and deacylation reactions.

3.5.4 Water dynamics

By taking more dynamics effects of the system into account, calculated (free) energy barriers were iteratively improved. Thus, our calculations with four different setups reveal that the effect of the dynamics, especially water dynamics, plays an important role in the deacylation step. In this project, our final protocol that combines a short MD simulation, more QM water molecules, and the QM/MM FEP technique performs much better than the traditional QM/MM method for modelling the deacylation step. On the contrary, there is no water molecule directly involved in the acylation reaction. The dynamics of surrounding water molecules are not a major factor that influences the energy barrier. Thus, the traditional static QM/MM method works fine for simulating the acylation step.

3.5.5 Comparison of calculated values with experimental values

In the end, we can compare our calculated free energy barriers with the experimental values. Experimental and calculated values with our best protocol are listed in Table 3.8. In general, our calculation results agree with the experimental values. Based on the experimental values and the fact that acyl-enzyme complexes for two β-lactones and the phenyl ester can be detected by mass spectrometry, it can be derived that the deacylation step is the rate-determining step for β-lactones and the phenyl ester. This implies that the free energy barrier of the deacylation step is higher than that of the acylation step for these three ligands. Our calculations also support this finding (Table 3.8). On the other hand, the acyl-enzyme complex for the fluorescent substrate cannot be detected.

Table 3.8. Comparison of calculated free energy barriers with experimental values for acylation and deacylation reactions of SaClpP. (*: derived from ref. [24])

	Step	Lig24	Lig25	ML90	Sub
Experimental	Acylation	20.8	20.6	21.0	18.5*
	Deacylation	23.7	23.9	24.0	N/A
Calculated	Acylation	14.4	20.7	16.9	25.9
	Deacylation	23.5	27.0	24.8	23.7

The acylation step is thus the rate-determining step in this case. Because the deacylation progress is simply a hydrolysis reaction of an ester and covalently bound complexes for all four ligands have similar structures, energy barriers of deacylation reactions should have similar values. Our calculated free energy barriers for the deacylation reaction are from 23.5 to 27.0 kcal mol^{-1}, which basically reproduces this assumption. Considering the fact that the energy barrier of the acylation reaction is higher than that of the deacylation reaction for the fluorescent substrate and the other way around for the other three ligands, the energy barrier of the acylation reaction for the fluorescent substrate should be higher than energy barriers for β-lactones and the phenyl ester. Our calculation results support this inference as well. This means there is an intrinsic contradiction between the experimental value of the acylation reaction for the fluorescent (18.5 kcal mol^{-1}) and the experimental values (20.6–21.0 kcal mol^{-1}) for the other three ligands. One of the possible reason is that our model is still quite simplified and the real situation is more complicated. There might be other unknown factors which accelerate the acylation reaction for the fluorescent substrate.

3.6 Summary

Experimental free energy barriers for both acylation and deacylation reactions of SaClpP with three inhibitors were derived. Two-dimensional QM/MM simulations for SaClpP with these inhibitors and one fluorescent substrate were performed. Several semi-empirical QM methods were used to simulate acylation reactions. Our calculation results reveal that semi-empirical QM methods cannot describe the acylation reaction of SaClpP systems properly. For most semi-empirical QM methods, even qualitatively correct results (one-step reaction mechanism) for β-lactones and the phenyl ester were not obtained. AM1 is the most robust semi-empirical method for ClpP systems. But it overestimates energy barriers and doesn't provide correct reaction mechanism for the acylation reactions of β-lactones and the phenyl ester. Although DFTB calculations agree with DFT calculations for two β-lactones very well, potential energy surfaces calculated by DFTB for other two ligands show a different pattern compared to DFT potential energy surfaces.

Our calculation results with the best protocol correlate with experimental data very well. For acylation, the protocol is simply static 2D QM/MM calculations at the B3LYP/def2-TZVP//BP86/def2-SVP level. For deacylation, since water dynamics is the key to successful simulations, the best protocol combines a short MD simulation, more QM water molecules, and QM/MM FEP calculations.

It seems that QM/MM calculations can only provide semi-quantitatively correct results. For β-lactones and the phenyl ester, our calculation results reproduced the experimental finding that the deacylation step is the rate-determining step. While for the fluorescent substrate, our calculations reveal that the acylation step is the rate-determining step. In general, QM/MM calculations cannot distinguish small differences in free energy barriers for different ligands. For two β-lactones, the length of the aliphatic chain has no significant influence on energy barriers of both acylation and deacylation reactions.

Our calculation results also suggest that acylation reactions of SaClpP with β-lactones and the phenyl ester follow a one-step mechanism. On the contrary, the reaction mechanism for the fluorescent substrate is a traditional two-step mechanism. It is still not very clear which factor mainly determines the reaction mechanism for the deacylation reaction. Our simulations for the deacylation step reveal that if an intermediate exists, it is usually very shallow on the potential energy surface. In addition, no matter whether the reaction is a one-step or two-step mechanism, the proton transfer step is always after the nucleophilic attack step.

Chapter 4

A QM/MM study of HCV NS3/4A protease variants with the natural substrate NS4A/4B and MAVS

4.1 Systems preparation and calculation setup

4.1.1 The generally accepted reaction mechanism

The generally accepted reaction mechanism (Figure 4-1) of the acylation reaction for HCV NS3/4A protease variants with the NS4A/4B substrate and MAVS is almost the same as that for SaClpP with the fluorescent substrate. The only difference is the residue numbering for the catalytic triad.

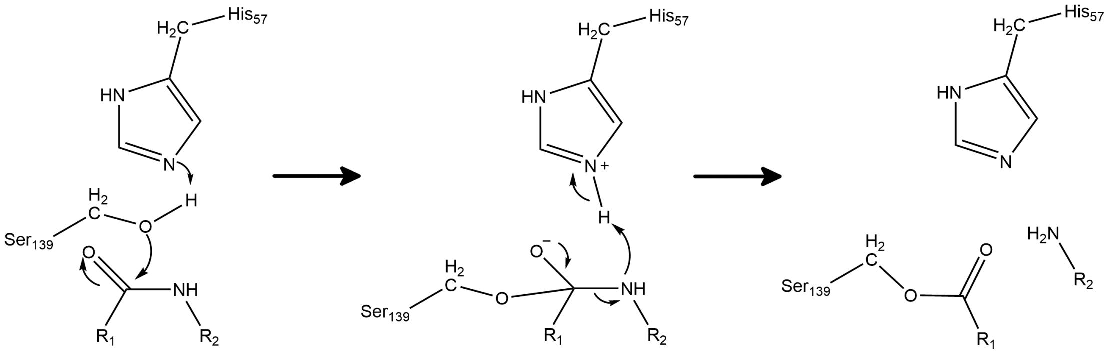


Figure 4-1. The generally accepted reaction mechanism of the acylation reaction for HCV NS3/4A serine protease variants with the NS4A/4B substrate and MAVS. The reaction is separated by two steps. The first step is nucleophilic attack of S139 to the substrate and the second step is proton transfer from H57 to the substrate with the cleavage of the peptide bond.

4.1.2 Molecular dynamics simulations

The initial structure for simulations was prepared from the PDB file 3RC5 [70], which is a crystal structure of the HCV NS3 protease fused with the NS4A cofactor and bound to the N-terminal cleavage product of MAVS. The missing C-terminal part of MAVS (residues P’1-4) was restored based on the crystal structure of the complex of bovine alpha chymotrypsin with Eglin C (PDB ID: 1ACB) [236]. The protonation state of histidine residues was determined by examining their direct environment manually. The histidine of MAVS (further referred to as H1’(p)) was modeled in its neutral form with a proton on the epsilon nitrogen of its imidazole ring. All amino acids were modelled in their standard protonation state at neutral pH, accordingly with predictions by propka 2.0 [237]. Point mutations in protein and substrates were introduced manually. The Amber03.r1 force field [145] was used for the protein and substrates. The structures of complexes were prepared for minimization using tleap from AmberTools15 [142] and all calculations were conducted in a neutralized, rectangular SPC/E water box [155] extending at least 14 Å from any protein atom at each side of the box. All calculations were performed with the Amber14/AmberTools15 suite of programs [142]. After energy minimization and adjustment of the box size to reach a target density of 1000 kg/m^3, the systems were gradually heated up to 300K in the NVT ensemble [238]. Finally, production runs were performed for 1000 ns with four replicas in the NPT ensemble at 300K and 1 bar. Both heat up and production runs employed a time step of 1 fs and SHAKE constraints [227] were applied to all bonds involving hydrogen atoms. Temperature and pressure were controlled with Langevin dynamics (collision frequency of 4.0 ps^{-1}) and isotropic position scaling (τ_p of 1 ps), respectively. All analyses of MD simulations were performed on trajectories frames extracted from the last 900 ns of NPT production runs, in order to remove bias towards the initial crystal structure.

4.1.3 Extraction of residue interaction networks from MD trajectories

Hydrogen bonds occurring within the protein or substrate during our simulations were extracted from analysis of trajectories using the “hbond” command in cpptraj [239], setting a maximum donor-acceptor distance of 3.5 Å and a minimum donor-hydrogen-acceptor angle of 135°. Close contacts between carbon atoms were extracted using the "nativecontacts" command in cpptraj [239], with a maximum distance of 5 Å and saving the full native and non-native time series ("nonnative" option). The resulting timelines of hydrogen bonds and close contacts, which detail the presence or absence of an interaction in each frame during simulation, were processed using aifgen [240] and loaded into the network analysis plugin SenseNet (manuscript in review) for Cytoscape 3 [241].

4.1.4 Selection of starting structures for QM/MM calculations

To extract representative structures of particular hydrogen bond interaction pattern, we performed clustering of residue interaction networks in SenseNet. The clustering was performed on a subnetwork including all peptide (NS4A/4B or MAVS) residues, the catalytic triad H57, D81, and S139, residues Q41, R155, D168 and all residues directly connected to those mentioned via hydrogen bonds or close contacts. Each timeframe of the hydrogen bond subnetwork was transformed into an interaction matrix, and the distance between matrices was defined by their Frobenius norm. Clusters were assembled using the hierarchical agglomerative method until 10 clusters remained. Initial structures were selected from the largest cluster. If the largest cluster contained two conformations of the dihedral angle CA-CB-CG-CD2 of H1'(p) (see details in section 4.2.2), the cluster was further divided into two subgroups. Initial structures for both conformations were then selected from the corresponding subgroups. The value of the first reaction coordinate (equation (4-1)), which is negative in the reactant state, was calculated for each structure in this cluster. The structure with the maximum of the calculated value (namely the minimum in terms of the absolute value) was chosen as the initial structure.

4.1.5 QM/MM calculations

The acylation reaction of NS3/4A protease variants was studied within the hybrid quantum mechanics/molecular mechanics (QM/MM) framework. After choosing the initial structure as described above, the system was first optimized at the MM level while keeping the QM region frozen. The QM region consists of side chains of the catalytic triad (H57, D81, and S139), P1, and P1' residues in the substrate and their neighboring CO and NH groups [77]. All residues and water molecules further than 30 Å away from the oxygen of the S139 hydroxyl group were cut out using an in-house python script (Figure 4-2). The remaining termini were capped by either ACE or NME residues. If the residues placed directly before and after the deleted residue were inside the sphere, the deleted residue was substituted by a glycine. The outmost 5 Å layer was fully fixed for the calculations. All QM/MM calculations were driven by ChemShell software package [217], which provides an interface to Turbomole [135] and dl_poly [228]. Geometry optimizations were performed with the dl_find [205] module of ChemShell. The maximum number of optimization cycles was set to 1000. The RI approximation was used for density functional theory (DFT) calculations. The criterion of SCF convergence was 10^{-6} Hartree. The link-atom scheme was used to treat the QM-

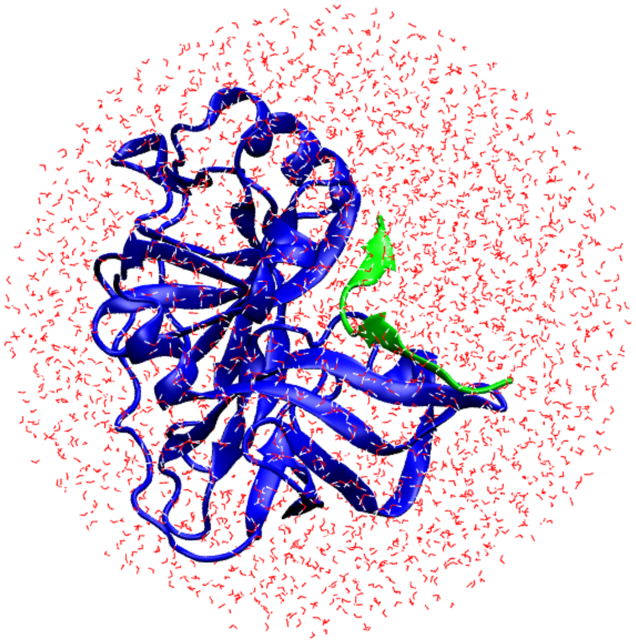


Figure 4-2. Illustration of the prepared model of the WT HCV NS3/4A protease-MAVS complex (good conformation) with surrounding water molecules for the QM/MM calculation. The HCV NS3/4A protease is shown in blue and the MAVS peptide is shown in green.

MM boundary and electrostatic embedding was used to deal with electrostatic QM-MM interactions. In addition, point charges of MM frontier atoms were shifted away and point dipoles were added to compensate [158]. Electrostatic interactions between MM atoms were evaluated for all atom pairs. All potential energy surfaces were calculated in two dimensions. Two reaction coordinates were used for the study (Figure 4-3). The first reaction coordinate (nucleophilic attack reaction coordinate) is defined as

$$\xi 1 = d(\mathrm{O_{Ser}} - \mathrm{H_{Ser}}) - d(\mathrm{N_{His}} - \mathrm{H_{Ser}}) - d(\mathrm{O_{Ser}} - \mathrm{C_{Cys}}) \tag{4-1}$$

and the second reaction coordinate (proton transfer reaction coordinate) is defined as

$$\xi 2 = d(\mathrm{C_{Cys}} - \mathrm{N_{P1'}}) - d(\mathrm{N_{P1'}} - \mathrm{H_{Ser}}) \tag{4-2}$$

After preparing the model for QM/MM calculations, an initial energy minimization was performed while keeping the QM region frozen. A 2D potential energy surface was determined by restrained geometry optimizations at grid points. The calculation started from the grid point at the bottom left corner. The first series of geometry optimizations were performed by gradually increasing $\xi 2$. Then, several parallel calculations were

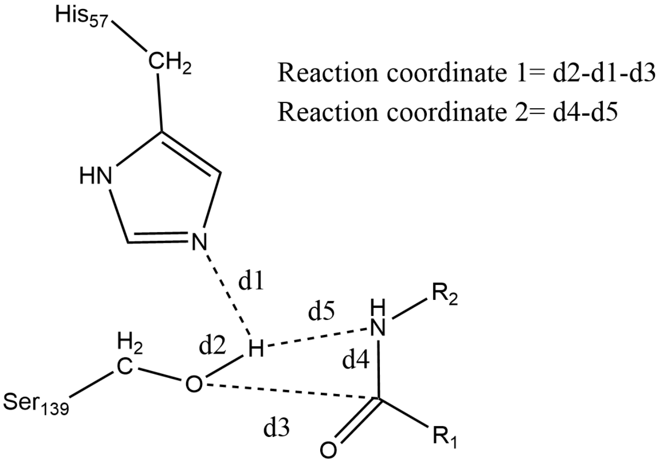


Figure 4-3. The setup of reaction coordinates for the acylation reaction of HCV NS3/4A protease variants bound to MAVS and the NS4A/4B substrate

conducted by fixing $\xi2$ and increasing $\xi1$ to obtain the bottom part of the surface. The top left part of the surface was not calculated due to very high potential energies in this area. In this case, the proton of S139 is close to the substrate and the peptide bond is partially broken whereas the ester bond is not formed. The top right part of the potential energy surface was obtained by several parallel geometry optimizations along the reaction coordinate 2. The interval between grid points along the reaction coordinate 1 is 0.1 Å and along the reaction coordinate 2 is 0.2 Å. At each grid point, both reaction coordinates were restrained by a harmonic potential with a force constant of 3.0 Hartree/Bohr^2. The geometry was optimized at the BP86/def2-SVP level [117, 133]. Afterwards, single point calculations were performed for all optimized structures at the B3LYP/def2-TZVP level [115, 118, 120, 122, 133] to obtain the final potential energy surface. D3 dispersion correction [125] was applied to all geometry optimization and single point calculations.

4.1.6 QM/MM free-energy perturbation

Free energies of activation and free energy differences of the acylation reaction were estimated by performing QM/MM free energy perturbations (FEP) [215, 216]. The qmmmfep module implemented in ChemShell was used for this purpose [217]. We slightly modified the module to use RESP charges for QM atoms instead of default ESP charges. The 200 closest MM atoms were used as reference points when fitting the RESP charges. The fitting of QM charges was based on DFT calculations at the B3LYP/def2-TZVP level. The QM part, MM frontier atoms and outmost 5 Å layer were kept frozen. The method of link atom perturbation was set to 4, implying that QM boundary atoms, link atoms and MM boundary atoms were all perturbed [216]. Free water molecules were kept internally rigid by using the SHAKE algorithm [227]. In each window, the system was heated up and equilibrated at 300 K for 10 ps in the NVT ensemble with 1 fs time step. The Nosé-Hoover chain thermostat [210, 211] was used with a chain length of 4 and time constant (τ_T) of 0.02 ps. The FEP production run was performed for forward perturbation with a time step of 1 fs for 10 to 15 ps depending on the convergence check [216, 231]. All hydrogen atoms were assigned the mass of deuterium to ensure energy conservation with a time step of 1 fs [216].

4.2 Calculation Results and Discussion

4.2.1 Hydrogen bonds networks

We performed MD simulations for several NS3/4A protease variants, i.e., wild type (WT), D168A, Q41R, Q41R-D168A, and R155K, either bound to the NS4A/4B natural substrate or to the MAVS peptide. The resulting trajectories, which spanned a total simulation time of 4 µs over four replicas, were further analyzed with respect to the networks of hydrogen bonds occurring over the course of the simulation. Figure 4-4 illustrates hydrogen bond networks near the active site for all protease-substrate systems. The network can be subdivided into three parts: The first part is the key unit formed by the catalytic triad, G137, and C1(p), where the (p) denotes a location within the peptide substrate. Among the catalytic triad, there are one hydrogen bond between S139 and H57 and two hydrogen bonds between D81 and H57. The carbonyl group of C1(p) sits in the oxyanion hole forming two hydrogen bonds to the backbone of G137 and S139. This subnetwork encompassing 5 hydrogen bonds is quite stable and exists in all systems (Figures 4-4 and 4-5). The second part includes three residues, namely H57, Q41 (or Q41R), and a residue at the P1' position (H1'(p) for MAVS and S1'(p) for the NS4A/4B substrate). Q41 or Q41R connects H57 and P1' by a hydrogen bond and π-π interactions. In the following text, this part will be called the lower part. The last one consists of residues around R155 and D168 including E4(p) or E2(p), D81, and R123. Hydrogen bonds in this part vary a lot depending on protease variants. This part will be called the upper part in this chapter.

Two representative 3D structures of the wildtype HCV NS3/4A protease bound to MAVS are shown in Figure 4-5A, B. We will further discuss these two different conformations in the next section. As can be seen from the picture, hydrogen bonds in the central part are very solid in both conformations. However, some hydrogen bonds in the upper and the lower part including the hydrogen bond between Q41 and H57 (0.38), the hydrogen bond between Q41 and H57 (0.43), the hydrogen bond between D81 and R155 (0.77), and the hydrogen bond between D168 and R123 (0.66) may

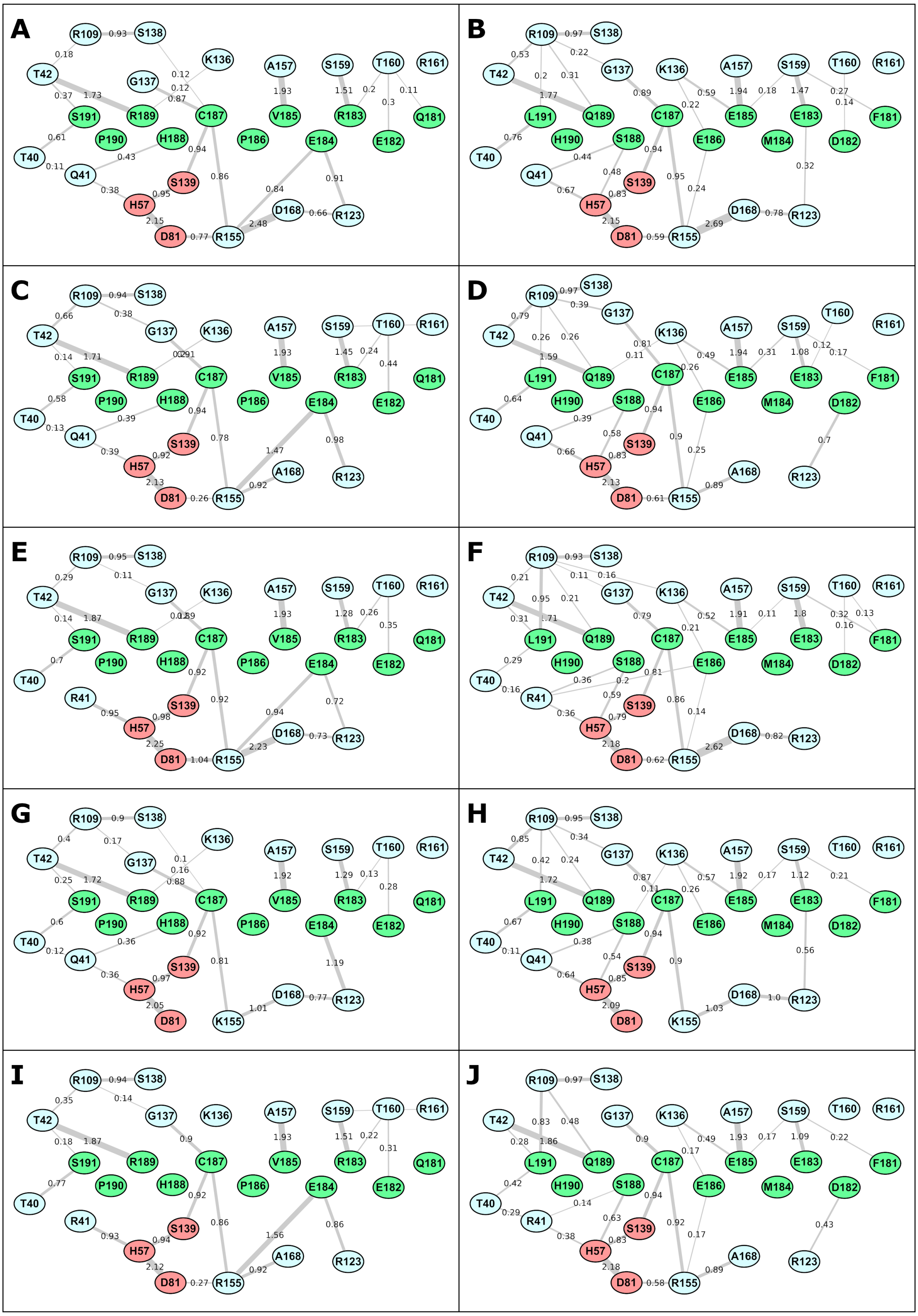


Figure 4-4. Illustration of hydrogen bond networks of protease-substrate complex simulations for HCV NS3/4A protease variants. (A) WT-MAVS (B) WT-NS4A/4B (C) D168A-MAVS (D) D168A-NS4A/4B (E) Q41R-MAVS (F) Q41R-NS4A/4B (G) R155K-MAVS (H) R155K-NS4A/4B (I) Q41R-D168A-MAVS (J) Q41R-D168A-NS4A/4B.

change from time to time (Figures 4-4A and 4-5A, B). The D168A mutant bound to MAVS shows a different pattern in the upper part (Table 4.1 and Figure 4-4C). Since the aspartate residue is substituted by a hydrophobic alanine residue, there is no hydrogen

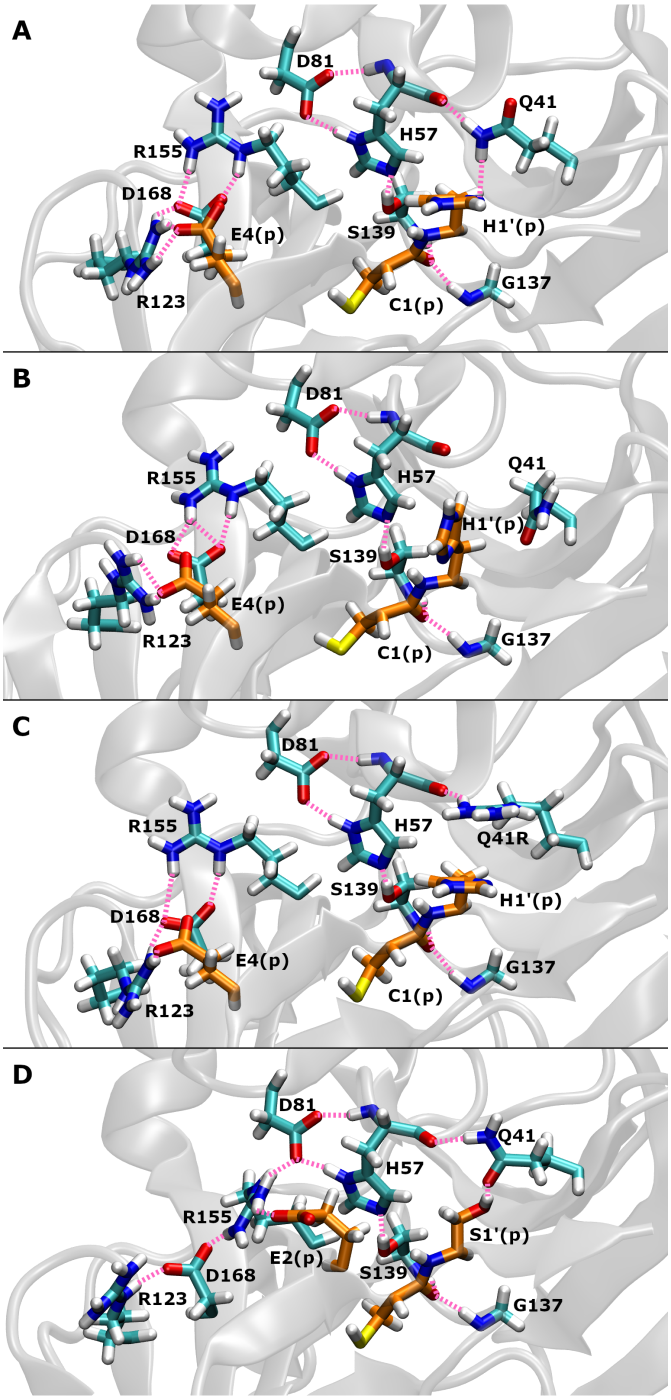


Figure 4-5. Illustration of initial structures of protease-substrate complexes for QM/MM simulations showing the binding site and hydrogen bonds networks around the catalytic triad. (A) WT-MAVS with good conformation (B) WT-MAVS with bad conformation (C) Q41R-MAVS (D) WT-NS4AB. Hydrogen bonds are shown in pink. Carbon atoms in the protease are shown in cyan and carbon atoms in the substrate are shown in orange.

bond between D168A and R123. In addition, hydrogen bonds connecting to R155 also change a lot. The guanidine group of R155 tends to interact more with E4(p) instead of D168A. So there is only one backbone H-bond remaining between R155 and D168A. In the meanwhile, H-bond between R155 and D81 also decreases from 0.77 to 0.26 (Figure 4-4C). While for the upper part, there is almost no change of hydrogen bonds between Q41 and H1'(p) and between Q41 and H57. In contrast, the Q41R mutant bound to MAVS mainly affects the lower part. Q41R forms a stable hydrogen bond to the backbone of H57 and the hydrogen bond to H1'(p) is replaced by strong π-π interactions (Figures 4-4E and 4-5C). For the R155K mutant bound to MAVS, the side chain of the lysine residue cannot interact with D81 and E4(p). Only two backbone hydrogen bonds to C1(p) and D168 remain the same (Figure 4-4G). The double mutant Q41R-D168A bound to MAVS can be simply considered as the combination of two single mutants (Q41R and D168A) bound to MAVS (Figure 4-4I).

The hydrogen bonds network of the wildtype protease bound to the NS4A/4B substrate shows a slightly different pattern compared to the network of the wildtype protease with MAVS (Figures 4-4B and 4-5D). There is a hydrogen bond between Q41 and S1'(p) (0.48) and R155 forms a weak hydrogen bond to E2(p) (0.24) instead of E4(p) in MAVS. Hydrogen bonds among the catalytic triad are still quite stable since the natural substrate is optimized for the protease. Regarding the D168A mutant bound to the NS4A/4B substrate, there is no hydrogen bond between D168A and R123 and only one backbone H-bond remaining between R155 and D168A, which is similar to the D168A mutant bound to MAVS (Figure 4-4D). For the R155K mutant bound to NS4A/4B, the R155K residue only keeps two hydrogen bonds to C1(p) and D168 respectively (Figure 4-4F). The Q41R mutant bound to NS4A/4B slightly reduces hydrogen bonds to S1'(p) and H57 (Figure 4-4H). In the end, the double mutant Q41R-D168A bound to the NS4A/4B substrate can be still treated as the combination of two single mutants (Figure 4-4J).

Comparison of hydrogen bonds networks of HCV NS3/4A protease mutants bound to MAVS and the NS4A/4B substrate with the corresponding hydrogen bonds network of the wildtype protease is summarized in Table 4.1. In short, hydrogen bonds near the active site decrease for the NS4A/4B substrate bound to NS3/4A protease mutants. While for MAVS bound to NS3/4A protease mutants, the D168A mutant only shows an increase of hydrogen bond between R155 and E4(p) and the Q41R mutant shows an increase of hydrogen bond between Q41R and H57. For both substrates, it seems that the Q41R mutant only influences the lower part of the network and the D168A mutant only influences the upper part of the network. This means Q41R and D168A change

Table 4.1. Comparison of hydrogen bonds networks for HCV NS3/4A protease variants (wildtype as reference, “=”, reference or almost no change; “+”, an increase of hydrogen bond; “-”, a decrease of hydrogen bond; “--”, a decrease of hydrogen bond to almost 0).

Substrate	Variant	D81-R155	R155-D168	R155-E4(p)	D168-R123	Q41-H57	Q41-H1’(p)
MAVS	WT	=	=	=	=	=	=
	D168A	-	-	+	--	=	=
	R155K	--	-	--	=	=	=
	Q41R	=	=	=	=	+	--
	Q41R-D168A	-	-	+	--	+	--
Substrate	Variant	D81-R155	R155-D168	R155-E2(p)	D168-R123	Q41-H57	Q41-S1’(p)
NS4A/4B	WT	=	=	=	=	=	=
	D168A	=	-	=	--	=	=
	R155K	--	-	–	=	=	=
	Q41R	=	=	=	=	-	-
	Q41R-D168A	=	-	=	--	-	-

the hydrogen bonds networks independently. Another interesting point is that Q41R mutants make a stronger H-bond between H57 and Q41R for MAVS systems while Q41R mutants lead to a weaker H-bond between H57 and Q41R for NS4A/4B systems. This is because the movement of Q41R and H1’(p) in MAVS systems is largely restricted by π-π interactions between them and a hydrogen bond between Q41R and H57.

4.2.2 Q41 protease-MAVS complexes show two different conformations

In our previous work [88], we noticed that the Q41R mutant participated in π-π interactions with MAVS, which were close to the catalytic site. In this work, we extended MD simulations, observing that two different conformations of the P1’ histidine residue of MAVS exist for NS3-Q41 systems with respect to the dihedral angle of CA-CB-CG-CD2 of the histidine residue. We denote this residue as H1’(p) in the following text for simplicity. The root mean square deviation time series of our simulations are reported in Figure 4-6 and indicate that the systems reached equilibrium after approximately 100 ns. The histograms distribution of the dihedral angle of H1’(p) extracted from production MD trajectories (Figure 4-7), clearly outlines two major conformations of H1’(p) in WT, R155K, and D168A mutants. In contrast, only one conformation of this residue was observed in the case of Q41R and Q41R-D168A mutants.

Due to the close proximity of H1’(p) to the catalytic site, we hypothesized that this conformational difference may explain another observation made in ref. [88], namely

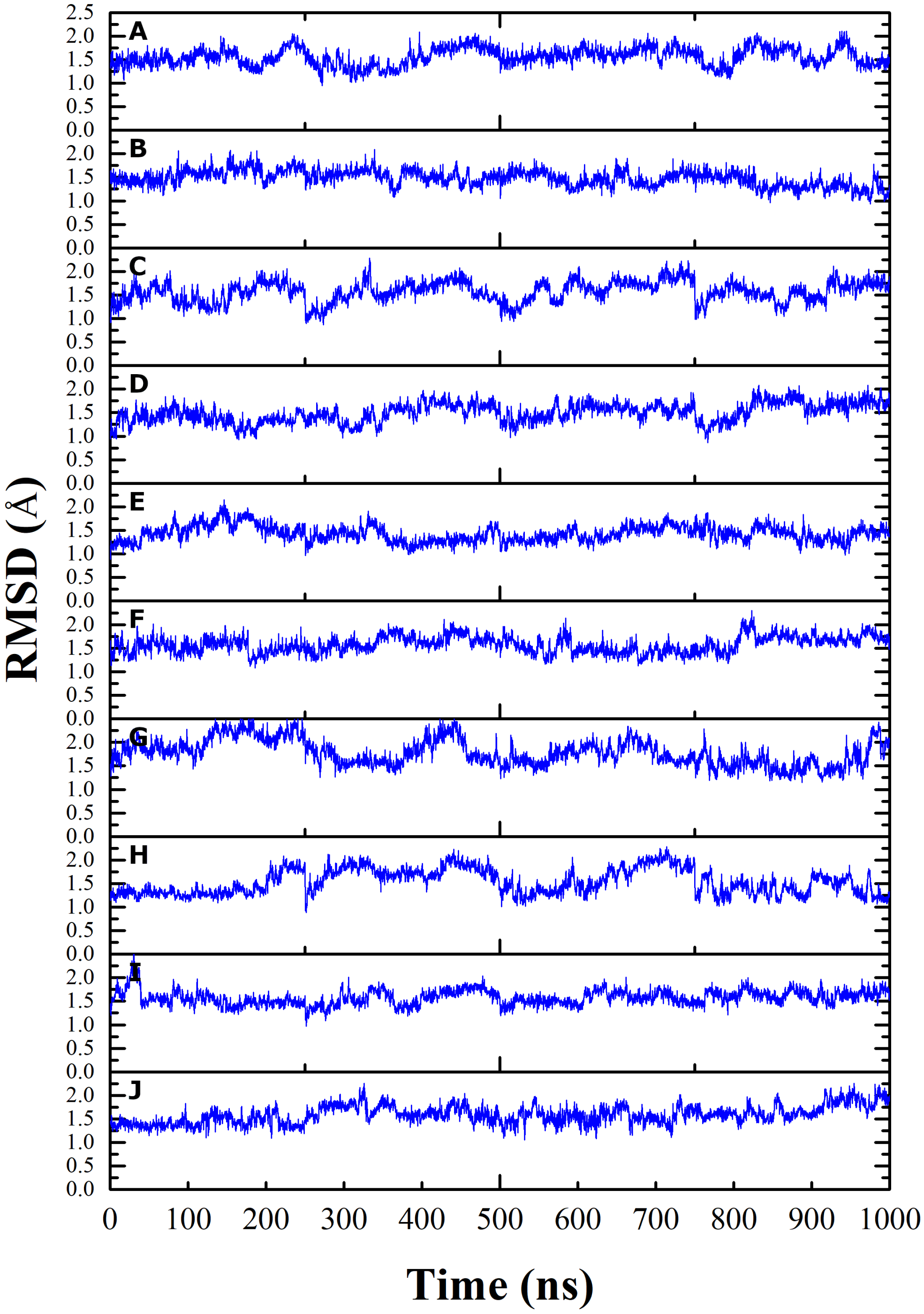


Figure 4-6. RMSD plots of the backbone of the protein-substrate complex with respect to its initial structure of production simulation for (A) WT-MAVS, (B) D168A-MAVS, (C) R155K-MAVS, (D) Q41R-MAVS, (E) Q41R-D168A-MAVS, (F) WT-NS4AB, (G) D168A-NS4AB, (H) R155K-NS4AB, (I) Q41R-NS4AB, and (J) Q41R-D168A-NS4AB.

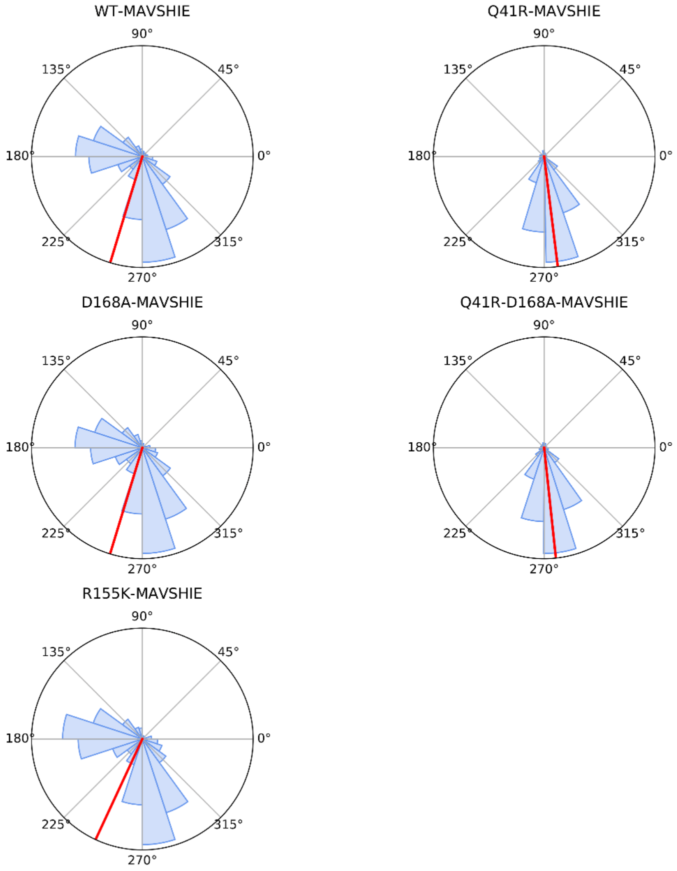


Figure 4-7. Histograms of the dihedral angle of CA-CB-CG-CD2 of H1'(p) for NS3/4A protease variants bound to MAVS extracted from production MD trajectories. Red lines show the mean of the dihedral angle.

that the substrate cleavage rate (Table 4.2) of Q41R and Q41R-D168A mutants (0.275 and 0.507 s^{-1}) is higher than that of WT and of the D168A mutant (0.058 and 0.029 s^{-1}). In order to test this hypothesis, QM/MM calculations for the acylation reaction of NS3/4A protease variants bound to MAVS were performed. Anticipating on our results and for the sake of simplification, the conformations with a dihedral angle of

Table 4.2. Experimental kinetic constants for NS3/4A protease variants [88].

Substrate	Variant	$K_M(\mu\text{M})$	k_{cat}(s^{-1})
MAVS	WT	7.82	0.058
	D168A	14.41	0.029
	Q41R	37.32	0.275
	Q41R-D168A	1.77	0.507
NS4A/4B	WT	3.11	0.19
	D168A	2.45	0.09
	Q41R	2.11	0.18
	Q41R-D168A	1.71	0.15

CA-CB-CG-CD2 of H1'(p) between 250° and 300° will be further referred to as good conformations and the conformations with a dihedral angle between 125° and 200° as bad conformations (Figure 4-7). Initial structures for subsequent calculations were selected based on the clustering of residue interaction networks. In the structures of wild type protease bound to MAVS with good conformation (Figure 4-5A), Q41 forms two hydrogen bonds with H57 and H1'(p). In contrast, these two hydrogen bonds are not present in the bad conformation and H1'(p) lies in a different orientation. In this case, the imidazole ring of H1'(p) is mostly parallel to the plane of the amide group of Q41 (Figure 4-5B). Figure 4-5C shows the sole conformation of the Q41R mutant bound to MAVS observed in our simulation. A π-π stacking interaction is formed between R41 and H1'(p) and a hydrogen bond between R41 and the backbone of H57 further stabilizes the conformation, which in turn locks the orientation of H1'(p). Hence, only the good conformation is available for the Q41R mutant. In addition, we also performed QM/MM calculations for the NS4A/4B substrate to verify the effectiveness of our computational model. In the case of the natural substrate NS4A/4B, the residue at the substrate's 1' position is a serine, i.e. S1'(p), which did not change its conformation substantially during simulation (Figure 4-5D).

4.2.3 Two-dimensional potential energy surfaces calculated by QM/MM methods

The potential energy surfaces calculated at the B3LYP/def2-TZVP level are depicted in Figures 4-8 and 4-9 for MAVS or the NS4A/4B, respectively. The top left part for all potential energy surfaces was not calculated due to unrealistic high energies and non-physical structures in this area. The overall pattern of the surface is generally similar from one system to another. Unlike reported in another study performed at the semi-empirical SCC-DFTB level [8], our calculation still support the hypothesis of a metastable tetrahedral intermediate. A shallow minimum is indeed present on the B3LYP/def2-TZVP//BP86/def2-SVP potential energy surfaces for the acylation reaction of NS3/4A protease variants bound to MAVS and the NS4A/4B substrate. We note, however, that such shallow minimum may vanish at room temperature, in agreement with the calculations in ref. [8]. The reactant state is in the bottom-left corner and the product state is in the upper-right region. The first transition state and the intermediate are in the bottom-right region and the second transition state is on the right. Nearly all potential energy surfaces show two separate transition states connected by a tetrahedral intermediate. The bad conformation of R155K is the only exception, where the reaction occurs in a single step with a prohibiting energy barrier of 37.6 kcal mol^{-1}. The control of

Table 4.3. Calculated QM/MM energy barriers and energy differences of the acylation reaction for HCV NS3/4A protease variants. Free energies calculated by QM/MM FEP are listed in parenthesis. (unit: kcal mol^{-1}, *: data from ref. [242])

Substrate	Variant	Experimental	$\Delta E^{\neq 1}$	ΔE_{int}	$\Delta E^{\neq 2}$	ΔE
MAVS	WT(good)	19.3	21.2(18.4)	17.7(14.0)	19.6(14.7)	10.4(7.9)
	WT(bad)	N.A.	20.7(19.2)	20.3(17.3)	26.0(23.9)	13.2(14.2)
	D168A(good)	19.7	23.6(21.9)	22.0(18.5)	24.0(20.3)	10.5(11.4)
	D168A(bad)	N.A.	34.6(31.2)	34.0(30.4)	34.6(30.6)	15.0(16.9)
	R155K(good)	17.7*	26.5(20.4)	26.4(20.1)	27.8(21.2)	9.7(7.5)
	R155K(bad)	N.A.	-	-	37.6(28.7)	12.9(16.4)
	Q41R	18.3	23.2(18.2)	22.8(16.8)	23.3(17.3)	5.4(7.4)
	Q41R-D168A	18.0	21.8(18.8)	21.5(18.2)	23.5(21.5)	7.0(7.6)
NS4A/4B	WT	18.6	23.9(22.2)	20.8(19.5)	21.1(19.6)	13.1(12.3)
	D168A	19.0	23.3(21.7)	22.2(20.7)	24.3(21.7)	10.3(11.5)
	R155K	17.7*	24.4(20.5)	22.5(17.2)	25.4(19.2)	10.3(8.0)
	Q41R	18.6	19.3(19.1)	19.2(18.8)	24.9(22.1)	7.9(9.0)
	Q41R-D168A	18.7	23.9(20.9)	21.2(17.6)	23.9(19.7)	11.2(7.9)

the reaction via two coordinates bears several advantages: i) surfaces contain information regarding the whole acylation reaction, including nucleophilic attack and proton transfer steps; ii) the topology of the surfaces and energetics of stationary points can be compared fairly easily; iii) discontinuities on the potential energy surface due to uncontrolled regions of the molecule system are greatly reduced.

Calculated energy barriers and energy differences of the acylation reaction for NS3/4A protease variants bound to MAVS and to the NS4A/4B substrate are listed in Table 4.3. Experimental values in the table were derived from rate constants (Table 4.2) as measured in our previous study [88]. The corresponding value for R155K is from another study [242] with different experimental condition and thus is only listed here for comparison. The protease variants with Q41 (i.e. wild type, R155K and D168A) show a higher energy barrier for the acylation reaction when in the bad conformation compared to the good conformation (4.8 to 10.6 kcal mol^{-1} higher). This tends to indicate that the acylation reaction is more likely to proceed via the good conformation for these systems. Excluding the systems with the bad conformation, all others have similar energy barriers. In general, our calculated potential energy barriers (21.2–27.8 kcal mol^{-1}) are slightly higher than the experimental values (17.7–19.7 kcal mol^{-1}). For the acylation reaction of the NS4A/4B substrate bound to NS3/4A protease variants, calculated energy barriers are in a very narrow range (23.9–25.4 kcal mol^{-1}). These values are also slightly higher than the experimental values (17.7–19.0 kcal mol^{-1}). On the other hand, previous calculated energy barriers of the acylation reaction for different serine proteases using DFT QM/MM at the B3LYP level have a wide range from 13.2 to 30.1 kcal mol^{-1} [72, 74, 78, 81, 82, 84, 85]. Our calculation results for the NS4A/4B substrate suggest

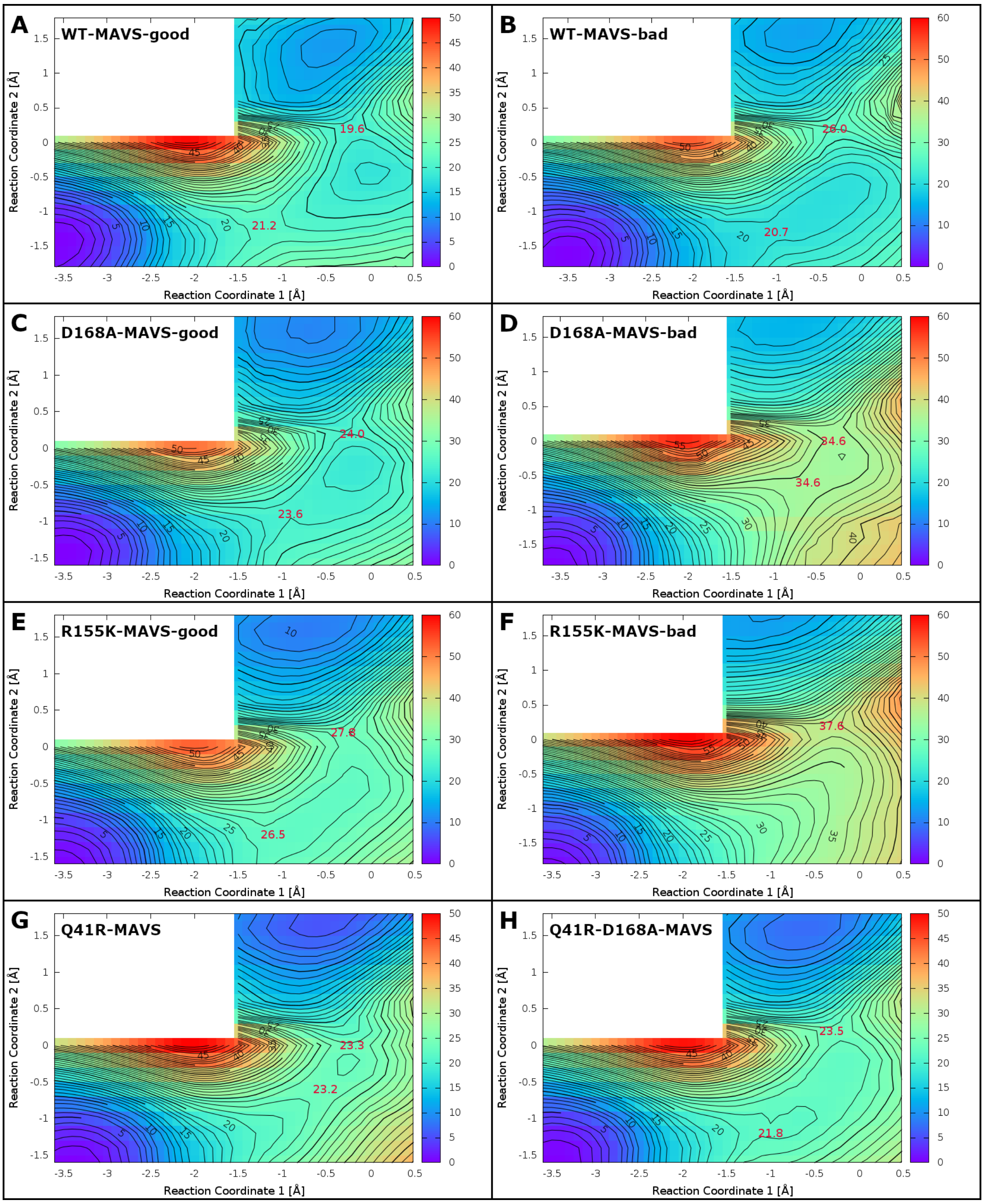


Figure 4-8. Potential energy surfaces of the acylation reaction for HCV NS3/4A protease variants bound to MAVS calculated at the B3LYP/def2-TZVP level. Values of energy barriers are shown in red numbers at positions of transition states. (A) WT with good conformation (B) WT with bad conformation (C) D168A with good conformation (D) D168A with bad conformation (E) R155K with good conformation (F) R155K with bad conformation (G) Q41R (H) Q41R-D168A.

that these NS3/4A protease variants have almost the same energy barriers of the acylation reaction. In addition, these values are very close to calculated values of MAVS systems with good conformation suggesting that when H1’(p) of MAVS is in a proper position,

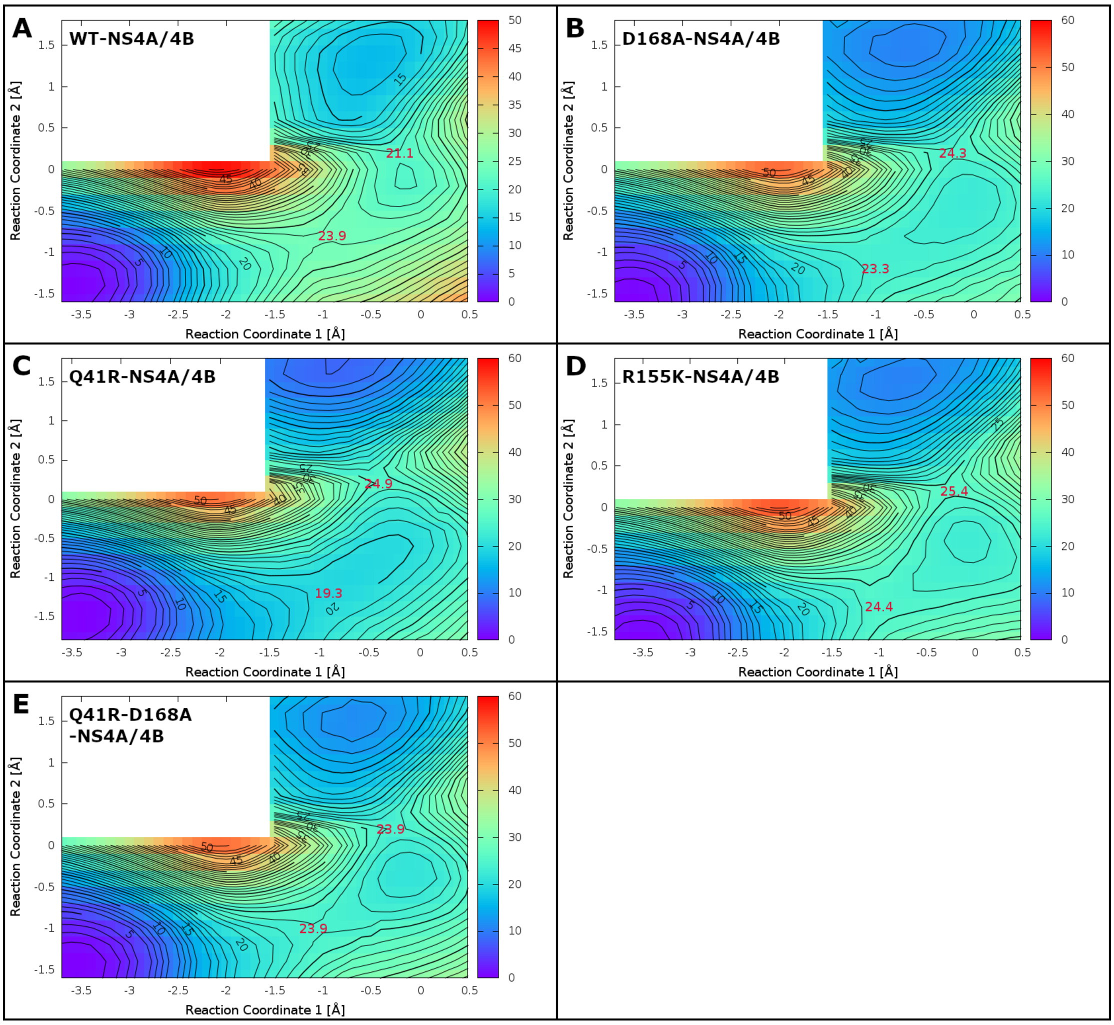


Figure 4-9. Potential energy surfaces of the acylation reaction for HCV NS3/4A protease variants bound to the NS4A/4B substrate calculated at the B3LYP/def2-TZVP level. Values of energy barriers are shown in red numbers at positions of transition states. (A) WT (B) D168A (C) Q41R (D) R155K (E) Q41R-D168A.

NS3/4A protease variants bound to MAVS and the natural substrate NS4A/4B have roughly the same energy barriers of the acylation reaction.

In order to seek the source of higher energy barriers for bad conformations, single point calculations were performed at the B3LYP/def2-TZVP level on the isolate QM region (i.e., without MM point-charges). A comparison of QM and QM/MM energies (Figure 4-10) shows that QM energies are correlated to QM/MM energies, indicating that the higher energy barriers for bad conformations are mainly originating from the QM part. We propose that the repulsion between the negatively charged ND1 atom in H1'(p) and negatively charged oxygen and nitrogen atoms in the reaction center at the transition state is mainly responsible for the higher barrier calculated in the bad conformations.

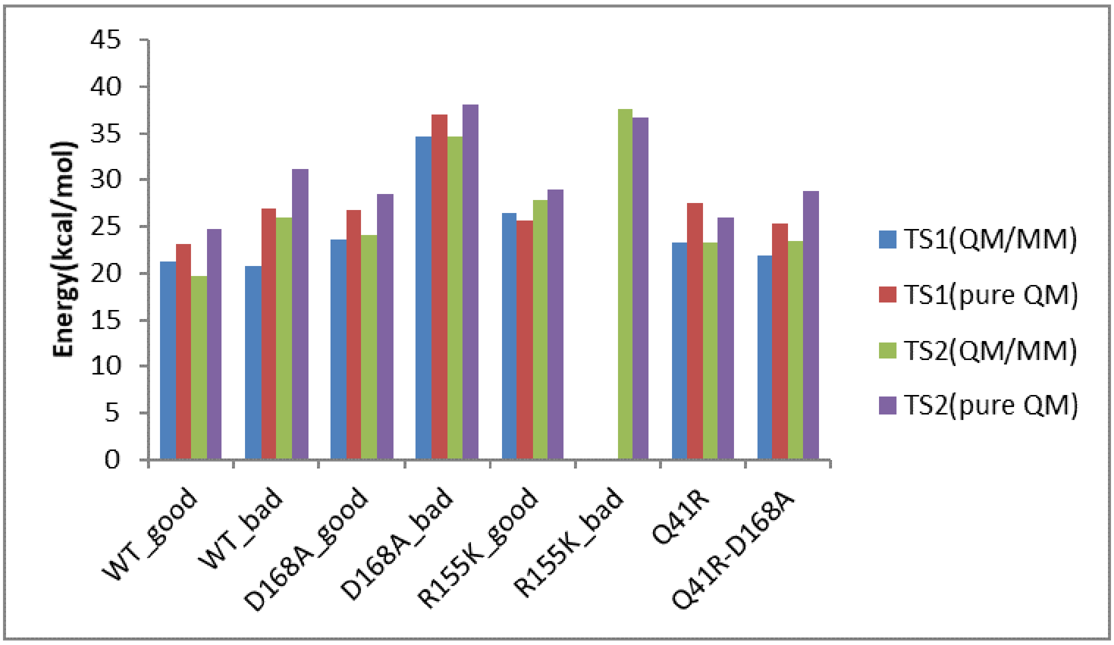


Figure 4-10. Relative QM/MM single point energies and pure QM energies (without MM part) of both transition states for HCV NS3/4A protease variants bound to MAVS.

4.2.4 Characteristic structures of the acylation reaction

Some characteristic structures of the acylation reaction calculated at the B3LYP/def2-TZVP level are shown in Figure 4-11. Mayer's bond orders (BOs) were also calculated from the density matrix elements [234, 235] using an in-house python script to analyze the progress of the acylation reaction. We illustrate the progress using structure corresponding to the wild type protease bound to MAVS in the good conformation. The reaction starts from the reactant state (Figure 4-11A), which is very similar to the initial structure (Figure 4-5A) extracted from the MD simulation. The bond order of the peptide bond at the reactant state is 1.31 showing a partial double bond character, and there is no bond (BO = 0.10) between S139 and the substrate. The nucleophilic attack of S139 to the carbonyl carbon of the ligand occurs in a concerted manner with the proton transfer from S139 to H57. H57 serves as a general base to accept a proton from S139. The negatively charged oxygen in the carbonyl group at the first transition state is stabilized by the backbone NH groups of G137 and S139 (Figure 4-11B). At this transition state, a bond is formed (BO = 0.77) between H57 and the proton and D81 stabilizes the positively charged H57. The bond order of the peptide bond decreases to 1.03 showing that the π conjugation of the bond is disrupted, while a bond between S139 and the substrate is partially formed (BO = 0.56) (Figure 4-11B). The reaction then proceeds to reach a shallow minimum corresponding to the tetrahedral intermediate state (Figure 4-11C). The peptide bond starts breaking (BO = 0.90). Meanwhile, the ester

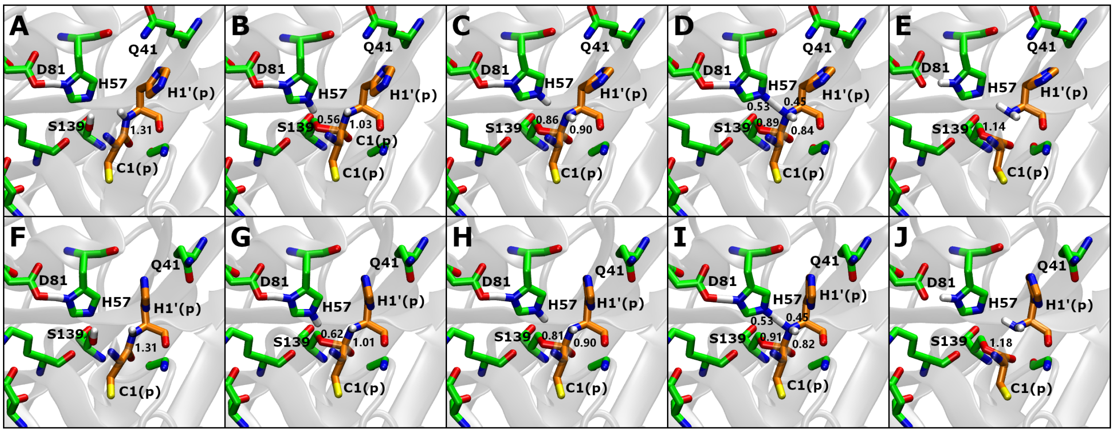


Figure 4-11. Illustration of characteristic structures and some important bond orders for the acylation reaction. (A) the reactant state, (B) the first transition state, (C) the intermediate, (D) the second transition state and (E) the product state of the wild type NS3/4A protease bound to MAVS with good conformation. (F) the reactant state, (G) the first transition state, (H) the intermediate, (I) the second transition state and (J) the product state of the wild type NS3/4A protease bound to MAVS with bad conformation. Carbon atoms in protease are shown in green and carbon atoms in the substrate are shown in orange.

bond is basically formed (BO = 0.86). Afterwards, the system follows to the second transition state, which corresponds to the step of proton transfer from H57 to H1'(p). The proton is roughly in the middle of two histidine residues. The bond orders of the proton with the nitrogen atoms of H57 and H1'(p) are 0.53 and 0.45, respectively. The scissile bond starts to break (BO = 0.84) while the bond order of the ester bond slightly increases to 0.89 (Figure 4-11D). In the end, the scissile bond is fully broken (BO = 0.12) to form a covalently bound complex with the release of the N-terminal peptide (Figure 4-11E). The bond order of the ester bond increases to 1.14 due to the π conjugation. In addition, the first transition state and the intermediate show fully protonated H57. Our calculations indicate that the proton transfer to the substrate happens after the nucleophilic attack of S139 is completed. The wild type NS3/4A protease bound to MAVS with bad conformation shows similar characteristic structures of the reaction (Figure 4-11F-J). Especially, key bond orders at both transition states (Figure 4-11G, I) are nearly identical to those obtained with the good conformation (Figure 4-11B, D). The major difference is the found in the side chain of H1'(p), which indicates once more that the conformation of H1'(p) is the main factor that determines energy barrier of the acylation reaction for NS3/4A protease variants bound to MAVS.

4.2.5 QM/MM FEP calculations

In order to further estimate free energy barriers and include dynamics effect of the surroundings near the active site, calculations of QM/MM free energy perturbation were performed. Due to restrictions of computational resources, we performed these calculations only in one-dimension rather than on the whole surface. We assumed that QM/MM FEP calculations will not change the positions of the stationary points (i.e., reactant, transition states, intermediate, and product) on the potential energy surface. FEP calculations require to choose a path connecting the reactant state to the state of interest and any path should yield the same results because free energy is a state function. We performed a test simulation of the wild type NS3/4A protease bound to MAVS in the bad conformation. The test simulation shows that free energy curves and energy curves share a similar trend (Figure 4-12), which indicates that our assumption mentioned above is reasonable. In addition, the free energy difference between two paths is less than 0.5 kcal mol^{-1} (Figure 4-12). Therefore, the influence of selecting different paths for QM/MM FEP calculations is relatively small.

Calculated free energy barriers and free energy differences of the acylation reaction for NS3/4A protease variants bound to MAVS and the NS4A/4B substrate are listed in Table 4.3. The calculation results show again that for MAVS systems with two conformations, free energies of activation for bad conformation are 5.5–9.3 kcal mol^{-1} higher than those for corresponding good conformation. This translates into a rate constant for bad conformation that is at least 3 orders of magnitudes lower than that for good conformation. The protease-MAVS complexes with bad conformation, therefore, non-reactive in contrast to the good conformation. The calculated free energy barriers

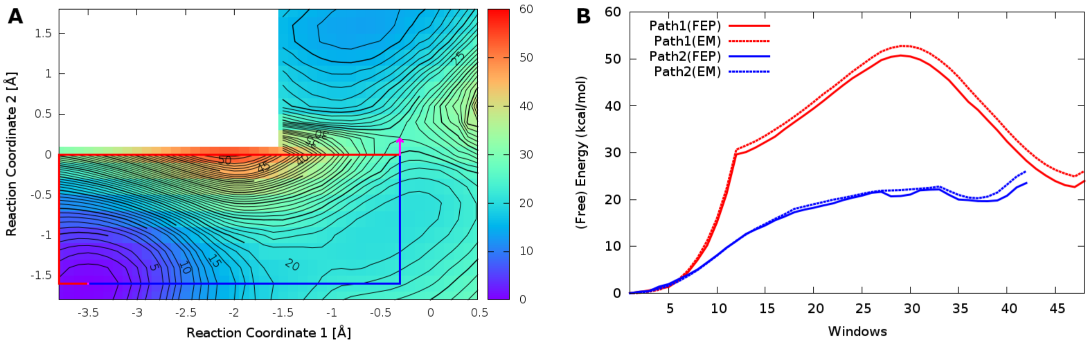


Figure 4-12. (A) Two paths connecting the reactant state and the second transition state on the potential energy surface of the acylation reaction for the wildtype HCV NS3/4A protease bound to MAVS with bad conformation calculated at the B3LYP/def2-TZVP level. (B) Free energy differences (solid lines) calculated by QM/MM FEP and energy differences (dashed lines) calculated by single point QM/MM calculations along two different paths. The scan path is shown in red and another path is shown in blue.

(18.2–22.2 kcal mol^{-1}) are in better agreement with the experimental values (17.7–19.7 kcal mol^{-1}) than static QM/MM values. The trends discussed in the previous section, however, remain unchanged. The results also indicate that two famous PI-resistant mutations, D168A and R155K, have no significant influence on the rate constant of the acylation reaction.

Previous theoretical studies have shown that both nucleophilic attack of serine or proton transfer can be the rate-limiting step [73, 76, 80–85]. The data reported in Table 4.2 show that the rate-limiting step in our simulations varies from one system to another. For the D168A and the Q41 mutants bound to MAVS, the rate-determining step changes from the proton transfer step to the nucleophilic attack step after QM/MM FEP calculations. This was also observed in a previous study when the method changed from HF/MM to MP2/MM [76]. At the moment, it is still not very clear which factor mainly determines the rate-limiting step of the acylation reaction.

4.2.6 Kinetic model

In order to explain the experimental data by our calculation results, we propose a simple kinetic model (Figure 4-13) for HCV NS3/4A protease variants bound to MAVS with two different conformations. The enzyme binds to a substrate yielding a protease-substrate complex which can exist in either good or bad conformation with respect to the H1'(p) residue. There exists an equilibrium between the unbound enzyme, complex with good conformation, and complex with bad conformation with equilibrium constants of K_{M}, K_{C}, and $K_{\mathrm{M}}K_{\mathrm{C}}$. The protease-substrate complex in one of the two conformations can release the product with rate constant of either k_{cat} or k_{bad}. The two conformational states of the protease-MAVS complex are treated as two distinct species.

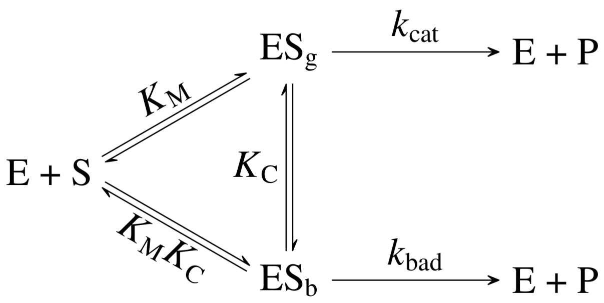


Figure 4-13. Proposed kinetics scheme of the protease-substrate complex with two conformations. K_{M} and K_{C} are equilibrium constants. k_{cat} is the rate constant for the breakdown of ES with good conformation and k_{bad} is the rate constant for the bad conformation.

As the conservation law of the enzyme, the total concentration of the enzymee $[\mathrm{E_t}]$ can be written as:

$$[\mathrm{E_t}] = [\mathrm{E}] + [\mathrm{ES_g}] + [\mathrm{ES_b}] \tag{4-3}$$

where [E], $[\mathrm{ES_g}]$, and $[\mathrm{ES_b}]$ are the concentration of unbound enzyme, the concentration of enzyme-substrate complex with good conformation, and the concentration of enzyme-substrate complex with bad conformation, respectively. The concentration of enzyme-substrate complex with good conformation can be expressed as:

$$[\mathrm{ES_g}] = \frac{[\mathrm{E}] \cdot [\mathrm{S}]}{K_M} \tag{4-4}$$

where [S] is the concentration of the unbound substrate and K_{M} is Michaelis–Menten constant for the enzyme-substrate complex with good conformation. The relationship between concentrations of both conformations can be described by the following equation:

$$[\mathrm{ES_b}] = \frac{[\mathrm{ES_g}]}{K_C} \tag{4-5}$$

where K_{C} is the equilibrium constant and in principle can be estimated by MD simulations of the enzyme-substrate complex. Inserting equations (4-4) and (4-5) into (4-3) yields

$$[\mathrm{E}] = [\mathrm{E_t}] \cdot \frac{1}{1 + (K_C + 1)/(K_M K_C)[\mathrm{S}]} \tag{4-6}$$

The reaction velocity is determined by $[\mathrm{ES_g}]$ and $[\mathrm{ES_b}]$ and can be expressed by:

$$v = k_{cat} \cdot [\mathrm{ES_g}] + k_{bad} \cdot [\mathrm{ES_b}] \tag{4-7}$$

where k_{cat} is the rate constant of enzyme-substrate with good conformations while k_{bad} is the rate constant of the enzyme-substrate complex with bad conformation. The reaction velocity can be expressed further, combining equations (4-4) to (4-7) together:

$$v = [\mathrm{E_t}] \cdot \frac{k_{cat}/K_M[\mathrm{S}] + k_{bad}/(K_M K_C)[\mathrm{S}]}{1 + (K_C + 1)/(K_M K_C)[\mathrm{S}]} \tag{4-8}$$

Since k_{bad} is significantly smaller than k_{cat} based on our QM/MM calculations, equation (4-8) can be simplified to:

$$v = [\mathrm{E_t}] \cdot \frac{k_{cat}[\mathrm{S}]}{K_M + (K_C + 1)/K_C[\mathrm{S}]} = \frac{k_{cat}}{1 + 1/K_C} \cdot [\mathrm{E_t}] \cdot \frac{[\mathrm{S}]}{K_M/(1 + 1/K_C) + [\mathrm{S}]} \tag{4-9}$$

We further define apparent kinetics parameters as follows:

$$k_{cat}^{\text{app}} = \frac{k_{cat}}{1 + 1/K_C} \tag{4-10}$$

and

$$K_M^{\text{app}} = \frac{K_M}{1 + 1/K_C} \tag{4-11}$$

The reaction velocity becomes:

$$v = k_{cat}^{\text{app}} \cdot [\text{E}_\text{t}] \cdot \frac{[\text{S}]}{K_M^{\text{app}} + [\text{S}]} \tag{4-12}$$

has the same form as Michaelis-Menten equation. Therefore, the relationship between real kinetics parameters and experimentally measured apparent kinetics parameters is obtained. Since the denominator in equation (4-10) is larger than 1, the apparent rate constant is smaller than the real rate constant when the system contains a non-reactive enzyme-substrate complex. This is the case for protease variants bound to MAVS with the wild type Q41 residue. For Q41R and Q41R-D168A mutants bound to MAVS, there is no non-reactive enzyme-substrate complex, thus the apparent rate constant is equal to the real rate constant. Within the assumption that all HCV NS3/4A protease variants bound to MAVS have roughly the same real rate constants, our QM/MM calculation combined with this kinetic model provide a valid explanation of experimental observations: i.e., the measured rate constants for wild type and the D168A mutant are smaller than those for the Q41R mutant and the Q41R-D168A mutant (Table 4.2).

The ratio of good and bad conformations in our simulations with Q41 (i.e., WT, R155K, and D168A) is roughly 2:1, which suggest an energy difference of only 0.4 kcal mol^{-1} between the two conformations. The existence of two conformational states of the enzyme-substrate complex influences the kinetics behaviour of the enzyme catalysed reaction. This occurs when two conformational states show similar potential energies around the reactant state while they have quite different potential energies near transition states. In our case, it is caused by the change of the electrostatic environment from the reactant state to the transition state, which induces strong repulsive forces in the latter state. In general, an increase in the population of non-reactive conformations will tend to lower the rate constant of the enzymatic reaction. The kinetics of the enzyme-substrate complex with a non-reactive conformation eventually has the same format as the kinetics of an uncompetitive inhibitor. The presence of non-reactive conformations of the enzyme-substrate complexes thus play a role similar role to that of uncompetitive inhibitors in modulating reaction's kinetics.

4.3 Summary

Classical molecular dynamics combined with QM/MM calculations have been applied to study the acylation reaction of HCV NS3/4A protease variants bound to MAVS and the NS4A/4B substrate. Molecular dynamics simulations of 1 μs with four replicas clearly indicate the existence of two conformational states of the P1' histidine residue of the substrate for wild type NS3/4A protease, D168A mutant, and R155K mutant bound to MAVS. Our QM/MM and QM/MM free energy perturbation calculations indicate that the conformation of H1'(p) is the major factor affecting the energy barrier of the acylation reaction for HCV protease variants bound to MAVS. Two main conformations of this residue are present in all HCV-protease complexed with MAVS, except for variants including the Q41R point mutation. The conformation that is inexistent in the latter mutants was shown to be non-reactive based on our estimation of free energy barriers. The two famous PI-resistant mutations, D168A and R155K, have nearly no impact on the rate constant of the acylation reaction for both MAVS and NS4A/4B systems. Using a simple kinetic model described in this paper, we demonstrate that the existence of a non-reactive conformation of the enzyme-substrate complex eventually lowers the apparent rate constant k_{cat} in a manner similar to the mechanism of action of an uncompetitive inhibitor. We also suggest an explanation at the molecular level for the difference in experimental rate constants between Q41 proteases and Q41R proteases bound to MAVS. In short, Q41R mutants restrict the conformational space of the substrate by strong π-π interactions between R41 and H1'(p) and a hydrogen bond between R41 and H57, which increases the probability of finding the system in a reactive conformation. Although the Q41R mutation is not directly located at the active site, it has a significant effect on the cleavage rate of the acylation reaction. This finding also potential implies that the enzymatic reaction may be modulated through mutating residues which can change the protein's interaction networks or via the interaction of a small molecule with the distant Q41 amino acid.

Chapter 5

Conclusion and Perspectives

Although many previous theoretical research articles used semi-empirical QM methods to study acylation and deacylation reactions for different serine proteases and achieved some successes, our simulation results of acylation reactions for SaClpP systems by semi-empirical QM methods and the DFT method reveal that semi-empirical QM methods are not good enough to simulate acylation reactions of serine proteases. In most cases, simulations conducted by semi-empirical QM methods cannot provide even qualitatively correct results. AM1 and DFTB methods perform better than other semi-empirical QM methods (PM3, PM6, and RM1). But the AM1 method overestimates energy barriers of the acylation step and doesn't provide the correct mechanism of acylation reactions for the phenyl ester and the fluorescent substrate. Potential energy surfaces for the phenyl ester and the fluorescent substrate calculated by the DFTB method show different pattern compared to DFT potential energy surfaces despite similar results obtained for β-lactones.

A solid framework for calculating two-dimensional potential energy surfaces was established to analyze both acylation and deacylation reactions. QM/MM FEP calculations were performed to further estimate free energy barriers. Taking dynamics of the system into account is one of the key factors in successful QM/MM calculations. Calculation results of ClpP systems with our best protocol correlate to experimental values very well. However, our simulations are still limited by the accuracy of calculation results. At the moment, QM/MM calculations can only provide semi-quantitatively correct results. The main factor that determines the kinetics of a reaction can be verified. Based on our calculation results of SaClpP systems, one major factor which influences the energy barrier of both acylation and deacylation reactions for serine proteases is the ability of leaving groups. When a peptide bond is cleaved by a serine protease, the leaving group is an amide anion, which is worse than an alcohol anion. Thus the energy barrier of the acylation step for the fluorescent substrate is higher than that of the deacylation step. This also agrees with the general concept that the acylation step is the rate-determining step for hydrolysis reactions with natural substrates catalyzed by a serine protease. On the contrary, the deacylation step is the rate-determining step for phenyl ester-type inhibitors because the phenol anion is a good leaving group compared to the alcohol anion. The opening of the four-membered ring for β-lactones can be treated as the departure of a very good leaving group. Therefore, the deacylation step is also the rate-determining step for β-lactones.

Regarding simulations of HCV NS3/4A protease variants, when the substrate is the natural substrate NS4A/4B, the cleavage capability is basically not changed when the protease is mutated (Q41R, R155K, D168A, and Q41R-D168A). When MAVS is the substrate, MD simulations reveal that there are two distinct conformations of the histidine residue at the P1' position. Our QM/MM calculation results indicate that the HCV NS3/4A protease bound to MAVS with bad conformation has a higher free energy barrier than the corresponding good conformation. The existence of a non-reactive conformation can be seen as an uncompetitive inhibitor. The protease variant with the Q41R mutant relatively increases the rate constant of the acylation reaction by ruling out the bad conformation. This might explain the source of the competitive advantage of the Q41R mutant *in vivo*. On the other hand, two famous PI-resistance mutants, R155K and D168A, have no direct influence on the kinetics of the acylation reaction for MAVS.

Despite some technical differences (force fields and the radius of cut spheres) between QM/MM simulations for ClpP and HCV systems, our simulation framework of two-dimensional potential energy surfaces shows a general pattern of the acylation reaction for both systems and the deacylation reaction for ClpP. The reaction always starts from the bottom left corner of the surface, passing through one or two transition states, to the product state at the upper right part. The area around the center of the surface always has very high energies. All acylation reactions involving a peptide bond show two transition states and one intermediate although the energy gap between the intermediate and the lower transition state is very tiny in some cases. Whether an intermediate exists for the deacylation step is still not very clear at the moment.

QM/MM technique is, in general, a very powerful tool to study the mechanism of enzyme-catalyzed reactions. It helps people to explain experimental phenomenon and acquire more information about biomolecular systems. Based on QM/MM simulations conducted in the present study, the dynamics of the entire system is very important to accurate predict the kinetics behavior of both acylation and deacylation reactions. We believe that the further development of the QM/MM method should also combine with molecular dynamics simulations of the system somehow. A long QM/MM MD simulation is still a challenge under current cost restrictions. We think there are two paths for the future development of the QM/MM method. The first one is the further development of the hardware. Long DFT or higher level QM/MM MD might become affordable. Another potential path is to parameterize a semi-empirical method for a specific system by fitting the potential energy surface to the higher level QM method. Then people can obtain accurate results with relatively low computational cost by performing re-parameterized semi-empirical QM/MM MD simulations.

Bibliography

[1] K. Oda. New families of carboxyl peptidases: serine-carboxyl peptidases and glutamic peptidases. *J. Biochem.*, 151(1):13–25, 2012.

[2] L. Hedstrom. Serine protease mechanism and specificity. *Chem. Rev.*, 102(12):4501–24, 2002.

[3] D. M. Blow. The tortuous story of Asp…His…Ser: Structural analysis of α-chymotrypsin. *Trends Biochem. Sci.*, 22(10):405–408, 1997.

[4] S. Ba-Saif, A. K. Luthra, and A. Williams. Concerted acetyl-group transfer between substituted phenolate ion nucleophiles: variation of transition-state structure as a function of substituent. *J. Am. Chem. Soc.*, 111(7):2647–2652, 1989.

[5] D. Stefanidis, S. Cho, S. Dhe-Paganon, and W. P. Jencks. Structure-reactivity correlations for reactions of substituted phenolate anions with acetate and formate esters. *J. Am. Chem. Soc.*, 115(5):1650–1656, 1993.

[6] A. C. Hengge and R. A. Hess. Concerted or Stepwise Mechanisms for Acyl Transfer Reactions of p-Nitrophenyl Acetate? Transition State Structures from Isotope Effects. *J. Am. Chem. Soc.*, 116(25):11256–11263, 1994.

[7] W. W. Cleland and A. C. Hengge. Mechanisms of phosphoryl and acyl transfer. *FASEB J.*, 9(15):1585–94, 1995.

[8] J. Á. Martínez-González, M. González, L. Masgrau, and R. Martínez. Theoretical Study of the Free Energy Surface and Kinetics of the Hepatitis C Virus NS3/NS4A Serine Protease Reaction with the NS5A/5B Substrate. Does the Generally Accepted Tetrahedral Intermediate Really Exist? *ACS Catal.*, 5(1):246–255, 2014.

[9] P. Bross, B. S. Andresen, I. Knudsen, T. A. Kruse, and N. Gregersen. Human ClpP protease: cDNA sequence, tissue-specific expression and chromosomal assignment of the gene. *FEBS Lett.*, 377(2):249–252, 1995.

[10] A. Y. Yu and W. A. Houry. ClpP: a distinctive family of cylindrical energy-dependent serine proteases. *FEBS Lett.*, 581(19):3749–57, 2007.

[11] S. Gottesman. Regulation by proteolysis: Developmental switches. *Curr. Opin. Microbiol.*, 2(2):142–147, 1999.

[12] S. Wickner, M. R. Maurizi, and S. Gottesman. Posttranslational quality control: folding, refolding, and degrading proteins. *Science*, 286(5446):1888–93, 1999.

[13] Y. Katayama-Fujimura, S. Gottesman, and M. R. Maurizi. A multiple-component, ATP-dependent protease from Escherichia coli. *J. Biol. Chem.*, 262(10):4477–4485, 1987.

[14] D. Frees, S. N. Qazi, P. J. Hill, and H. Ingmer. Alternative roles of ClpX and ClpP in Staphylococcus aureus stress tolerance and virulence. *Mol. Microbiol.*, 48(6):1565–78, 2003.

[15] D. Frees, K. Sorensen, and H. Ingmer. Global virulence regulation in Staphylococcus aureus: pinpointing the roles of ClpP and ClpX in the sar/agr regulatory network. *Infect. Immun.*, 73(12):8100–8, 2005.

[16] M. R. Maurizi, M. W. Thompson, S. K. Singh, and S. H. Kim. Endopeptidase Clp: ATP-dependent Clp protease from Escherichia coli. *Methods Enzymol.*, 244:314–31, 1994.

[17] S. Gottesman, W. P. Clark, V. de Crecy-Lagard, and M. R. Maurizi. ClpX, an alternative subunit for the ATP-dependent Clp protease of Escherichia coli. Sequence and in vivo activities. *J. Biol. Chem.*, 268(30):22618–22626, 1993.

[18] S. Wickner, S. Gottesman, D. Skowyra, J. Hoskins, K. McKenney, and M. R. Maurizi. A molecular chaperone, ClpA, functions like DnaK and DnaJ. *Proc. Natl. Acad. Sci. U.S.A.*, 91(25):12218–22, 1994.

[19] Y. I. Kim, I. Levchenko, K. Fraczkowska, R. V. Woodruff, R. T. Sauer, and T. A. Baker. Molecular determinants of complex formation between Clp/Hsp100 ATPases and the ClpP peptidase. *Nat. Struct. Biol.*, 8(3):230–3, 2001.

[20] S. A. Joshi, G. L. Hersch, T. A. Baker, and R. T. Sauer. Communication between ClpX and ClpP during substrate processing and degradation. *Nat. Struct. Mol. Biol.*, 11(5):404–11, 2004.

[21] T. A. Baker and R. T. Sauer. ClpXP, an ATP-powered unfolding and protein-degradation machine. *Biochim. Biophys. Acta.*, 1823(1):15–28, 2012.

[22] S. Gottesman, E. Roche, Y. Zhou, and R. T. Sauer. The ClpXP and ClpAP proteases degrade proteins with carboxy-terminal peptide tails added by the SsrA-tagging system. *Genes Dev.*, 12(9):1338–47, 1998.

[23] S. R. Geiger, T. Bottcher, S. A. Sieber, and P. Cramer. A conformational switch underlies ClpP protease function. *Angew. Chem. Int. Ed.*, 50(25):5749–52, 2011.

[24] J. Zhang, F. Ye, L. Lan, H. Jiang, C. Luo, and C. G. Yang. Structural switching of

Staphylococcus aureus Clp protease: a key to understanding protease dynamics. *J. Biol. Chem.*, 286(43):37590–601, 2011.

[25] M. Gersch, A. List, M. Groll, and S. A. Sieber. Insights into structural network responsible for oligomerization and activity of bacterial virulence regulator caseinolytic protease P (ClpP) protein. *J. Biol. Chem.*, 287(12):9484–94, 2012.

[26] F. Ye, J. Zhang, H. Liu, R. Hilgenfeld, R. Zhang, X. Kong, L. Li, J. Lu, X. Zhang, D. Li, H. Jiang, C. G. Yang, and C. Luo. Helix unfolding/refolding characterizes the functional dynamics of Staphylococcus aureus Clp protease. *J. Biol. Chem.*, 288(24):17643–53, 2013.

[27] A. P. Magiorakos, A. Srinivasan, R. B. Carey, Y. Carmeli, M. E. Falagas, C. G. Giske, S. Harbarth, J. F. Hindler, G. Kahlmeter, B. Olsson-Liljequist, D. L. Paterson, L. B. Rice, J. Stelling, M. J. Struelens, A. Vatopoulos, J. T. Weber, and D. L. Monnet. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin. Microbiol. Infect.*, 18(3):268–81, 2012.

[28] T. Bottcher and S. A. Sieber. Beta-lactones as privileged structures for the active-site labeling of versatile bacterial enzyme classes. *Angew. Chem. Int. Ed.*, 47(24):4600–3, 2008.

[29] T. Bottcher and S. A. Sieber. Beta-lactones as specific inhibitors of ClpP attenuate the production of extracellular virulence factors of Staphylococcus aureus. *J. Am. Chem. Soc.*, 130(44):14400–1, 2008.

[30] T. Bottcher and S. A. Sieber. Structurally refined beta-lactones as potent inhibitors of devastating bacterial virulence factors. *Chembiochem*, 10(4):663–6, 2009.

[31] M. W. Hackl, M. Lakemeyer, M. Dahmen, M. Glaser, A. Pahl, K. Lorenz-Baath, T. Menzel, S. Sievers, T. Bottcher, I. Antes, H. Waldmann, and S. A. Sieber. Phenyl Esters Are Potent Inhibitors of Caseinolytic Protease P and Reveal a Stereogenic Switch for Deoligomerization. *J. Am. Chem. Soc.*, 137(26):8475–83, 2015.

[32] M. Gersch, R. Kolb, F. Alte, M. Groll, and S. A. Sieber. Disruption of oligomerization and dehydroalanine formation as mechanisms for ClpP protease inhibition. *J. Am. Chem. Soc.*, 136(4):1360–6, 2014.

[33] J. Wang, J. A. Hartling, and J. M. Flanagan. The Structure of ClpP at 2.3 Å Resolution Suggests a Model for ATP-Dependent Proteolysis. *Cell*, 91(4):447–456, 1997.

[34] S. G. Kang, M. N. Dimitrova, J. Ortega, A. Ginsburg, and M. R. Maurizi. Human

mitochondrial ClpP is a stable heptamer that assembles into a tetradecamer in the presence of ClpX. *J. Biol. Chem.*, 280(42):35424–32, 2005.

[35] M. S. Kimber, A. Y. Yu, M. Borg, E. Leung, H. S. Chan, and W. A. Houry. Structural and theoretical studies indicate that the cylindrical protease ClpP samples extended and compact conformations. *Structure*, 18(7):798–808, 2010.

[36] B. G. Lee, E. Y. Park, K. E. Lee, H. Jeon, K. H. Sung, H. Paulsen, H. Rubsamen-Schaeff, H. Brotz-Oesterhelt, and H. K. Song. Structures of ClpP in complex with acyldepsipeptide antibiotics reveal its activation mechanism. *Nat. Struct. Mol. Biol.*, 17(4):471–8, 2010.

[37] World Health Organization. *Global Hepatitis Report 2017*. Geneva, 2017.

[38] N. Toshikuni, T. Arisawa, and M. Tsutsumi. Hepatitis C-related liver cirrhosis - strategies for the prevention of hepatic decompensation, hepatocarcinogenesis, and mortality. *World J. Gastroenterol.*, 20(11):2876–87, 2014.

[39] I. Rusyn and S. M. Lemon. Mechanisms of HCV-induced liver cancer: what did we learn from in vitro and animal studies? *Cancer Lett.*, 345(2):210–5, 2014.

[40] C. Ferri, M. Sebastiani, D. Giuggioli, M. Colaci, P. Fallahi, A. Piluso, A. Antonelli, and A. L. Zignego. Hepatitis C virus syndrome: A constellation of organ- and non-organ specific autoimmune disorders, B-cell non-Hodgkin's lymphoma, and cancer. *World J. Hepatol.*, 7(3):327–43, 2015.

[41] L. Chatel-Chaix, M. Baril, and D. Lamarre. Hepatitis C Virus NS3/4A Protease Inhibitors: A Light at the End of the Tunnel. *Viruses*, 2(8):1752–65, 2010.

[42] N. J. Liverton. Evolution of HCV NS3/4a Protease Inhibitors. In *HCV: The Journey from Discovery to a Cure*, Topics in Medicinal Chemistry, pages 231–259. 2019.

[43] R. B. Seth, L. Sun, C. K. Ea, and Z. J. Chen. Identification and characterization of MAVS, a mitochondrial antiviral signaling protein that activates NF-kappaB and IRF 3. *Cell*, 122(5):669–82, 2005.

[44] T. Kawai, K. Takahashi, S. Sato, C. Coban, H. Kumar, H. Kato, K. J. Ishii, O. Takeuchi, and S. Akira. IPS-1, an adaptor triggering RIG-I- and Mda5-mediated type I interferon induction. *Nat. Immunol.*, 6(10):981–8, 2005.

[45] L. G. Xu, Y. Y. Wang, K. J. Han, L. Y. Li, Z. Zhai, and H. B. Shu. VISA is an adapter protein required for virus-triggered IFN-beta signaling. *Mol. Cell*, 19(6):727–40, 2005.

[46] E. Meylan, J. Curran, K. Hofmann, D. Moradpour, M. Binder, R. Bartenschlager,

and J. Tschopp. Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. *Nature*, 437(7062):1167–72, 2005.

[47] X. D. Li, L. Sun, R. B. Seth, G. Pineda, and Z. J. Chen. Hepatitis C virus protease NS3/4A cleaves mitochondrial antiviral signaling protein off the mitochondria to evade innate immunity. *Proc. Natl. Acad. Sci. U.S.A.*, 102(49):17717–22, 2005.

[48] S. M. Horner, H. M. Liu, H. S. Park, J. Briley, and Jr. Gale, M. Mitochondrial-associated endoplasmic reticulum membranes (MAM) form innate immune synapses and are targeted by hepatitis C virus. *Proc. Natl. Acad. Sci. U.S.A.*, 108(35):14590–5, 2011.

[49] M. Yoneyama, M. Kikuchi, T. Natsukawa, N. Shinobu, T. Imaizumi, M. Miyagishi, K. Taira, S. Akira, and T. Fujita. The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat. Immunol.*, 5(7):730–7, 2004.

[50] S. Akira, S. Uematsu, and O. Takeuchi. Pathogen recognition and innate immunity. *Cell*, 124(4):783–801, 2006.

[51] H. Kato, O. Takeuchi, S. Sato, M. Yoneyama, M. Yamamoto, K. Matsui, S. Uematsu, A. Jung, T. Kawai, K. J. Ishii, O. Yamaguchi, K. Otsu, T. Tsujimura, C. S. Koh, C. Reis e Sousa, Y. Matsuura, T. Fujita, and S. Akira. Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature*, 441(7089):101–5, 2006.

[52] T. Kawai and S. Akira. Innate immune recognition of viral infection. *Nat. Immunol.*, 7(2):131–7, 2006.

[53] Y. M. Loo, J. Fornek, N. Crochet, G. Bajwa, O. Perwitasari, L. Martinez-Sobrido, S. Akira, M. A. Gill, A. Garcia-Sastre, M. G. Katze, and Jr. Gale, M. Distinct RIG-I and MDA5 signaling by RNA viruses in innate immunity. *J. Virol.*, 82(1):335–45, 2008.

[54] O. Takeuchi and S. Akira. Innate immunity to virus infection. *Immunol. Rev.*, 227(1):75–86, 2009.

[55] M. Yoneyama and T. Fujita. RNA recognition and signal transduction by RIG-I-like receptors. *Immunol. Rev.*, 227(1):54–65, 2009.

[56] Y. M. Loo and Jr. Gale, M. Immune signaling by RIG-I-like receptors. *Immunity*, 34(5):680–92, 2011.

[57] A. Le Bon and D. F. Tough. Links between innate and adaptive immunity via type I interferon. *Curr. Opin. Immunol.*, 14(4):432–436, 2002.

[58] A. Iwasaki and R. Medzhitov. Regulation of adaptive immunity by the innate immune system. *Science*, 327(5963):291–5, 2010.

[59] C. Lin, A. D. Kwong, and R. B. Perni. Discovery and development of VX-950, a novel, covalent, and reversible inhibitor of hepatitis C virus NS3.4A serine protease. *Infect. Disord. Drug Targets*, 6(1):3–16, 2006.

[60] P. Revill, N. Serradell, J. Bolós, and E. Rosa. Telaprevir. *Drugs Future*, 32(9), 2007.

[61] B. Degertekin and A. S. Lok. Update on viral hepatitis: 2007. *Curr. Opin. Gastroenterol.*, 24(3):306–11, 2008.

[62] F. G. Njoroge, K. X. Chen, N. Y. Shih, and J. J. Piwinski. Challenges in modern drug discovery: a case study of boceprevir, an HCV protease inhibitor for the treatment of hepatitis C virus infection. *Acc. Chem. Res.*, 41(1):50–9, 2008.

[63] X. Tong, A. Arasappan, F. Bennett, R. Chase, B. Feld, Z. Guo, A. Hart, V. Madison, B. Malcolm, J. Pichardo, A. Prongay, R. Ralston, A. Skelton, E. Xia, R. Zhang, and F. G. Njoroge. Preclinical characterization of the antiviral activity of SCH 900518 (narlaprevir), a novel mechanism-based inhibitor of hepatitis C virus NS3 protease. *Antimicrob. Agents Chemother.*, 54(6):2365–70, 2010.

[64] C. Reviriego. Asunaprevir. *Drugs Future*, 37(4), 2012.

[65] T. I. Lin, O. Lenz, G. Fanning, T. Verbinnen, F. Delouvroy, A. Scholliers, K. Vermeiren, A. Rosenquist, M. Edlund, B. Samuelsson, L. Vrang, H. de Kock, P. Wigerinck, P. Raboisson, and K. Simmen. In vitro activity and preclinical profile of TMC435350, a potent hepatitis C virus protease inhibitor. *Antimicrob. Agents Chemother.*, 53(4):1377–85, 2009.

[66] J. A. McCauley, C. J. McIntyre, M. T. Rudd, K. T. Nguyen, J. J. Romano, J. W. Butcher, K. F. Gilbert, K. J. Bush, M. K. Holloway, J. Swestock, B. L. Wan, S. S. Carroll, J. M. DiMuzio, D. J. Graham, S. W. Ludmerer, S. S. Mao, M. W. Stahlhut, C. M. Fandozzi, N. Trainor, D. B. Olsen, J. P. Vacca, and N. J. Liverton. Discovery of vaniprevir (MK-7009), a macrocyclic hepatitis C virus NS3/4a protease inhibitor. *J. Med. Chem.*, 53(6):2443–63, 2010.

[67] S. Harper, J. A. McCauley, M. T. Rudd, M. Ferrara, M. DiFilippo, B. Crescenzi, U. Koch, A. Petrocchi, M. K. Holloway, J. W. Butcher, J. J. Romano, K. J. Bush, K. F. Gilbert, C. J. McIntyre, K. T. Nguyen, E. Nizi, S. S. Carroll, S. W. Ludmerer, C. Burlein, J. M. DiMuzio, D. J. Graham, C. M. McHale, M. W. Stahlhut, D. B. Olsen, E. Monteagudo, S. Cianetti, C. Giuliano, V. Pucci, N. Trainor, C. M. Fandozzi, M. Rowley, P. J. Coleman, J. P. Vacca, V. Summa, and N. J. Liver-

ton. Discovery of MK-5172, a Macrocyclic Hepatitis C Virus NS3/4a Protease Inhibitor. *ACS Med. Chem. Lett.*, 3(4):332–6, 2012.

[68] I. Gentile, F. Borgia, A. R. Buonomo, E. Zappulo, G. Castaldo, and G. Borgia. ABT-450: a novel protease inhibitor for the treatment of hepatitis C virus infection. *Curr. Med. Chem.*, 21(28):3261–70, 2014.

[69] C. Zheng, M. Schneider, A. Marion, and I. Antes. The Q41R mutation in the HCV-protease enhances the reactivity towards MAVS by suppressing non-reactive pathways. *Phys. Chem. Chem. Phys.*, 24(4):2126–2138, 2022.

[70] K. P. Romano, J. M. Laine, L. M. Deveau, H. Cao, F. Massi, and C. A. Schiffer. Molecular mechanisms of viral and host cell substrate recognition by hepatitis C virus NS3/4A protease. *J. Virol.*, 85(13):6106–16, 2011.

[71] A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227–249, 1976.

[72] Y. Zhang, J. Kua, and J. A. McCammon. Role of the Catalytic Triad and Oxyanion Hole in Acetylcholinesterase Catalysis: An ab initio QM/MM Study. *J. Am. Chem. Soc.*, 124(35):10572–10577, 2002.

[73] T. Ishida and S. Kato. Theoretical Perspectives on the Reaction Mechanism of Serine Proteases: The Reaction Free Energy Profiles of the Acylation Process. *J. Am. Chem. Soc.*, 125(39):12035–12048, 2003.

[74] M. P. Gleeson, I. H. Hillier, and N. A. Burton. Theoretical analysis of peptidyl α-ketoheterocyclic inhibitors of human neutrophil elastase: Insight into the mechanism of inhibition and the application of QM/MM calculations in structure-based drug design. *Org. Biomol. Chem.*, 2(16):2275–2280, 2004.

[75] T. Ishida and S. Kato. Role of Asp102 in the Catalytic Relay System of Serine Proteases: A Theoretical Study. *J. Am. Chem. Soc.*, 126(22):7111–7118, 2004.

[76] C. Oliva, A. Rodríguez, M. González, and W. Yang. A quantum mechanics/molecular mechanics study of the reaction mechanism of the hepatitis C virus NS3 protease with the NS5A/5B substrate. *Proteins*, 66(2):444–455, 2006.

[77] A. Rodríguez, C. Oliva, M. González, M. van der Kamp, and A. J. Mulholland. Comparison of different quantum mechanical/molecular mechanics boundary treatments in the reaction of the hepatitis C virus NS3 protease with the NS5A/5B substrate. *J. Phys. Chem. B*, 111(44):12909–15, 2007.

[78] T. Rungrotmongkol, P. Decha, P. Sompornpisut, M. Malaisree, P. Intharathep, N. Nunthaboot, T. Udommaneethanakit, O. Aruksakunwong, and S. Hannongbua.

Combined QM/MM mechanistic study of the acylation process in furin complexed with the H5N1 avian influenza virus hemagglutinin's cleavage site. *Proteins*, 76(1):62–71, 2009.

[79] L. Bellucci, T. Laino, A. Tafi, and M. Botta. Metadynamics Simulations of Enantioselective Acylation Give Insights into the Catalytic Mechanism of Burkholderia cepacia Lipase. *J. Chem. Theory Comput.*, 6(4):1145–1156, 2010.

[80] A. Rodríguez, C. Oliva, and M. González. A comparative QM/MM study of the reaction mechanism of the Hepatitis C virus NS3/NS4A protease with the three main natural substrates NS5A/5B, NS4B/5A and NS4A/4B. *Phys. Chem. Chem. Phys.*, 12(28):8001, 2010.

[81] Y. Zhou, S. Wang, and Y. Zhang. Catalytic Reaction Mechanism of Acetylcholinesterase Determined by Born-Oppenheimer Ab Initio QM/MM Molecular Dynamics Simulations. *J. Phys. Chem. B*, 114(26):8817–8825, 2010.

[82] Y. Zhou and Y. Zhang. Serine protease acylation proceeds with a subtle reorientation of the histidine ring at the tetrahedral intermediate. *Chem. Commun.*, 47(5):1577–1579, 2011.

[83] M. C. P. Lima and G. M. Seabra. Reaction mechanism of the dengue virus serine protease: a QM/MM study. *Phys. Chem. Chem. Phys*, 18(44):30288–30296, 2016.

[84] Y. Zhou, D. Xie, and Y. Zhang. Amide Rotation Hindrance Predicts Proteolytic Resistance of Cystine-Knot Peptides. *J. Phys. Chem. Lett.*, 7(7):1138–1142, 2016.

[85] W. Wei, Y. Chen, D. Xie, and Y. Zhou. Molecular insight into chymotrypsin inhibitor 2 resisting proteolytic degradation. *Phys. Chem. Chem. Phys.*, 21(9):5049–5058, 2019.

[86] G. Jindal, D. Mondal, and A. Warshel. Exploring the Drug Resistance of HCV Protease. *J. Phys. Chem. B*, 121(28):6831–6840, 2017.

[87] M. Topf and W. G. Richards. Theoretical studies on the deacylation step of serine protease catalysis in the gas phase, in solution, and in elastase. *J. Am. Chem. Soc.*, 126(44):14631–41, 2004.

[88] G. Dultz, T. Shimakami, M. Schneider, K. Murai, D. Yamane, A. Marion, T. M. Zeitler, C. Stross, C. Grimm, R. M. Richter, K. Baumer, M. Yi, R. M. Biondi, S. Zeuzem, R. Tampe, I. Antes, C. M. Lange, and C. Welsch. Extended interaction networks with HCV protease NS3-4A substrates explain the lack of adaptive capability against protease inhibitors. *J. Biol. Chem.*, 295(40):13862–13874, 2020.

[89] E. Schrödinger. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys. Rev.*, 28(6):1049–1070, 1926.

[90] M. Born and R. Oppenheimer. Zur Quantentheorie der Molekeln. *Ann. Phys.*, 389(20):457–484, 1927.

[91] V. Fock. Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems. *Z. Phys.*, 61(1-2):126–148, 1930.

[92] V. Fock. "Selfconsistent field" mit Austausch für Natrium. *Z. Phys.*, 62(11-12):795–805, 1930.

[93] W. Hartree, D. R.; Hartree. Self-consistent field, with exchange, for beryllium. *Proc. R. Soc. Lond. A Math. Phys. Sci.*, 150(869):9–33, 1935.

[94] J. C. Slater. The Theory of Complex Spectra. *Phys. Rev.*, 34(10):1293–1322, 1929.

[95] R. G. Parr, R. Breslow, and M. Karplus. *The Quantum Theory of Molecular Electronic Structure*. Benjamin, 1963.

[96] C. C. J. Roothaan. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.*, 23(2):69–89, 1951.

[97] I. N. Levine. *Quantum Chemistry*. Pearson Education, Boston, 7th edition, 2014.

[98] J. A. Pople, D. P. Santry, and G. A. Segal. Approximate Self-Consistent Molecular Orbital Theory. I. Invariant Procedures. *J. Chem. Phys.*, 43(10):S129–S135, 1965.

[99] J. A. Pople, D. L. Beveridge, and P. A. Dobosh. Approximate Self-Consistent Molecular-Orbital Theory. V. Intermediate Neglect of Differential Overlap. *J. Chem. Phys.*, 47(6):2026–2033, 1967.

[100] J. A. Pople and G. A. Segal. Approximate Self-Consistent Molecular Orbital Theory. II. Calculations with Complete Neglect of Differential Overlap. *J. Chem. Phys.*, 43(10):S136–S151, 1965.

[101] M. J. S. Dewar and W. Thiel. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.*, 99(15):4899–4907, 1977.

[102] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, and J. J. P. Stewart. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.*, 107(13):3902–3909, 1985.

[103] J. J. P. Stewart. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.*, 10(2):209–220, 1989.

[104] J. J. Stewart. Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J. Mol. Model.*, 13(12):1173–213, 2007.

[105] J. Řezáč and P. Hobza. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.*, 8(1):141–51, 2012.

[106] J. J. Stewart. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.*, 19(1):1–32, 2013.

[107] G. B. Rocha, R. O. Freire, A. M. Simas, and J. J. Stewart. RM1: a reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.*, 27(10):1101–11, 2006.

[108] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, Th Frauenheim, S. Suhai, and G. Seifert. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B*, 58(11):7260–7268, 1998.

[109] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.*, 136(3B):B864–B871, 1964.

[110] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140(4A):A1133–A1138, 1965.

[111] R. G. Parr and Y. Weitao. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, 1989.

[112] F. Bloch. Bemerkung zur Elektronentheorie des Ferromagnetismus und der elektrischen Leitfähigkeit. *Z. Phys.*, 57(7-8):545–555, 1929.

[113] P. A. M. Dirac. Note on Exchange Phenomena in the Thomas Atom. *Math. Proc. Camb. Philos. Soc.*, 26(3):376–385, 1930.

[114] D. M. Ceperley and B. J. Alder. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.*, 45(7):566–569, 1980.

[115] S. H. Vosko, L. Wilk, and M. Nusair. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.*, 58(8):1200–1211, 1980.

[116] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A Gen. Phys.*, 38(6):3098–3100, 1988.

[117] J. P. Perdew. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B*, 33(12):8822–8824, 1986.

[118] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37(2):785–789, 1988.

[119] J. P. Perdew and Y. Wang. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B*, 45(23):13244–13249, 1992.

[120] A. D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98(7):5648–5652, 1993.

[121] K. Kim and K. D. Jordan. Comparison of Density Functional and MP2 Calculations on the Water Monomer and Dimer. *J. Phys. Chem.*, 98(40):10089–10094, 1994.

[122] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.*, 98(45):11623–11627, 1994.

[123] S. Grimme. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.*, 124(3):034108, 2006.

[124] L. Goerigk and S. Grimme. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals-Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.*, 7(2):291–309, 2011.

[125] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104, 2010.

[126] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *J. Chem. Phys.*, 72(1):650–654, 1980.

[127] A. D. McLean and G. S. Chandler. Contracted Gaussian basis sets for molecular calculations. I. Second row atoms, Z=11–18. *J. Chem. Phys.*, 72(10):5639–5648, 1980.

[128] R. Ditchfield, W. J. Hehre, and J. A. Pople. Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *J. Chem. Phys.*, 54(2):724–728, 1971.

[129] W. J. Hehre, R. Ditchfield, and J. A. Pople. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.*, 56(5):2257–2261, 1972.

[130] P. C. Hariharan and J. A. Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theor. Chim. Acta*, 28(3):213–222, 1973.

[131] A. Schäfer, H. Horn, and R. Ahlrichs. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *J. Chem. Phys.*, 97(4):2571–2577, 1992.

[132] A. Schäfer, C. Huber, and R. Ahlrichs. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *J. Chem. Phys.*, 100(8):5829–5835, 1994.

[133] F. Weigend and R. Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–305, 2005.

[134] F. Weigend. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.*, 8(9):1057–65, 2006.

[135] TURBOMOLE V7.1 2016, a development of University of Karlsruhe and-Forschungszentrum Karlsruhe GmbH, 1989-2007,TURBOMOLE GmbH, since 2007; available from `http://www.turbomole.com`.

[136] T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.*, 90(2):1007–1023, 1989.

[137] R. A. Kendall, T. H. Dunning, and R. J. Harrison. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.*, 96(9):6796–6806, 1992.

[138] D. E. Woon and T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.*, 98(2):1358–1371, 1993.

[139] K. A. Peterson, D. E. Woon, and T. H. Dunning. Benchmark calculations with correlated molecular wave functions. IV. The classical barrier height of the H+H2→H2+H reaction. *J. Chem. Phys.*, 100(10):7410–7415, 1994.

[140] A. K. Wilson, T. van Mourik, and T. H. Dunning. Gaussian basis sets for use in correlated molecular calculations. VI. Sextuple zeta correlation consistent basis sets for boron through neon. *J. Mol. Struct. THEOCHEM*, 388:339–349, 1996.

[141] F. Jensen. *Introduction to computational chemistry*. Wiley, 2 edition, 2007.

[142] D. A. Case, J. T. Berryman, R. M. Betz, D. S. Cerutti, III T. E. Cheatham, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, T. Luchko,

R. Luo, B. Madej, K. M. Merz, G. Monard, P. Needham, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, R. Salomon-Ferrer, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, D. M. York, and P. A. Kollman. Amber 15. University of California, San Francisco, 2015.

[143] D. A. Case, R. M. Betz, D. S. Cerutti, III T. E. Cheatham, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao, and P. A. Kollman. Amber 16. University of California, San Francisco, 2016.

[144] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–25, 2006.

[145] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24(16):1999–2012, 2003.

[146] M. C. Lee and Y. Duan. Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins*, 55(3):620–34, 2004.

[147] J. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074, 2000.

[148] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.*, 97(40):10269–10280, 1993.

[149] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, 11(8):3696–713, 2015.

[150] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development

and testing of a general amber force field. *J. Comput. Chem.*, 25(9):1157–74, 2004.

[151] A. Jakalian, B. L. Bush, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.*, 21(2):132–146, 2000.

[152] A. Jakalian, D. B. Jack, and C. I. Bayly. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.*, 23(16):1623–41, 2002.

[153] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision D.01. Gaussian Inc., Wallingford CT, 2016.

[154] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79(2):926–935, 1983.

[155] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *J. Phys. Chem.*, 91(24):6269–6271, 1987.

[156] I. S. Joung and 3rd Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B*, 112(30):9020–41, 2008.

[157] I. S. Joung and 3rd Cheatham, T. E. Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B*, 113(40):13279–90, 2009.

[158] H. M. Senn and W. Thiel. QM/MM methods for biomolecular systems. *Angew. Chem. Int. Ed.*, 48(7):1198–229, 2009.

[159] D. Bakowies and W. Thiel. Hybrid Models for Combined Quantum Mechanical

and Molecular Mechanical Approaches. *J. Phys. Chem.*, 100(25):10580–10594, 1996.

[160] I. Antes and W. Thiel. On the Treatment of Link Atoms in Hybrid Methods. In *Combined Quantum Mechanical and Molecular Mechanical Methods*, ACS Symposium Series, pages 50–65. 1998.

[161] U. Ryde. The coordination of the catalytic zinc in alcohol dehydrogenase studied by combined quantum-chemical and molecular mechanics calculations. *J. Comput. Aided Mol. Des.*, 10(2):153–64, 1996.

[162] U. Eichler, C. M. Klmel, and J. Sauer. Combiningab initio techniques with analytical potential functions for structure predictions of large systems: Method and application to crystalline silica polymorphs. *J. Comput. Chem.*, 18(4):463–477, 1997.

[163] T. K. Woo, L. Cavallo, and T. Ziegler. Implementation of the IMOMM methodology for performing combined QM/MM molecular dynamics simulations and frequency calculations. *Theor. Chem. Acc.*, 100(5-6):307–313, 1998.

[164] S. Dapprich, I. Komáromi, K. S. Byun, K. Morokuma, and M. J. Frisch. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J. Mol. Struct. THEOCHEM*, 461-462:1–21, 1999.

[165] A. H. de Vries, P. Sherwood, S. J. Collins, A. M. Rigby, M. Rigutto, and G. J. Kramer. Zeolite Structure and Reactivity by Combined Quantum-Chemical-Classical Calculations. *J. Phys. Chem. B*, 103(29):6133–6141, 1999.

[166] M. Eichinger, P. Tavan, J. Hutter, and M. Parrinello. A hybrid method for solutes in complex solvents: Density functional theory combined with empirical force fields. *J. Chem. Phys.*, 110(21):10452–10467, 1999.

[167] M. J. Field, M. Albe, C. Bret, F. Proust-De Martin, and A. Thomas. The dynamo library for molecular simulations using hybrid quantum mechanical and molecular mechanical potentials. *J. Comput. Chem.*, 21(12):1088–1100, 2000.

[168] M. Swart. AddRemove: A new link model for use in QM/MM studies. *Int. J. Quantum Chem.*, 91(2):177–183, 2003.

[169] M. J. Field, P. A. Bash, and M. Karplus. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *J. Comput. Chem.*, 11(6):700–733, 1990.

[170] N. Reuter, A. Dejaegere, B. Maigret, and M. Karplus. Frontier Bonds in QM/MM

Methods: A Comparison of Different Approaches. *J. Phys. Chem. A*, 104(8):1720–1735, 2000.

[171] N. Ferré and M. Olivucci. The amide bond: pitfalls and drawbacks of the link atom scheme. *J. Mol. Struct. THEOCHEM*, 632(1-3):71–82, 2003.

[172] U. C. Singh and P. A. Kollman. A combinedab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the CH3Cl + Cl? exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.*, 7(6):718–730, 1986.

[173] B. Waszkowycz, I. H. Hillier, N. Gensmantel, and D. W. Payling. A theoretical study of hydrolysis by phospholipase A2: the catalytic role of the active site and substrate specificity. *J. Chem. Soc, Perk. T. 2*, (7), 1990.

[174] B. Waszkowycz, I. H. Hillier, N. Gensmantel, and D. W. Payling. A combined quantum mechanical/molecular mechanical model of the potential energy surface of ester hydrolysis by the enzyme phospholipase A2. *J. Chem. Soc, Perk. T. 2*, (2), 1991.

[175] B. Waszkowycz, I. H. Hillier, N. Gensmantel, and D. W. Payling. A quantum mechanical/molecular mechanical model of inhibition of the enzyme phospholipase A2. *J. Chem. Soc, Perk. T. 2*, (11), 1991.

[176] V. V. Vasilyev. Tetrahedral intermediate formation in the acylation step of acetylcholinesterases. A combined quantum chemical and molecular mechanical model. *J. Mol. Struct. THEOCHEM*, 304(2):129–141, 1994.

[177] K. P. Eurenius, D. C. Chatfield, B. R. Brooks, and M. Hodoscek. Enzyme mechanisms with hybrid quantum and molecular mechanical potentials. I. Theoretical considerations. *Int. J. Quantum Chem.*, 60(6):1189–1200, 1996.

[178] P. D. Lyne, M. Hodoscek, and M. Karplus. A Hybrid QM-MM Potential Employing Hartree-Fock or Density Functional Methods in the Quantum Region. *J. Phys. Chem. A*, 103(18):3462–3471, 1999.

[179] H. Lin and D. G. Truhlar. Redistributed charge and dipole schemes for combined quantum mechanical and molecular mechanical calculations. *J. Phys. Chem. A*, 109(17):3991–4004, 2005.

[180] P. Sherwood, A. H. de Vries, S. J. Collins, S. P. Greatbanks, N. A. Burton, M. A. Vincent, and I. H. Hillier. Computer simulation of zeolite structure and reactivity using embedded cluster methods. *Faraday Discuss.*, 106:79–92, 1997.

[181] P. Sherwood. Hybrid quantum mechanics/molecular mechanics approaches. In Johannes Grotendorst, editor, *Modern Methods and Algorithms of Quantum*

Chemistry, pages 257–277. John von Neumann Institute for Computing, Jülich, 2000.

[182] P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjøvoll, A. Fahmi, A. Schäfer, and C. Lennartz. QUASI: A general purpose implementation of the QM/MM approach and its application to problems in catalysis. *J. Mol. Struct. THEOCHEM*, 632(1-3):1–28, 2003.

[183] P. H. König, M. Hoffmann, T. Frauenheim, and Q. Cui. A critical evaluation of different QM/MM frontier treatments with SCC-DFTB as the QM method. *J. Phys. Chem. B*, 109(18):9082–95, 2005.

[184] P. Amara and M. J. Field. Evaluation of an ab initio quantum mechanical/molecular mechanical hybrid-potential link-atom method. *Theor. Chem. Acc.*, 109(1):43–52, 2003.

[185] D. Das, K. P. Eurenius, E. M. Billings, P. Sherwood, D. C. Chatfield, M. Hodoščeк, and B. R. Brooks. Optimization of quantum mechanical molecular mechanical partitioning schemes: Gaussian delocalization of molecular mechanical charges and the double link atom method. *J. Chem. Phys.*, 117(23):10534–10547, 2002.

[186] V. Théry, D. Rinaldi, J. L. Rivail, B. Maigret, and G. G. Ferenczy. Quantum mechanical computations on very large molecular systems: The local self-consistent field method. *J. Comput. Chem.*, 15(3):269–282, 1994.

[187] X. Assfeld and J. L. Rivail. Quantum chemical computations on parts of large molecules: the ab initio local self consistent field method. *Chem. Phys. Lett.*, 263(1-2):100–106, 1996.

[188] G. Monard, M. Loos, V. Théry, K. Baka, and J. L. Rivail. Hybrid classical quantum force field for modeling very large molecules. *Int. J. Quantum Chem.*, 58(2):153–159, 1996.

[189] X. Assfeld, N. Ferré, and J. L. Rivail. The Local Self-Consistent Field Principles and Applications to Combined Quantum Mechanical-Molecular Mechanical Computations on Biomacromolecular Systems. In *Combined Quantum Mechanical and Molecular Mechanical Methods*, ACS Symposium Series, pages 234–249. 1998.

[190] N. Ferré, X. Assfeld, and J. L. Rivail. Specific force field parameters determination for the hybrid ab initio QM/MM LSCF method. *J. Comput. Chem.*, 23(6):610–24, 2002.

[191] J. Gao, P. Amara, C. Alhambra, and M. J. Field. A Generalized Hybrid Orbital (GHO) Method for the Treatment of Boundary Atoms in Combined QM/MM Calculations. *J. Phys. Chem. A*, 102(24):4714–4721, 1998.

[192] P. Amara, M. J. Field, C. Alhambra, and J. Gao. The generalized hybrid orbital method for combined quantum mechanical/molecular mechanical calculations: formulation and tests of the analytical derivatives. *Theor. Chem. Acc.*, 104(5):336–343, 2000.

[193] M. Garcia-Viloca and J. Gao. Generalized hybrid orbital for the treatment of boundary atoms in combined quantum mechanical and molecular mechanical calculations using the semiempirical parameterized model 3 method. *Theor. Chem. Acc.*, 111(2-6):280–286, 2003.

[194] J. Pu, J. Gao, and D. G. Truhlar. Combining Self-Consistent-Charge Density-Functional Tight-Binding (SCC-DFTB) with Molecular Mechanics by the Generalized Hybrid Orbital (GHO) Method. *J. Phys. Chem. A*, 108(25):5454–5463, 2004.

[195] J. Pu, J. Gao, and D. G. Truhlar. Generalized Hybrid Orbital (GHO) Method for Combining Ab Initio Hartree-Fock Wave Functions with Molecular Mechanics. *J. Phys. Chem. A*, 108(4):632–650, 2004.

[196] J. Pu, J. Gao, and D. G. Truhlar. Generalized hybrid-orbital method for combining density functional theory with molecular mechanicals. *Chemphyschem*, 6(9):1853–65, 2005.

[197] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Natl. Bur. Stand.*, 49(6), 1952.

[198] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Comput. J.*, 7(2):149–154, 1964.

[199] E. Polak and G. Ribiere. Note sur la convergence de méthodes de directions conjuguées. *Rev. Fr. Inform. Rech. O.*, 3(16):35–43, 1969.

[200] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA J. Appl. Math.*, 6(1):76–90, 1970.

[201] R. Fletcher. A new approach to variable metric algorithms. *Comput. J.*, 13(3):317–322, 1970.

[202] D. Goldfarb. A family of variable-metric methods derived by variational means. *Math. Comput.*, 24(109):23–23, 1970.

[203] D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Math. Comput.*, 24(111):647–647, 1970.

[204] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45(1-3):503–528, 1989.

[205] J. Kästner, J. M. Carr, T. W. Keal, W. Thiel, A. Wander, and P. Sherwood. DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations†. *J. Phys. Chem. A*, 113(43):11856–11865, 2009.

[206] D. Frenkel and B. Smit. *Understanding molecular simulation from algorithms to applications*. Academic Press, 2 edition, 2002.

[207] M. J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team. Gromacs user manual version 2019.4. http://www.gromacs.org.

[208] R. W. Hockney and J. W. Eastwood. *Computer Simulation Using Particles*. Taylor & Francis, 1988.

[209] H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[210] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52(2):255–268, 1984.

[211] W. G. Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A*, 31(3):1695–1697, 1985.

[212] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[213] S. Nosé and M. L. Klein. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, 50(5):1055–1076, 1983.

[214] P. Langevin. Theory of Brownian motion. *C. R. Acad. Sci.*, 146:508–533, 1908.

[215] Y. Zhang, H. Liu, and W. Yang. Free energy calculation on enzyme reactions with an efficient iterative procedure to determine minimum energy paths on a combinedab initioQM/MM potential energy surface. *J. Chem. Phys.*, 112(8):3483–3492, 2000.

[216] J. Kästner, H. M. Senn, S. Thiel, N. Otte, and W. Thiel. QM/MM Free-Energy Perturbation Compared to Thermodynamic Integration and Umbrella Sampling: Application to an Enzymatic Reaction. *J. Chem. Theory Comput.*, 2(2):452–61, 2006.

[217] S. Metz, J. Kästner, A. A. Sokol, T. W. Keal, and P. Sherwood. ChemShell-a modular software package for QM/MM simulations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 4(2):101–110, 2014.

[218] L. Michaelis, M. L. Menten, K. A. Johnson, and R. S. Goody. The original Michaelis constant: translation of the 1913 Michaelis-Menten paper. *Biochemistry*, 50(39):8264–9, 2011.

[219] D. L. Nelson and M. M. Cox. *Lehninger principles of biochemistry*. W. H. Freeman, New York, NY, 7 edition, 2017.

[220] J. F. Morrison and C. T. Walsh. The Behavior and Significance of Slow-Binding Enzyme Inhibitors. In Alton Meister, editor, *Advances in Enzymology and Related Areas of Molecular Biology*, volume 61, pages 201–301. John Wiley & Sons, Inc, 1988.

[221] M. Gersch, F. Gut, V. S. Korotkov, J. Lehmann, T. Bottcher, M. Rusch, C. Hedberg, H. Waldmann, G. Klebe, and S. A. Sieber. The mechanism of caseinolytic protease (ClpP) inhibition. *Angew. Chem. Int. Ed.*, 52(10):3009–14, 2013.

[222] M. W. Hackl and S. A. Sieber. unpublished work.

[223] I. Antes. DynaDock: A new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. *Proteins*, 78(5):1084–104, 2010.

[224] A. Szyk and M. R. Maurizi. Crystal structure at 1.9A of E. coli ClpP with a peptide covalently bound at the active site. *J. Struct. Biol.*, 156(1):165–74, 2006.

[225] C. E. A. F. Schafmeister, W. S. Ross, and V. Romanovski. Leap, 1995. San Francisco, University of California.

[226] T. Darden, D. York, and L. Pedersen. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[227] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–341, 1977.

[228] W. Smith and T. R. Forester. DL_poly_2.0: A general-purpose parallel molecular dynamics simulation package. *J. Mol. Graph.*, 14(3):136–141, 1996.

[229] F. Neese. The ORCA program system. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2(1):73–78, 2011.

[230] F. Neese. Software update: the ORCA program system, version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 8(1), 2017.

[231] S. K. Schiferl and D. C. Wallace. Statistical errors in molecular dynamics averages. *J. Chem. Phys.*, 83(10):5203–5209, 1985.

[232] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[233] E. Pellegrini and M. J. Field. A Generalized-Born Solvation Model for Macromolecular Hybrid-Potential Calculations. *J. Phys. Chem. A*, 106(7):1316–1326, 2002.

[234] I. Mayer. Bond order and valence: Relations to Mulliken's population analysis. *Int. J. Quantum Chem.*, 26(1):151–154, 1984.

[235] A. J. Bridgeman, G. Cavigliasso, L. R. Ireland, and J. Rothery. The Mayer bond order as a tool in inorganic chemistry†. *J. Chem. Soc., Dalton Trans.*, (14):2095–2108, 2001.

[236] F. Frigerio, A. Coda, L. Pugliese, C. Lionetti, E. Menegatti, G. Amiconi, H. P. Schnebli, P. Ascenzi, and M. Bolognesi. Crystal and molecular structure of the bovine alpha-chymotrypsin-eglin c complex at 2.0 A resolution. *J. Mol. Biol.*, 225(1):107–23, 1992.

[237] H. Li, A. D. Robertson, and J. H. Jensen. Very fast empirical prediction and rationalization of protein pKa values. *Proteins*, 61(4):704–21, 2005.

[238] E. R. Duell, M. Glaser, C. Le Chapelain, I. Antes, M. Groll, and E. M. Huber. Sequential Inactivation of Gliotoxin by the S-Methyltransferase TmtA. *ACS Chem. Biol.*, 11(4):1082–9, 2016.

[239] D. R. Roe and T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.*, 9(7):3084–3095, 2013.

[240] M. Schneider. Aifgen1.0. Technische Universität München, Freising, 2019.

[241] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.

[242] A. E. Ehrenberg, B. Schmuck, M. I. Anwar, S. S. Gustafsson, G. Stenberg, and U. H. Danielson. Accounting for strain variations and resistance mutations in the characterization of hepatitis C NS3 protease inhibitors. *J Enzyme Inhib. Med. Chem.*, 29(6):868–76, 2014.

Acknowledgement

First of all, I would like to express my sincere appreciation to my previous supervisor, Prof. Dr. Iris Antes, for her constant support and patient guidance through the past years. I owe earnest gratitude for all her contributions of time, ideas, and funding, which allowed me to complete this dissertation. Unfortunately, she passed away in the summer of 2021. I lost my "Doktormutter" and I will miss her forever.

I would like to pay my special regards to Prof. Dr. Aphrodite Kapurniotu who was my mentor in Graduate Center Weihenstephan and kindly took over my supervision. Her kind help and great support allow me to finish my doctoral study smoothly.

My thanks and appreciations also go to Prof. Dr. Martin Zacharias for being the second examiner of my dissertation and critically reviewing my manuscripts.

I am also thankful to Prof. Dr. Dieter Langosch for being my new mentor.

I am especially grateful for the encouragement, help and support from Asst. Prof. Dr. Antoine Marion, who helped me a lot to solve some technical problems and supported me in publishing our works.

I would like to thank our collaborators, Prof. Dr. Stephan A. Sieber and Dr. Mathias Hackl, who conducted extra experiments to verify my assumptions.

Thanks to all my current and past colleagues in the group of Theoretical Chemical Biology and Protein Modelling: Okke Melse, Markus Schneider, Maximilian Meixner, Martin Zachmann, Martijn Bemelmans, Simone Göppert, Manuel Glaser, Dr. Thomas Hoffmann, Dr. Ewa Chudyk, Dr. Atanas Patronov, Dr. Harpreet Shah, Lukas Wietbrock, Dr. Helmut Lutz, Dr. Ilke Ugur-Marion, Dr. Ina Bisha and all master and bachelor students. I also would like to thank past secretaries in our group: Bianca Riedl, Iwona Federspiel, Martina Uzoma, Maria Thole and Brigitte Stable.

Finally, an honorable mention goes to my wife Xiaofei Yin and my parents, for their unconditional love and support during good and difficult times throughout my whole doctoral life.