Fall 12-2014

Development and Application of Combined Quantum Mechanical and Molecular Mechanical Methods

Rui Lai
University of Nebraska-Lincoln, rlai3@huskers.unl.edu

Development and Application of Combined Quantum Mechanical and Molecular Mechanical Methods

by

Rui Lai

A THESIS

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Master of Science

Major: Chemistry

Under the Supervision of Professor Hui Li

Lincoln, Nebraska

December 2014

Development and Application of Combined Quantum Mechanical and Molecular Mechanical Methods

Rui Lai, M.S.

University of Nebraska 2014

Advisor: Hui Li

Compromising of computational cost and accuracy, combined quantum mechanical and molecular mechanical (QM/MM) methods are practical methods for studying large molecular systems. The use of induced dipole polarizable force fields can significantly improve the accuracy of MM and QM/MM methods. However, induced dipole models tend to overestimate the polarization energy at short interaction distances. Damping functions can be applied to reduce the over polarization. MM-MM damping schemes have been developed to correct the overestimated polarization between MM atoms; QM-MM damping scheme has not been developed. In this thesis, a QM-MM damping scheme is developed for the damping of the MM dipole polarizability when the MM atoms are in short interacting distance with QM atoms. With this damping scheme, the induced dipole polarization energies in QM/MM calculation can reproduce the values from accurate QM calculations. A general protocol for applying QM/MM methods to study enzyme catalysis is established, and applied to compute the activation free energy of the hydrogen abstraction reaction of camphor catalyzed by cytochrome enzyme P450cam. The estimated activation free energy is in good agreement with the experiments and the results obtained from other QM/MM methods.

Table of Contents

List of Abbreviations

AMBER	Assisted Model Building with Energy Refinement
B3LYP	A hybrid functional used in DFT with 20% HF and 80% Becke88 exchange combined with the Lee-Yang-Parr correlation functional
CHARMM	Chemistry at Harvard Molecular Mechanics
Cpd I	The active species of heme enzymes, Compound I
DFT	Density Function Theory
FCM	Force constant matrix
GAMESS	General Atomic and Molecular Electronic Structure System
HF	Hartree-Fock
MD	Molecular Dynamics
MM	Molecular Mechanics
MP2	Second order Møller-Plesset perturbation theory method
OPLS	Optimized Potentials for Liquid Simulations
PBC	Periodic Boundary Condition
P450cam	A bacterial enzyme belongs to the group of cytochrome P450 enzymes
QM	Quantum Mechanics
QuanPol	Quantum Chemistry Polarizable Force Field program
UB3LYP	Unrestricted B3LYP
6-31G*	A valence double-zeta polarized basis set defined for the atoms H-Zn

CHAPTER 1 Introduction

1.1   General Overview

Computational chemistry is a branch of chemistry in that computers are used to solve mathematic equations that describe behavior of chemical systems[1]. In computational chemistry, theoretical models are used to represent real systems. Various approximations are involved in the solution of electronic wavefunctions and the modeling of intermolecular interactions, as well as phase space sampling. It is always necessary to develop more accurate and more efficient computational chemistry methods to solve emerging chemical problems related to life, materials and energy.

According to the postulates of quantum mechanics (QM)[2], a system can be completely described by a wavefunction that satisfies time-dependent Schrödinger equitation[3]. QM methods that solve electronic Schrödinger equations have been proved to be very accurate for atomic and molecular systems. However, due to high computational costs, QM methods are not affordable for large molecular systems[4]. On the other hand, the empirical molecular mechanical (MM) methods are very efficient for simulating large molecular systems. However, because they are based on ball-spring model, MM methods cannot be used to describe electronic properties and their changes, such as chemical reactions[1]. The combined QM and MM methods (QM/MM) proposed by Warshel and Levitt in 1976[5] can be used to study local chemical reactions in a large molecular system, for example, enzyme catalyzed biological reactions.

The use of induced dipole polarizable force fields can significantly improve the accuracy of MM and QM/MM methods. However, induced dipole models tend to overestimate the polarization energy at short interaction distances. Damping functions

can be applied to reduce the over polarization. In chapter 2, a QM-MM damping scheme is developed for the damping of the MM dipole polarizability when the MM atoms are in short interaction distance with QM atoms. With this damping scheme, the induced dipole polarization energies in QM/MM calculation can reproduce the values from accurate QM calculations. In chapter 3, a general protocol for applying QM/MM methods to study enzyme catalysis is established, and applied to compute the activation free energy of the hydrogen abstraction reaction of camphor catalyzed by the enzyme P450cam.

1.2 Theoretical Background

1.2.1 Quantum Mechanical (QM) Methods

Based on the Schrödinger equation, a molecular system can be described by a wavefunction that satisfies the time dependent Schrödinger equation is:

$$\mathbf{H}(\mathbf{r},t)\Psi(\mathbf{r},t) = i\hbar \frac{\partial \Psi(\mathbf{r},t)}{\partial t} \tag{1-1}$$

In equation (1-1), Ψ is the wavefunction and $\mathbf{H}$ is the Hamiltonian operator, which is the sum of kinetic energy operator and the potential energy operator:

$$\mathbf{H}(\mathbf{r},t) = -\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r},t) \tag{1-2}$$

where ∇^2 is the Laplace operator:

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{1-3}$$

For the cases that the potential energy operators are independent of time, the wavefunction can be divided into a spatial part and a time part:

$$\Psi(\mathbf{r},t) = \psi(\mathbf{r})T(t) \tag{1-4}$$

The spatial part satisfies the time-independent Schrödinger equation:

$$\mathbf{H}(\mathbf{r})\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{1-5}$$

The total energy of the system can be obtained from the solution of equation (1-5). The total energy of the system typically contains five parts: the kinetic energy of electrons, the kinetic energy of nuclei, the electrons and nuclei attraction energy, the repulsion energy of electrons and the repulsion energy of nuclei[6]. Therefore, the Hamiltonian operator for a system that contains p electrons and q nuclei can be written as a combination of the accordingly five parts:

$$\mathbf{H} = -\sum_{i}^{p}\frac{1}{2}\nabla^2 - \sum_{k}^{q}\frac{1}{2m_k}\nabla^2 - \sum_{k}^{q}\sum_{i}^{p}\frac{Z_k}{r_{ik}} + \sum_{i<j}^{p}\frac{1}{r_{ij}} + \sum_{k<l}^{q}\frac{Z_k Z_l}{r_{kl}} \tag{1-6}$$

In equation (1-6), i and j represent two different electrons in the system, k and l are for two different nuclei, m_k is the mass of nucleus k, Z is the charge of nucleus, and r is the distance between electrons, nuclei, or the distance between electron and nucleus accordingly.

According to the Born-Oppenheimer approximation[7], the nuclei move much slower than electrons, so the molecular wavefunction can be separated into two parts: electronic part and nuclear part. The electronic Hamiltonian can be written as:

$$\mathbf{H}_e = -\sum_{i}^{p}\frac{1}{2}\nabla^2 - \sum_{k}^{q}\sum_{i}^{p}\frac{Z_k}{r_{ik}} + \sum_{i<j}^{p}\frac{1}{r_{ij}} + \sum_{k<l}^{q}\frac{Z_k Z_l}{r_{kl}} \tag{1-7}$$

The electronic Schrödinger equation can be written as:

$$\mathbf{H}_e\Psi_e = E_e\Psi_e \tag{1-8}$$

The electronic energy including nuclear repulsion energy can be obtained by solving equation (1-8) and the nuclear kinetic energy can be obtained by a Hessian vibrational analysis.

From the above description, the electronic energy can be computed simply by solving equation (1-8). However, equation (1-8) cannot be exactly solved for systems with multiple elctrons. Many approximations have been introduced to solve the equation. One of the most fundamental and widely used approximations is the Hartree-Fock method[8, 9]. In Hartree's method, the total electronic wavefunction of a multi-electron system is approximated as the product of independent one-electron wavefunctions:

$$\Psi_e = \psi_e(1)\psi_e(2)\psi_e(3)\cdots\psi_e(n) \tag{1-9}$$

It is called a Hartree product. $\psi_e(n)$ is the spatial orbitals. There are two problems for the Hartree product. One is that electrons are indistinguishable thus the wavefunction should be antisymmetric. The other is that electrons have the property of spin. In Hartree's method, the wavefunctions are symmetric and the spins of electrons are not included in the wavefunctions. To solve these problems, Fock[9] and Slater[10] made a correction to this method by using a determinant of spin orbitals, which is the product of a spatial orbital and a spin function:

$$\chi_i(j) = \begin{cases} \psi_i(j)\alpha(j) \\ \quad or \\ \psi_i(j)\beta(j) \end{cases} \tag{1-10}$$

For n electrons and n spin orbitals, the Slater determinant is:

$$\Psi = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_N(1) \\ \chi_1(2) & \chi_1(2) & \cdots & \chi_N(2) \\ \vdots & \vdots & \vdots & \vdots \\ \chi_1(n) & \chi_1(n) & \cdots & \chi_N(n) \end{vmatrix} \tag{1-11}$$

Given the Hamiltonian $\mathbf{H}_e$ and the determinant wavefunction, the best wavefunction that minimizes the energy can be obtained by using the variation principle.

The variation treatment starts from the energy expression. For example, the energy of a closed-shell molecule can be expressed as:

$$\begin{aligned} E_0 &= \langle \Psi_0 | \mathbf{H} | \Psi_0 \rangle \\ &= \sum_i^p h_i + \sum_{i=1}^p \sum_{j=1}^{i-1} (2J_{ij} - K_{ij}) + V_{qq} \end{aligned} \tag{1-12}$$

In equation (1-12), $\mathbf{h}_i$ is one-electron integral and V_{qq} is the repulsion of nuclei:

$$\begin{aligned} h_i &= \int \chi_i^*(1)\mathbf{h}_i \chi_i(1) dv_1 \\ \mathbf{h}_i &= -\frac{1}{2}\nabla_i^2 - \sum_i^p \sum_k^q \frac{Z_k}{r_{ik}} \\ V_{qq} &= \sum_k^n \sum_{l>k}^n \frac{Z_k Z_l}{r_{kl}} \end{aligned} \tag{1-13}$$

The two-electron integral J_{ij} is the Coulomb integral, which is repulsion energy between electrons:

$$J_{ij} = \int \chi_i^*(1)\chi_i(1)\frac{1}{r_{ij}}\chi_j^*(2)\chi_j(2) dv_1 dv_2 \tag{1-14}$$

K_{ij} is the exchange integral:

$$K_{ij} = \int \chi_i^*(1)\chi_j^*(2)\frac{1}{r_{ij}}\chi_i(2)\chi_j(1) dv_1 dv_2 \tag{1-15}$$

The Hartree-Fock equations can be obtained via a differentiation procedure:

$$\mathbf{F}_i \chi_e(i) = \varepsilon_i \chi_e(i) \tag{1-16}$$

$\mathbf{F}_i$ is the Fock operator:

$$\mathbf{F}_i = \mathbf{h}_i + \sum_{j=1}^{p/2} \left[2\mathbf{J}_j - \mathbf{K}_j \right] \tag{1-17}$$

$\mathbf{J}$ is the electronic coulomb operator and $\mathbf{K}$ is the exchange operator. When they are applied to the wavefunction, we have:

$$\mathbf{J}_j \chi_i(1) = \chi_i(1) \int d\mathbf{v}_2 \left| \chi_j(2) \right|^2 \frac{1}{r_{12}} \tag{1-18}$$

$$\mathbf{K}_j \chi_i(1) = \chi_i(1) \int d\mathbf{v}_2 \chi_j(2)^* \chi_j(2) \frac{1}{r_{12}} \tag{1-19}$$

The Hartree-Fock equation (1-16) is solved iteratively by using self-consistent field (SCF) method. The initial guess is updated during the iteration until the difference between two consecutive iterates reaches a certain criteria. The spatial orbitals can be expanded by linear combinations of basis functions:

$$\psi_i = \sum_p C_{ip} \mu_p \tag{1-20}$$

where C_{ip} are the molecular orbital expansion coefficients and μ_p are the basis functions. This method is the basis set approximation introduced by Roothaan[11]. As the basis set becomes larger and larger, the Hartree-Fock energy approaches the complete basis set limit. Using the basis set, the Hartree-Fock equations can be written as:

$$F_i \sum_p C_{ip} \mu_p = \varepsilon_i \sum_p C_{ip} \mu_p \tag{1-21}$$

where ε_i is expressed as:

$$\varepsilon_i = \left\langle \psi_i \left| \mathbf{F}_i \right| \psi_i \right\rangle = \mathbf{H}_i^{core} + \sum_{i=1}^{p} \sum_{j=1}^{i-1} (J_{ij} - K_{ij}) \tag{1-22}$$

Using the basis set, equation (1-21) leads to the Roothaan-Hall euqations[12]:

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

$$F_{\alpha\beta} = \langle \mu_\alpha | \mathbf{F} | \mu_\beta \rangle \qquad (1\text{-}23)$$

$$S_{\alpha\beta} = \langle \mu_\alpha | \mu_\beta \rangle$$

$\mathbf{S}$ is the overlap matrix contains the overlap integrals of basis functions. $\mathbf{F}$ is the matrix representation of Fock operator. The element $F_{\alpha\beta}$ in the Fock matrix can be expanded:

$$
\begin{aligned}
F_{\alpha\beta} &= \langle \mu_\alpha | \mathbf{F} | \mu_\beta \rangle = \langle \mu_\alpha | \mathbf{h} | \mu_\beta \rangle + \sum_j^{occ} \langle \mu_\alpha | \mathbf{J}_j - \mathbf{K}_j | \mu_\beta \rangle \\
&= \langle \mu_\alpha | \mathbf{h} | \mu_\beta \rangle + \sum_{\gamma\delta}^{M_{basis}} P_{\gamma\delta} \left(\langle \mu_\alpha \mu_\gamma | \frac{1}{r_{12}} | \mu_\beta \mu_\delta \rangle - \langle \mu_\alpha \mu_\gamma | \frac{1}{r_{12}} | \mu_\delta \mu_\beta \rangle \right)
\end{aligned}
\qquad (1\text{-}24)
$$

In equation (1-24), $P_{\gamma\delta}$ is the density matrix of the expansion coefficients:

$$P_{\gamma\delta} = \sum_j^{occ} C_{\gamma j} C_{\delta j} \qquad (1\text{-}25)$$

The coefficients are determined in a self-consistent manner. Therefore, the HF energy in basis set can be obtained.

The Slater determinant is written as the products of spatial orbital and spin functions (α or β). In a closed-shell system, all electrons are paired and the same spatial orbital function can be used for a pair of electrons with different spins α or β and the electrons in a pair have the same energy. This method is called as the restricted Hartree-Fock method (RHF).

There are two HF methods for open-shell systems. One is known as the unrestricted Hartree-Fock method (UHF). In UHF method, the electrons in the same orbitals have the freedom to have different spatial orbitals. Due to the contribution of higher states in lower states wavefunctions, the spin contamination problem exists in UHF method. In the other approach, known as the restricted open-shell HF method

(ROHF), two sets of wavefunctions for paired and unpaired electrons are considered. So the spin contamination problem does not occur.

HF method is the basis of other methods, such as second order Møller-Plesset perturbation theory (MP2) method, coupled cluster (CC) method and density functional theory (DFT) method.

1.2.2 Molecular Mechanical (MM) Methods

In MM methods, a molecular system is treated with ball-spring model. Thus, the electronic motions in a system are ignored and the energy of a system is studied as a function of the nuclear positions only[13]. These approximations enable the applicability of MM methods to describe large molecular systems. MM methods are widely used to perform molecular dynamics (MD) simulations, Monte Carlo simulations and ligand docking calculations[4, 6]. Since force fields are used to calculate the potential energy in MM methods, they are also known as force field methods. The potential energy given in some popular force fields (e.g. AMBER[14], CHARMM[15], GROMOS[16]) can be written as the sum of several individual energy terms:

$$E_{MM} = E_{bond} + E_{ang} + E_{tors} + E_{vdw} + E_{ele} \tag{1-26}$$

In equation (1-26), E_{bond} is the bond stretching energy. E_{ang} is the energy for bending a bond angle. E_{tors} is the torsional energy for the rotation of three connected bonds. E_{vdw} and E_{ele} describe the non-bonding interactions. E_{vdw} is the van der Waals interactions and is usually modeled by the Lenard-Jones potential. E_{ele} is the electrostatic interaction.

Including electronic polarization to MM force fields is an active area of research. Methods used to include polarization effects in force fields are: induced point dipole,

Drude oscillator, fluctuating charge, polarizable continuum treatment and electronic polarization via QM or QM/MM treatment. The polarizable force field described in this thesis is based on induced point dipole model. In this model, the induced dipole polarization energy is significantly overestimated when two atoms are in short interacting distance. Damping functions are needed to deal with this problem.

1.2.3   QM/MM Methods

Although MM methods are efficient for large molecular systems, the lack of capability in describing the electronic structure prevent them from being routinely used to model chemical reactions. QM methods can describe electronic structure changes in chemical reactions, but the computational cost is extraordinary high for large molecular systems. As a solution, the combined QM and MM methods (QM/MM) have been developed[5].

A typical QM/MM molecular system is divided into two regions: A QM region and a MM region. The QM region contains the reactive site of the system and the rest of the system is the MM region. The total energy of the system calculated in an additive scheme is the sum of the energies of the QM subsystems, energies of the MM subsystems and the interaction energies between QM and MM regions:

$$E = E_{QM} + E_{MM} + E_{QM-MM} \tag{1-27}$$

Typically, the interaction energies between QM and MM regions contain electrostatic terms and Van der Waals interaction terms.

There are several ways to include the electrostatic term[17]: mechanical embedding, electrostatic embedding and polarization embedding. Polarization effect is not considered in mechanical embedding since the QM partial charges are obtained by the gas phase

calculation of QM subsystems without the MM subsystems. In electrostatic embedding, the QM part is polarized by the partial charges of the MM region that are included in QM Hamiltonian as one-electron operators. In polarization embedding, QM Hamiltonian contains the polarization effect of MM region. When induced dipole polarizable force field is used, the polarization energy will be overestimated if the interaction distances of QM and MM atoms are short. Therefore, QM-MM damping functions are needed for short QM-MM interaction distances. A QM-MM damping function is introduced in this thesis.

CHAPTER 2 QM-MM Polarization Damping Function

2.1 Introduction

Induced dipole polarizable force fields proposed by Vesely[18], Stillinger[19, 20], Barnes[21] and Warshel[22] can be used to better describe the intermolecular interaction potential as compared to non-polarizable force fields. Systematically parameterized induced dipole polarizable force fields have been made available in various methods such as AMOEBA[23-25], AMBER[26, 27], OPLS/PFF[28, 29]. It is well known that induced dipole polarization is inaccurate at short interaction distances and damping functions are required.[30] Several damping functions have been proposed to correct short-distance interactions between induced dipoles.[30-33] When induced dipole polarizable force fields are used to formulate combined quantum mechanical and molecular mechanical (QM/MM) methods[5, 34-36], the induced dipole polarization energy of the MM atoms can also be overestimated at short QM-MM distances, especially when the QM atoms have significant charges. Since there is no QM-MM dipole-dipole interaction, the existing MM-MM damping function methods[30-33] cannot be applied to QM-MM cases. Currently, no QM-MM damping function for induced dipole MM methods has been reported in the literature.

In this chapter, a QM-MM damping function is introduced for MM induced dipole polarization. The design of this two-parameter Gaussian-type damping function allows for efficient evaluation of analytical energy gradients for both QM and MM atoms. By adjusting the two parameters, this damping function can produce polarization energies that are very similar to those from QM calculations. To deal with the polarization at short distances for MM systems, we also implemented the damping

function schemes proposed by Thole[30], van Duijnen and Swart[37] and Ren and Ponder[24].

2.2   Theory

In a typical QM/MM calculation, the total Hamiltonian of the system can be written as:

$$H = H_{QM} + H_{QM/MM}^{ele} + H_{QM/MM}^{pol} + H_{QM/MM}^{vdw} + H_{MM} \qquad (2\text{-}1)$$

In this equation, H_{QM} is the time independent, nonrelativistic Hamiltonian of the QM subsystem. $H_{QM/MM}^{ele}$, $H_{QM/MM}^{pol}$ and $H_{QM/MM}^{vdw}$ are the operators for the electrostatic interaction, polarization interaction and Van der Waals interactions between QM and MM respectively. H_{MM} is the operator for all the energy terms (binding and nonbonding interactions) of the MM region. Therefore, the inclusion of induced dipole polarizable force field in QM/MM calculations will cause the change of polarization interaction between QM and MM regions.

2.2.1   Polarization Energy

In the quantum chemistry polarizable force field (QuanPol)[38] program, induced dipole polarizable force field calculations, both pure MM and QM/MM, the following linear polarization equation is used[39]:

$$\mathbf{Dp} = \mathbf{E} \qquad (2\text{-}2)$$

Here $\mathbf{p}$ is a set of induced dipoles, $\mathbf{E}$ is a set of electric fields at the polarizability points due to MM charges, QM electrons and nuclei. The electric fields due to induced dipoles are not included in $\mathbf{E}$. $\mathbf{D}$ is a matrix in which the elements act on induce dipoles to

produce electric fields. The diagonal elements of the **D** matrix is the inverses of the polarizability tensors:

$$\mathbf{D}_{ii} = (\alpha_i)^{-1} = \begin{pmatrix} \alpha_{i,xx} & \alpha_{i,xy} & \alpha_{i,xz} \\ \alpha_{i,yx} & \alpha_{i,yy} & \alpha_{i,yz} \\ \alpha_{i,zx} & \alpha_{i,zy} & \alpha_{i,zz} \end{pmatrix}^{-1}$$

(2-3)

The off-diagonal elements in the **D** matrix are:

$$\mathbf{D}_{ij} = -\mathbf{T}_{ij} \qquad (i \neq j)$$

(2-4)

where the matrix operator $\mathbf{T}_{ij}$ is a symmetric matrix:

$$\mathbf{T}_{ij} = - \begin{pmatrix} \dfrac{1}{r_{ij}^3} - \dfrac{3}{r_{ij}^5} x_{ij}x_{ij} & -\dfrac{3}{r_{ij}^5} x_{ij}y_{ij} & -\dfrac{3}{r_{ij}^5} x_{ij}z_{ij} \\ -\dfrac{3}{r_{ij}^5} y_{ij}x_{ij} & \dfrac{1}{r_{ij}^3} - \dfrac{3}{r_{ij}^5} y_{ij}y_{ij} & -\dfrac{3}{r_{ij}^5} y_{ij}z_{ij} \\ -\dfrac{3}{r_{ij}^5} z_{ij}x_{ij} & -\dfrac{3}{r_{ij}^5} z_{ij}y_{ij} & \dfrac{1}{r_{ij}^3} - \dfrac{3}{r_{ij}^5} z_{ij}z_{ij} \end{pmatrix}$$

(2-5)

where (x_i, y_i, z_i) and (x_j, y_j, z_j) are the Cartesian coordinates of points i and j polarizability points and $x_{ij} = x_i - x_j$, $y_{ij} = y_i - y_j$, $z_{ij} = z_i - z_j$.

In both pure MM and QM/MM calculation the polarization energy of the induced dipoles in the field is the reversible work required to charge the field **E** from zero to full strength[39]:

$$G_{pol} = -\frac{1}{2}\mathbf{E}^T\mathbf{D}^{-1}\mathbf{E}$$

(2-6)

2.2.2 MM-MM Damping Schemes

One way to introduce damping to the interactions between induced dipoles is to scale down the interactions between the induced dipoles in the **T** matrix:

$$\mathbf{T}_{ij} = -\begin{pmatrix} \dfrac{f_e}{r_{ij}^3} - \dfrac{3f_t}{r_{ij}^5} x_{ij}x_{ij} & -\dfrac{3f_t}{r_{ij}^5} x_{ij}y_{ij} & -\dfrac{3f_t}{r_{ij}^5} x_{ij}z_{ij} \\[3mm] -\dfrac{3f_t}{r_{ij}^5} y_{ij}x_{ij} & \dfrac{f_e}{r_{ij}^3} - \dfrac{3f_t}{r_{ij}^5} y_{ij}y_{ij} & -\dfrac{3f_t}{r_{ij}^5} y_{ij}z_{ij} \\[3mm] -\dfrac{3f_t}{r_{ij}^5} z_{ij}x_{ij} & -\dfrac{3f_t}{r_{ij}^5} z_{ij}y_{ij} & \dfrac{f_e}{r_{ij}^3} - \dfrac{3f_t}{r_{ij}^5} z_{ij}z_{ij} \end{pmatrix} \tag{2-7}$$

Here f_e and f_t are MM-MM damping functions.

The damping function of linear Thole scheme are in the following forms[30]:

$$\begin{aligned} &If(v \geq 1) \quad f_e = 1.0 \quad\quad f_t = 1.0 \\ &If(v < 1) \quad f_e = 4v^3 - 3v^4 \quad f_t = v^4 \end{aligned} \tag{2-8}$$

where:

$$\begin{aligned} v &= u/a \\ u &= r_{ij}/(\alpha_i\alpha_j)^{1/6} \end{aligned} \tag{2-9}$$

a is the damping factor which is a constant but may vary for different damping schemes and interacting points. α_i and α_j are the polarizabilities for points i and j respectively. This damping function is not smooth, so energy will not be conserved in molecular dynamics (MD) simulation. An improvement to the linear Thole scheme is exponential Thole scheme[37]:

$$\begin{aligned} f_e &= 1 - \left(\frac{a^2u^2}{2} + au + 1\right)\exp(-au) \\ f_t &= 1 - \left(\frac{1}{6}a^3u^3 + \frac{a^2u^2}{2} + au + 1\right)\exp(-au) \end{aligned} \tag{2-10}$$

The Tinker-exponential damping functions are in the following forms[24]:

$$\begin{aligned} f_e &= 1 - \exp(-au^3) \\ f_t &= 1 - (1 + au^3)\exp(-au^3) \end{aligned} \tag{2-11}$$

2.2.3 QM-MM Damping Function

The MM-MM damping scheme cannot be used as a QM-MM damping scheme because there are no induced dipoles in the QM region. Here we propose the use of a QM-MM damping scheme that scales down the polarizability of the MM atoms when they are near QM atoms. The MM polarizability point i is scaled down if its distance to a QM atom j is smaller than a cutoff distance r_d:

$$\alpha_{i,f} = f_{ij}\alpha_i \tag{2-12}$$

The damping function f_{ij} is a Gaussian function, which is based on the distance between the two atoms (r_{ij}) and the cut off distance (r_d):

$$f_{ij} = \exp\left[-d \times \left(r_{ij} - r_d\right)^2\right] \tag{2-13}$$

d is a constant, which should be specific to the types of atoms i and j. If the MM atom i is close to n QM atoms, the polarizability i is scaled by all of the n damping functions:

$$\alpha_{i,f} = \alpha_i \prod_{j=1}^{n} f_{ij} \tag{2-14}$$

where f_{ij} is the damping function for point i and QM atom j.

2.2.4 Gradients of Polarization Energy

Based on equation (2-6), the derivative of the polarization energy with respect to a coordinate x can be written as:

$$\frac{\partial G_{pol}}{\partial x} = -\frac{1}{2}\left(\frac{\partial \mathbf{E}}{\partial x}\right)^T \mathbf{D}^{-1}\mathbf{E} - \frac{1}{2}\mathbf{E}^T\left(\frac{\partial \mathbf{D}^{-1}}{\partial x}\right)\mathbf{E} - \frac{1}{2}\mathbf{E}^T\mathbf{D}^{-1}\left(\frac{\partial \mathbf{E}}{\partial x}\right) \tag{2-15}$$

Since $\mathbf{p} = \mathbf{D}^{-1}\mathbf{E}$, so equation (2-15) can be written as:

$$\frac{\partial G_{pol}}{\partial x} = -\frac{1}{2}\left(\frac{\partial \mathbf{E}}{\partial x}\right)^T \mathbf{p} - \frac{1}{2}\mathbf{E}^T\left(\frac{\partial \mathbf{D}^{-1}}{\partial x}\right)\mathbf{E} - \frac{1}{2}\mathbf{E}^T\mathbf{D}^{-1}\left(\frac{\partial \mathbf{E}}{\partial x}\right) \tag{2-16}$$

Since

$$-\frac{1}{2}\mathbf{E}^T\mathbf{D}^{-1}\left(\frac{\partial \mathbf{E}}{\partial x}\right) = -\frac{1}{2}\left(\frac{\partial \mathbf{E}}{\partial x}\right)^T\left(\mathbf{D}^{-1}\right)^T\mathbf{E} \tag{2-17}$$

if we define $\tilde{\mathbf{p}} = \left(\mathbf{D}^{-1}\right)^T\mathbf{E}$, equation (2-16) becomes:

$$\frac{\partial G_{pol}}{\partial x} = -\left(\frac{\partial \mathbf{E}}{\partial x}\right)^T\left(\frac{\mathbf{p}+\tilde{\mathbf{p}}}{2}\right) - \frac{1}{2}\mathbf{E}^T\left(\frac{\partial \mathbf{D}^{-1}}{\partial x}\right)\mathbf{E} \tag{2-18}$$

For symmetric polarizability tensors, $\tilde{\mathbf{p}}$ is equal to $\mathbf{p}$. In order to evaluate the derivative

of the inverse of $\mathbf{D}$, the relation shown below is used.

$$\frac{\partial \mathbf{D}^{-1}}{\partial x} = -\mathbf{D}^{-1}\left(\frac{\partial \mathbf{D}}{\partial x}\right)\mathbf{D}^{-1} \tag{2-19}$$

So for symmetric polarizability tensors, we have:

$$-\frac{1}{2}\mathbf{E}^T\left(\frac{\partial \mathbf{D}^{-1}}{\partial x}\right)\mathbf{E} = \frac{1}{2}\mathbf{E}^T\mathbf{D}^{-1}\left(\frac{\partial \mathbf{D}}{\partial x}\right)\mathbf{D}^{-1}\mathbf{E} = \frac{1}{2}\mathbf{p}^T\left(\frac{\partial \mathbf{D}}{\partial x}\right)\mathbf{p} \tag{2-20}$$

And the gradient of polarization energy becomes:

$$\frac{\partial G_{pol}}{\partial x} = -\left(\frac{\partial \mathbf{E}}{\partial x}\right)^T\mathbf{p} + \frac{1}{2}\mathbf{p}^T\left(\frac{\partial \mathbf{D}}{\partial x}\right)\mathbf{p} \tag{2-21}$$

In the following part, we derive the explicit matrix element expression of the

gradient for a particular coordinate x_i. If run over all of the N induced dipoles in the

system, equation (2-21) should be expressed as:

$$\frac{\partial G_{pol}}{\partial x_i} = -\sum_n^N \left(\frac{\partial \mathbf{E}_n}{\partial x_i}\right)^T \mathbf{p}_n + \frac{1}{2}\sum_n^N \sum_m^N \mathbf{p}_n^T \left(\frac{\partial D_{nm}}{\partial x_i}\right)\mathbf{p}_m$$

$$= -\sum_{n=i}\left(\frac{\partial \mathbf{E}_n}{\partial x_i}\right)^T \mathbf{p}_n - \sum_{n\neq i}\left(\frac{\partial \mathbf{E}_n}{\partial x_i}\right)^T \mathbf{p}_n + \frac{1}{2}\sum_{n=i}\sum_{m=i}\mathbf{p}_n^T \left(\frac{\partial D_{nm}}{\partial x_i}\right)\mathbf{p}_m$$

$$+\frac{1}{2}\sum_{n=i}^N\sum_{m\neq i}^N\mathbf{p}_n^T \left(\frac{\partial D_{nm}}{\partial x_i}\right)\mathbf{p}_m + \frac{1}{2}\sum_{n\neq i}^N\sum_{m=i}^N\mathbf{p}_n^T \left(\frac{\partial D_{nm}}{\partial x_i}\right)\mathbf{p}_m + \frac{1}{2}\sum_{n\neq i}^N\sum_{m\neq i}^N\mathbf{p}_n^T \left(\frac{\partial D_{nm}}{\partial x_i}\right)\mathbf{p}_m \qquad (2\text{-}22)$$

$$= -\left(\frac{\partial \mathbf{E}_i}{\partial x_i}\right)^T \mathbf{p}_i - \sum_{n\neq i}^N\left(\frac{\partial \mathbf{E}_n}{\partial x_i}\right)^T \mathbf{p}_n + \frac{1}{2}\mathbf{p}_i^T\left(\frac{\partial D_{ii}}{\partial x_i}\right)\mathbf{p}_i$$

$$+\frac{1}{2}\sum_{m\neq i}^N\mathbf{p}_i^T\left(\frac{\partial D_{im}}{\partial x_i}\right)\mathbf{p}_m + \frac{1}{2}\sum_{n\neq i}^N\mathbf{p}_n^T\left(\frac{\partial D_{ni}}{\partial x_i}\right)\mathbf{p}_i + \frac{1}{2}\sum_{n\neq i}^N\sum_{m\neq i}^N\mathbf{p}_n^T\left(\frac{\partial D_{nm}}{\partial x_i}\right)\mathbf{p}_m$$

where m and n are different induced dipoles in the system.

When point i moves, only the electric field at point i is changed. The electric fields at all the other points are not subject to change. Consequently, only $\partial \mathbf{E}_i / \partial x_i$ is non-zero term. Because the change of x_i only results in the change of elements in line i and column i of D matrix, only $\partial \mathbf{D}_{im} / \partial x_i$ and $\partial \mathbf{D}_{ni} / \partial x_i$ are non-zero terms. So the gradient can be written as:

$$\frac{\partial G_{pol}}{\partial x_i} = -\left(\frac{\partial \mathbf{E}_i}{\partial x_i}\right)^T \mathbf{p}_i + \frac{1}{2}\mathbf{p}_i^T\left(\frac{\partial \mathbf{D}_{ii}}{\partial x_i}\right)\mathbf{p}_i + \frac{1}{2}\sum_{m\neq i}^N\mathbf{p}_i^T\left(\frac{\partial \mathbf{D}_{im}}{\partial x_i}\right)\mathbf{p}_m + \frac{1}{2}\sum_{n\neq i}^N\mathbf{p}_n^T\left(\frac{\partial \mathbf{D}_{ni}}{\partial x_i}\right)\mathbf{p}_i \quad (2\text{-}23)$$

The third and fourth terms in equation (2-23) are the same, so

$$\frac{\partial G_{pol}}{\partial x_i} = -\left(\frac{\partial \mathbf{E}_i}{\partial x_i}\right)^T \mathbf{p}_i + \frac{1}{2}\mathbf{p}_i^T\left(\frac{\partial \mathbf{D}_{ii}}{\partial x_i}\right)\mathbf{p}_i + \sum_{j\neq i}^N\mathbf{p}_i^T\left(\frac{\partial \mathbf{D}_{ij}}{\partial x_i}\right)\mathbf{p}_j \qquad (2\text{-}24)$$

Then,

$$\frac{\partial D_{ij}}{\partial x_i} = -\frac{\partial \mathbf{T}_{ij}}{\partial x_i} = \frac{1}{r_{ij}^3}\frac{\partial f_e}{\partial x_i}\mathbf{I} - \left(\frac{3}{r_{ij}^5}\frac{\partial f_t}{\partial x_i} - \frac{15 f_t}{r_{ij}^7}\right)\mathbf{F}_{ij} - \frac{3}{r_{ij}^5}\mathbf{C}_{ij} \qquad (2\text{-}25)$$

where $\mathbf{I}$ is the unit matrix, $\mathbf{F}_{ij}$ and $\mathbf{C}_{ij}$ are in the following forms:

$$\mathbf{F}_{ij} = \begin{bmatrix} x_{ij}^2 & x_{ij}y_{ij} & x_{ij}z_{ij} \\ y_{ij}x_{ij} & y_{ij}^2 & y_{ij}z_{ij} \\ z_{ij}x_{ij} & z_{ij}y_{ij} & z_{ij}^2 \end{bmatrix} \tag{2-26}$$

$$\mathbf{C}_{ij} = \begin{bmatrix} (f_e + 2f_t)x_{ij} & f_t y_{ij} & f_t z_{ij} \\ f_t x_{ij} & (f_e + 2f_t)y_{ij} & f_t z_{ij} \\ f_t x_{ij} & f_t y_{ij} & (f_e + 2f_t)z_{ij} \end{bmatrix} \tag{2-27}$$

Since f_e and f_t are damping functions that depend on the distance of two interacting points as described in equations (2-8), (2-10) and (2-11), their derivatives can be computed straightforwardly.

When QM-MM damping function is used, the derivative of $\mathbf{D}_{ii}$ is not zero. As described in equation (2-14), if a polarizability point interacts with n QM atoms and the distance between two of these points are smaller than the cutoff distance r_d, the polarizability will be determined by all the atoms interacting with this point. In this way, the gradient of matrix $\mathbf{D}_{ii}$ should be expressed as:

$$\frac{\partial \mathbf{D}_{ii}}{\partial x_i} = \frac{\partial (\alpha_{i,f})^{-1}}{\partial x_i} = \frac{\partial \left(\alpha_i \prod_{j=1}^{n} f_{ij}\right)^{-1}}{\partial x_i} = -\left(\alpha_i \prod_{j=1}^{n} f_{ij}\right)^{-2} \frac{\partial \left(\alpha_i \prod_{j=1}^{n} f_{ij}\right)}{\partial x_i}$$
$$= -(\alpha_{i,f})^{-2} \alpha_i \sum_{j=1}^{n} \left(\frac{\partial f_{ij}}{\partial x_i} \prod_{k \neq j}^{n} f_{ik}\right) \tag{2-28}$$

Or simply written as:

$$\frac{\partial \mathbf{D}_{ii}}{\partial x_i} = -(\alpha_{i,f})^{-1} \sum_{j=1}^{n} \left(\frac{1}{f_{ij}} \frac{\partial f_{ij}}{\partial x_i}\right) \tag{2-29}$$

According to equation (2-13), the derivatives of the QM-MM damping function can be written as:

$$\frac{\partial f_{ij}}{\partial x_i} = -2d_{ij} \times \left(r_{ij} - r_{d,ij}\right) \exp\left[-d_{ij} \times \left(r_{ij} - r_{d,ij}\right)^2\right] \frac{\partial r_{ij}}{\partial x_i} \tag{2-30}$$

or written as:

$$\frac{1}{f_{ij}} \frac{\partial f_{ij}}{\partial x_i} = -2d_{ij} \times \left(r_{ij} - r_{d,ij}\right) \frac{\partial r_{ij}}{\partial x_i} \tag{2-31}$$

When QM-MM damping function is used, the derivative of $\mathbf{D}_{ii}$ with respect to the QM coordinate x_j can be similarly derived:

$$\frac{\partial \mathbf{D}_{ii}}{\partial x_j} = -\left(\alpha_{i,f}\right)^{-1} \sum_{j=1}^{n} \left(\frac{1}{f_{ij}} \frac{\partial f_{ij}}{\partial x_j}\right) \tag{2-32}$$

$$\frac{1}{f_{ij}} \frac{\partial f_{ij}}{\partial x_j} = -2d_{ij} \times \left(r_{ij} - r_{d,ij}\right) \frac{\partial r_{ij}}{\partial x_j} \tag{2-33}$$

In this way, the gradient of the polarization energy with damping functions can be easily computed based on equation (2-24), equation (2-25) and equation (2-29).

2.3   Implementation and Computational Methods

The method presented above has been implemented in the quantum chemistry polarizable force field (QuanPol)[38] program, which is integrated in the general atomic and molecular electronic system (GAMESS) package[40, 41]. Numerical tests show that the analytic gradients are accurate to 10^{-6} hartree/bohr. All calculations were performed using the QuanPol program and the GAMESS package.

A QM/MM system is used to demonstrate the energy conservation. In this system, the QM region is a water molecule, which is treated with MP2/6-31++G(d,p)[42] level of theory. For the MM region, there are 512 water molecules. This region is obtained by an equilibrium MD stimulation that is performed in NPT ensembles (the pressure and the

temperature are scaled, the energy may fluctuate). The MD step size is chosen to be 10^{-15} s. The three-site induced-dipole water model POL3[43, 44] polarizable water model is used in the simulations. The parameters for POL3 water model are shown in table 2-1 (LT: Linear Thole; ET: Exponential Thole; TE: Tinker-exponential). RATTLE[45] is used to fix the internal geometry of the water molecules. Thereafter, one of the water molecules is chosen as the QM region and then QM/MM simulations with the run type as MD are performed in NVE ensembles (the volume and total energy of the system is constant, the pressure and temperature may fluctuate). When the MM damping functions are included, the polarization interactions of 1-2 and 1-3 atom pairs are considered in the simulation. In order to minimize the MD integration errors in the demonstration of energy conservation, the MD step size is chosen to be 2.5×10^{-16} s. The size of the periodical boundary condition (PBC) box is different for each damping schemes: Linear Thole 24.87 Å, Exponential Thole 24.89 Å and Tinker-exponential 24.61 Å. In order to make a comparison, a QM/MM simulation only with the QM-MM Gaussian damping function is performed and a pure MM simulation without any damping functions is conducted. During the QM/MM simulation, the damping constant d is chosen to be 0.0863 bohr^{-2}. The cutoff distances in the QM-MM damping function are chosen to be 3.0 Å.

To find out the best damping constant d and cutoff distance r_d for the Gaussian damping function in QM/MM polarization model, a QM/MM study on M-H$_2$O and M-Cl$^-$ polarization interaction is used (where M is Na$^+$, K$^+$, Mg^{2+} or Ca^{2+}). The cations (Na$^+$, Mg^{2+} and Ca^{2+}) are considered as the QM subsystems and treated with MP2/aug-cc-pVTZ[46, 47] method. For K$^+$ QM subsystems, the K$^+$ is treated with modified Wachters's triple zeta valence (TZV) basis set[48, 49]. No frozen cores are considered for MP2

calculation. The water molecule is treated as QP302 flexible polarizable water model[38] except the polarizability. The polarizability of oxygen atom is chosen to be the experimental value (1.44 $Å^3$) instead of the assigned polarizability (0.8 $Å^3$) in QP302 since it has been observed to be suitable to demonstrate the polarization energy at longer distances[50, 51]. Partial charges of - 0.7160 e and +0.3580 e are placed on O and H atoms for water molecule. The Van der Waals interactions between QM and MM subsystems are modeled with Lennard-Jones (L-J) interactions. These parameters are shown in Table 2-2. For Cl^-, a polarizability of 5.5 $Å^3$ is placed on it[47, 52, 53].

The polarization energies obtained by the localized molecular orbital energy decomposition analysis (LMO-EDA)[54] that treats the two subsystems as two interacting QM monomers are used to compare with the polarization energies obtained by QM/MM simulations. For LMO-EDA calculations, MP2/aug-cc-pVTZ is used for Na^+, Mg^{2+} and Ca^{2+} related systems and MP2/aug-pc4 (polarization consistent basis sets)[55-58] is employed for K^+ and Cl^- system. The basis set super position error corrected values of LMO-EDA are used for comparison. In the LMO-EDA method, the polarization energy is defined as the orbital relaxation energy on going from monomer spin orbitals to the super molecular spin orbitals.

The distances of two interacting points chosen to demonstrate the polarization energies are based on the optimized distances obtained by using MP2/ aug-cc-pVTZ for all the systems except K^+-Cl^- system. For K^+-Cl^-, due to the limitation of basis set, we use MP2/TZV for K^+ and MP2/ aug-cc-pVTZ for Cl^-. The optimized distances between two interacting points are shown in Table 2-3.

Table 2-1. Parameters for POL3 and QP302 polarizable water model

	POL3	QP302
Bond length d_{O-H} (Å)	1.0000	1.0300
Bond angle θ_{H-O-H} (°)	109.47	109.47
Lennard-Jones ε (kcal/mol)	0.1560	0.1520
Lennard-Jones $R_{min}/2$ (Å)	1.7980	1.8142
Charge q_O (e)	-0.7300	-0.7160
Charge q_H (e)	0.3650	0.3580
Dipole polarizabilities α_O (Å^3)	0.5280	1.4400
Dipole polarizabilities α_H (Å^3)	0.1700	N/A
Damping factor a (LT)	2.4410	N/A
Damping factor a (ET)	1.3305	N/A
Damping factor a (TE)	0.4246	N/A

Table 2-2. Force field parameters for MM subsystems in the QM/MM simulation

	Na$^+$	Mg^{2+}	K$^+$	Ca^{2+}	Cl$^-$
L-J ε (kcal/mol)	0.002770	0.875044	0.000328	0.120000	0.100000
L-J $R_{min}/2$ (Å)	1.868000	0.922928	2.658000	1.367000	2.470000
Charge q (e)	1.000000	2.000000	2.000000	2.000000	-1.00000
Polarizabilities α (Å^3)	N/A	N/A	N/A	N/A	5.5

2.4 Results and Discussions

2.4.1 MD Energy Conservation

After 1 million steps of the NPT equilibration, the QM/MM simulations are performed within the NVE ensembles. Thereafter, the total energies in the QM/MM simulations within NVE ensembles are extracted for each 1000 steps. Clearly, energy is

conserved very well in these QM/MM simulations with the QM-MM Gaussian damping function and the Exponential-Thole or Tinker-exponential damping schemes for the MM polarization damping. Due to the MD integration errors, there is an energy drift in 100000 MD steps for about 0.02 kcal/mol, which is also observed in the pure MM simulation. The total energy of all the modeled systems were conserved in 100000 time steps, with a standard deviation of 0.03 kcal/mol for Exponential-Thole damping scheme (Figure 2-2) and 0.04 kcal/mol for Tinker-exponential damping scheme (Figure 2-3). For Linear-Thole scheme, the damping function is not smooth so the energy is not conserved (Figure 2-4). The average temperatures are also consistent with a standard deviation of 10 K for Exponential-Thole and Tinker-exponential damping schemes.
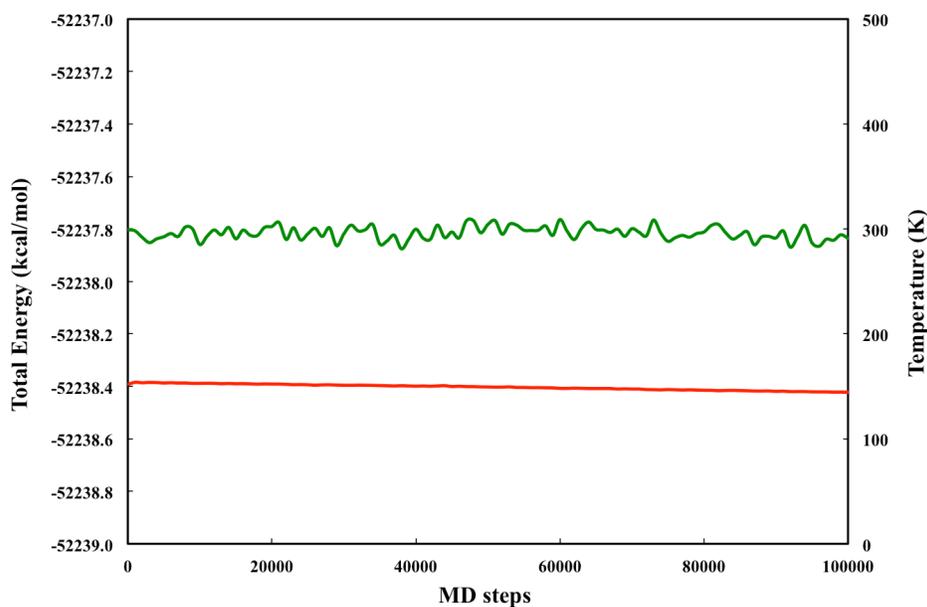
Figure 2-1. Energy (red) and temperature (green) in QM/MM simulation with the QM-MM damping function and without MM damping function
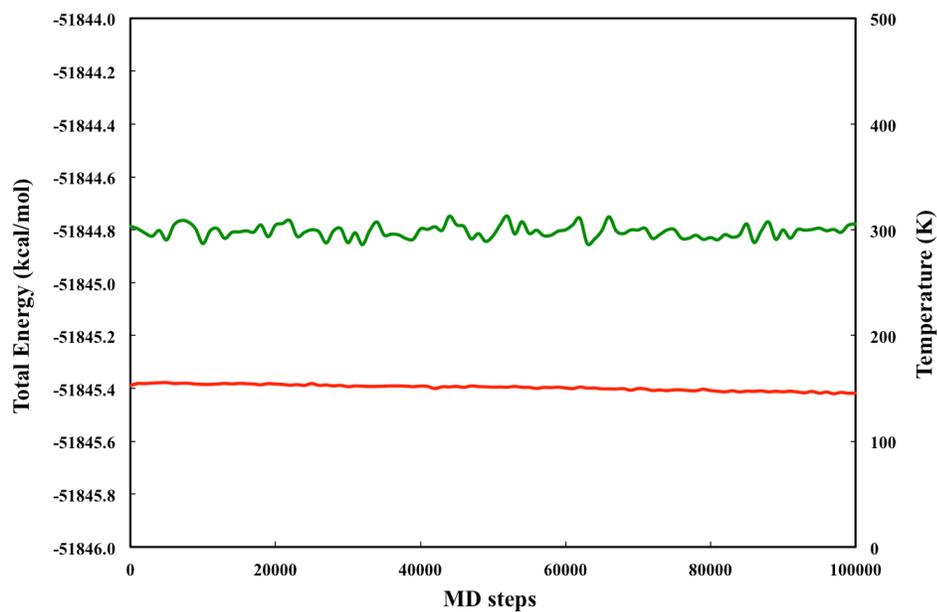
Figure 2-2. Energy (red) and temperature (green) in QM/MM simulation with the QM-MM damping function and the Exponential-Thole MM damping function
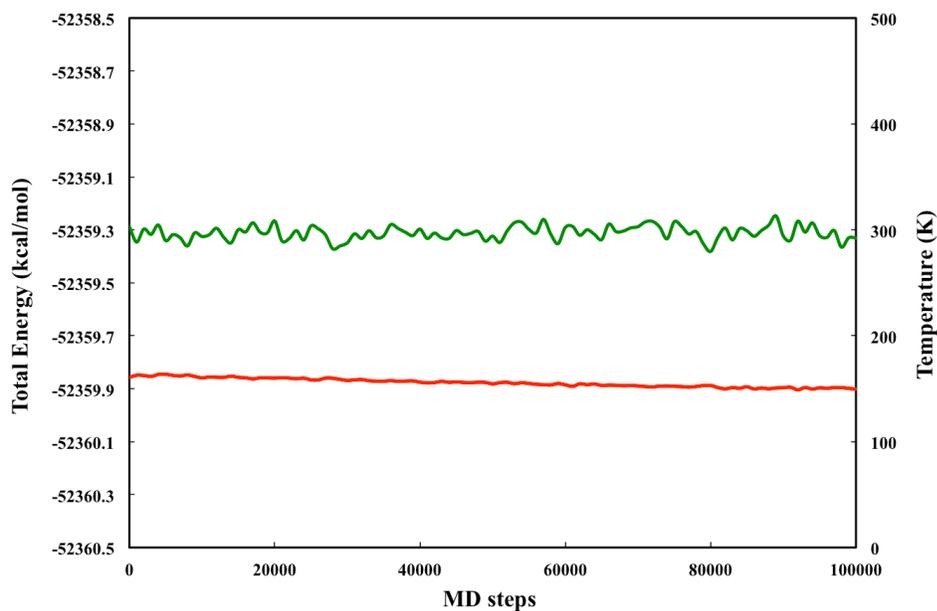
Figure 2-3. Energy (red) and temperature (green) in QM/MM simulation with the QM-MM damping function and the Tinker-Exponential MM damping function
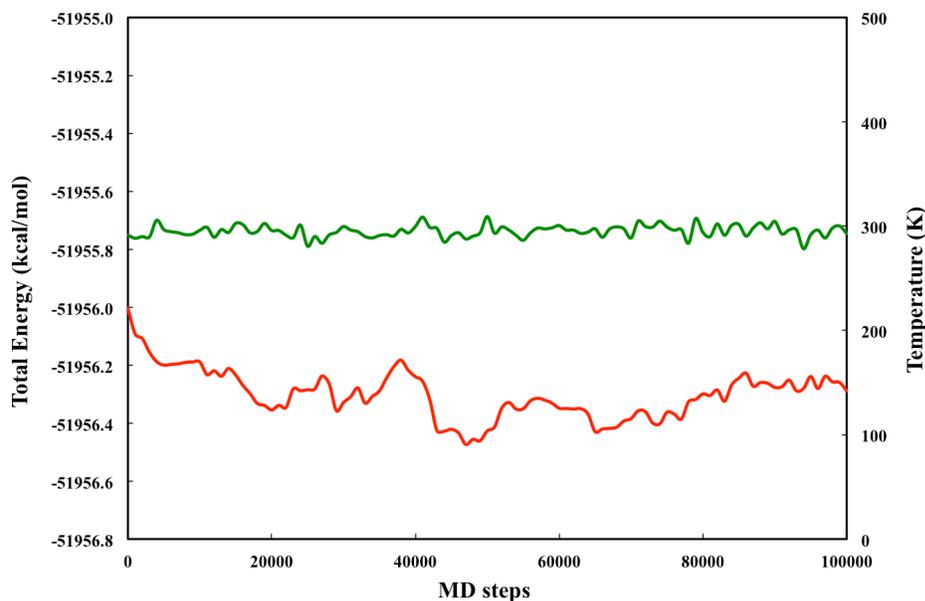
Figure 2-4. Energy (red) and temperature (green) in QM/MM simulation with the QM-MM Gaussian damping function and Linear-Thole MM damping function

2.4.2 Application of QM-MM Damping Function

Since the QM-MM damping function is sensitive to the damping constant d and the cutoff distance r_d, it is important to estimate d and r_d in this damping function. The optimized distances between the cations and the oxygen atom in water molecule or Cl$^-$ are obtained by using MP2/aug-cc-pVTZ[46, 47] method (for K$^+$, it is treated with TZV basis set). The obtained optimized distances are shown in Table 2-3. Based on the optimized

Table 2-3. Polarization energy of optimized structures

	r (Å)	d (bohr^{-2})	r_d (Å)	$E^{pol}_{LMO-EDA}$	E^{pol}_{damp}	$E^{pol}_{no-damp}$
Na$^+$-H$_2$O	2.27	0.0863	3.45	-5.99	-5.97	-9.17
Mg^{2+}-H$_2$O	1.93	0.140	3.00	-38.58	-39.65	-70.99
Ca^{2+}-H$_2$O	2.45	0.090	3.30	-19.55	-21.31	-26.64
Na$^+$-Cl$^-$	2.41	0.090	4.00	-12.52	-12.67	-29.15
K$^+$-Cl$^-$	2.96	0.050	4.00	-9.93	-10.10	-12.29

distances, the LMO-EDA calculations and QM/MM calculations are set in a range of most possible distances between the two interacting points. By changing d and the cutoff distances r_d simultaneously, we have obtained the best fitting polarization energy (shown in Table 2-3, Figure 2-5 to Figure 2-9) for these models. By using suitable d and r_d for the QM-MM damping function, the polarization energy is comparable to what calculated by using LMO-EDA method.

Na^+-H_2O. The optimized distance for Na^+ and O is 2.27 Å. In LMO-EDA calculation, if the distance between Na^+ and O is 2.27 Å, the polarization energy is given as -5.99 kcal/mol. By changing d and r_d in QM-MM damping function, we find that for r_d =3.45 Å and d=0.0863 bohr^{-2}, the polarization energy of the system is -5.97 kcal/mol, which is in concert with the LMO-EDA calculation. Meanwhile, if the damping function
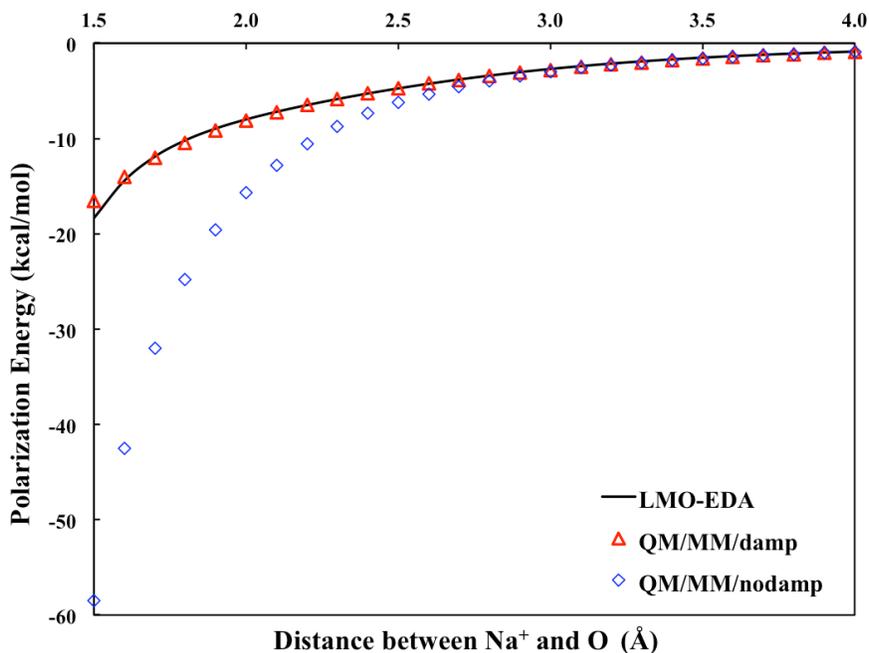
Figure 2-5. Polarization energy of O in water molecule calculated using LMO-EDA and QM/MM (with and without QM-MM damping function) for Na^+-H_2O pair.

is not included in the QM/MM calculation, the polarization energy is -9.17 kcal/mol, which is nearly twice of what calculated by LMO-EDA method. So with the damping function, the QM/MM polarization energy is more reliable. By changing the distance between Na^+ and O, a series of polarization energies are obtained by using LMO-EDA method and QM/MM with r_d =3.45 Å and d=0.0863 bohr^{-2}. The results show that the polarization energies calculated by QM/MM are in good agreement with the LMO-EDA results (Figure 2-5). For cases that Na^+ and O are in short interaction distances, Figure 2-5 shows that the QM/MM polarization energy calculations without the damping function are not trustable.

Mg^{2+}-H_2O. The optimized distance between Mg^{2+} and O is 1.93 Å. By using the optimized geometry, the polarization energy of optimized geometry produced by LMO-EDA calculation is -38.58 kcal/mol. A series of QM/MM simulation show that the best
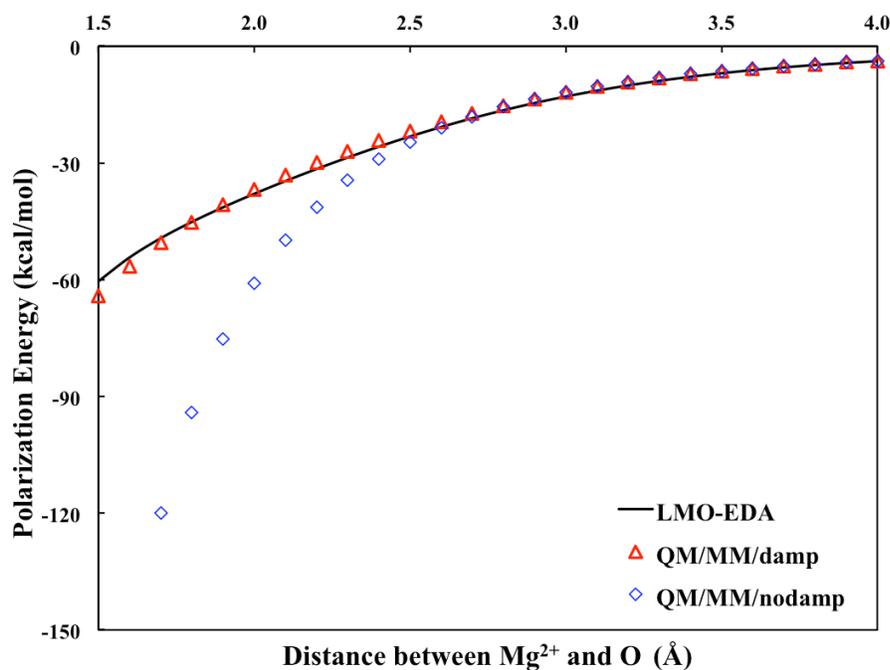
Figure 2-6. Polarization energy of O in water molecule calculated using LMO-EDA and QM/MM (with and without QM-MM damping function) for Mg^{2+}-H_2O pair.

fitting constants for this system are d=0.140 bohr^{-2} and r_d=3.00 Å. The fitting graph is shown in Figure 2-6. When the damping function is not used in the QM/MM calculation, the polarization energy is much larger than those calculated by LMO-EDA method at short distances. After using the QM-MM damping function, the polarization energy fits very well with those calculated by using LMO-EDA method (Figure 2-6). The polarization energy of the optimized system is obtained as -39.65 kcal/mol when the QM-MM damping function is used. This value has a deviation of 1.07 kcal/mol compared to the LMO-EDA one. However, if the QM-MM damping function is not used in QM/MM calculation, the polarization energy is calculated as -70.99 kcal/mol, nearly the twice of what estimated by LMO-EDA method.

Ca^{2+}-H$_2$O. The optimized distance between Ca^{2+} and O is 2.45 Å. An LMO-EDA calculation shows that the polarization energy of the optimized structure is -19.55
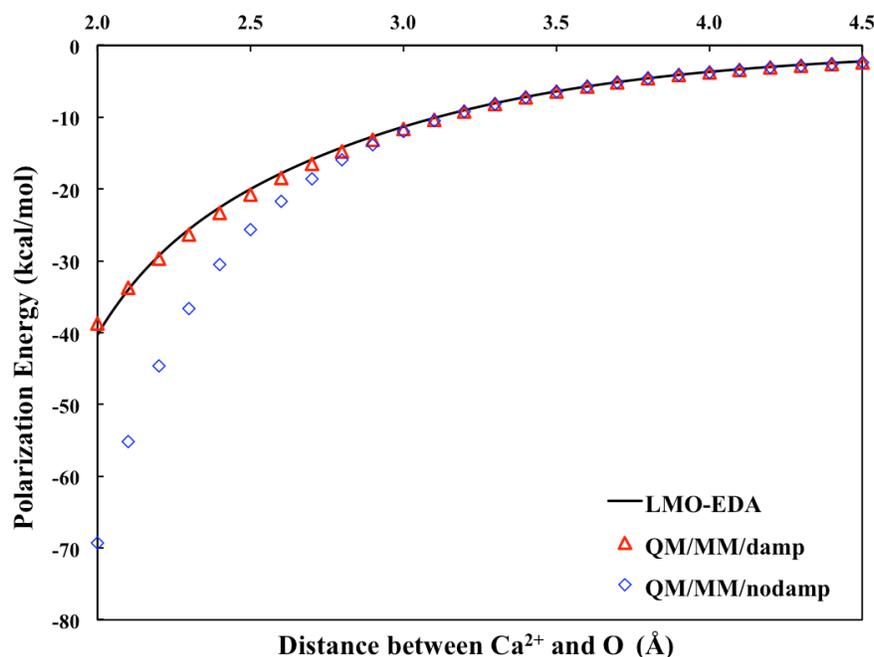
Figure 2-7. Polarization energy of O in water molecule calculated using LMO-EDA and QM/MM (with and without QM-MM damping function) for Ca^{2+}-H$_2$O pair.

kcal/mol. A series of QM/MM calculations with the new damping function give the best pair of constants as d=0.090 bohr^{-2} and r_d =3.30 Å. By using this pair of constants, polarization energies are obtained for different distances between Ca^{2+} and O (Figure 2-7). For the optimized geometry, the polarization energy is calculated as -21.31 kcal/mol with 1.76 kcal/mol difference compared to LMO-EDA result. If the damping function is not used, the QM/MM calculation gives the polarization energy at the optimized distance as -26.64 kcal/mol (a 7.09 kcal/mol difference with the LMO-EDA result). Moreover, Figure 2-7 shows that if the new damping function is not used, the QM/MM polarization energy is quite unreliable at short interaction distances.

Na$^+$-Cl$^-$. The optimized distance between Na$^+$ and Cl$^-$ is 2.41 Å. The LMO-EDA calculation produces the polarization energy for the optimized structure as -12.52
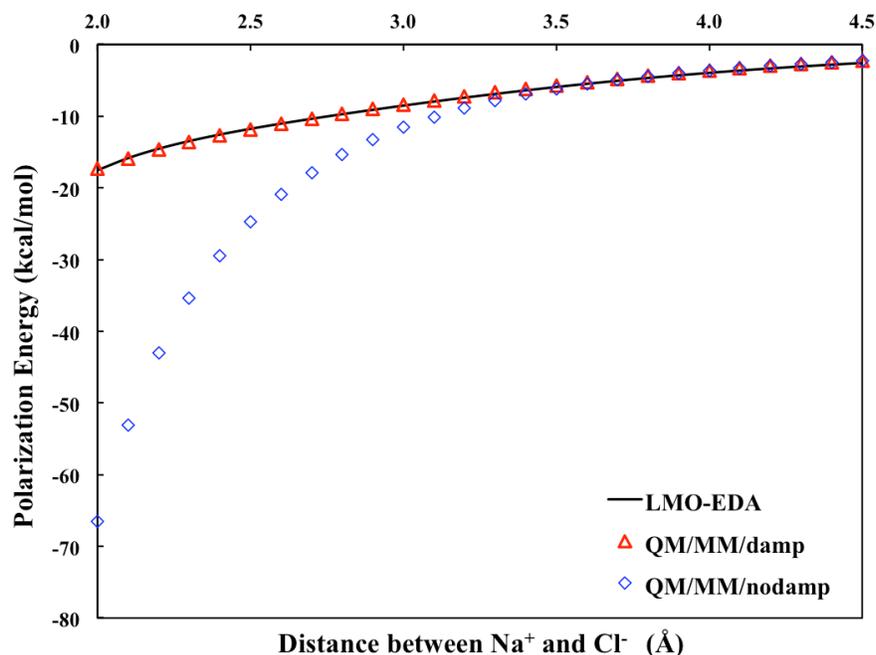
Figure 2-8. Polarization energy of Cl calculated using LMO-EDA and QM/MM (with and without QM-MM damping function) for Na$^+$-Cl$^-$ pair.

kcal/mol. If the QM-MM damping function is not used in the QM/MM calculation, the polarization energy given by QM/MM calculation is -29.15 kcal/mol, which is more than twice as great as the LMO-EDA result. With the constants d=0.090 bohr^{-2} and r_d=4.00 Å for the QM-MM damping function, the polarization energy is corrected to be -12.67 kcal/mol at the equilibrium distance. By changing the distances between Na$^+$ and Cl$^-$, the QM/MM polarization energies of Na$^+$-Cl$^-$ systems are obtained (Figure 2-8). The results show that the using of QM-MM damping function corrected the overestimation of the polarization energy of Cl$^-$.

K$^+$-Cl$^-$. The optimized distance between K$^+$ and Cl$^-$ is 2.96 Å. The changes of distances between K$^+$ and Cl$^-$ produce the polarization energies for systems in QM/MM calculations and LMO-EDA calculations. When the QM-MM damping function constants
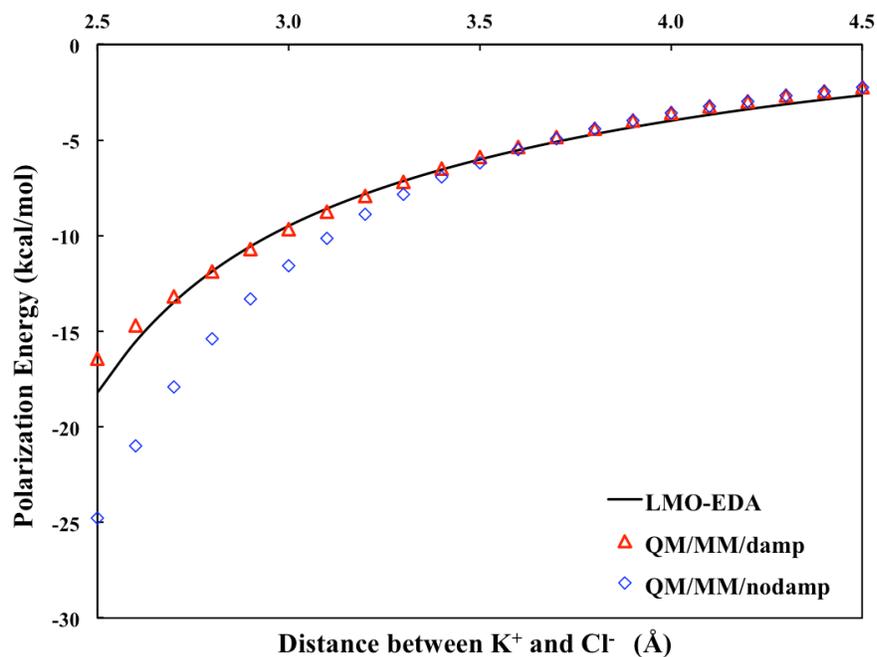
Figure 2-9. Polarization energy of Cl calculated using LMO-EDA and QM/MM (with and without QM-MM damping function) for K$^+$-Cl$^-$ pair.

d=0.050 bohr^{-2} and r_d=4.00 Å are used, the polarization energies approach to those calculated by using LMO-EDA method (Figure 2-9). The use of the new damping function corrects the errors in the polarization energies at short distances for QM/MM calculation (Figure 2-9). With the use of damping functions, the polarization energy of the optimized geometry calculated by QM/MM is damped from -12.29 kcal/mol to -10.10 kcal/mol, which is closer to that (-9.93 kcal/mol) calculated by LMO-EDA method.

Multi-atoms system. In order to know how the QM-MM damping function works for multiple QM atoms interacting with MM atoms, a series of polarization energies are obtained by QM/MM calculations on Na$^+$-Cl$^-$ systems with two Na$^+$ ion in QM regions by using the previously parameterized constants d=0.090 bohr^{-2} and r_d=4.00 Å. The model is shown in Figure 2-10. In this model, the two Na$^+$ ions and the Cl$^-$ ion form an isosceles triangle with the angle as 90˚. With the change of distances between Na$^+$ and Cl$^-$ (the distances between different Na$^+$ and Cl$^-$ are same), a series of polarization energies are obtained by QM/MM and LMO-EDA method. The results are shown in Figure 2-11. Clearly, the polarization energies are overestimated at short interacting distances if the QM-MM damping function is not used in QM/MM calculations. As we could see from equation (2-14), the damping function is repeatedly working on the polarizabilities of a polarizable atom. From Figure 2-11, with the using of QM-MM damping function, the polarization energy is over-corrected at short interaction distances. However, with the distance of Na$^+$ and Cl$^-$ between 2.0 Å to 4.0 Å, most of the polarization energies are more acceptable if compared to those calculated by QM/MM methods without damping function. Thus, the QM-MM damping function works well, but
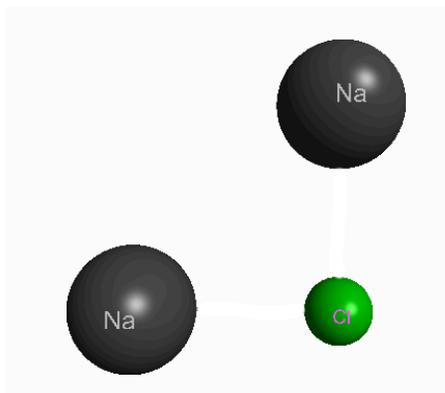
Figure 2-10. Model for two Na$^+$ ions and one Cl$^-$ ion

a more effective and accurate way should be introduced to deal with the interactions between multiple polarizable atoms.

Clearly, the polarization energy is more reliable by using the damping function with an adjustable d constant and a suitable cutoff distances r_d in QM/MM simulations
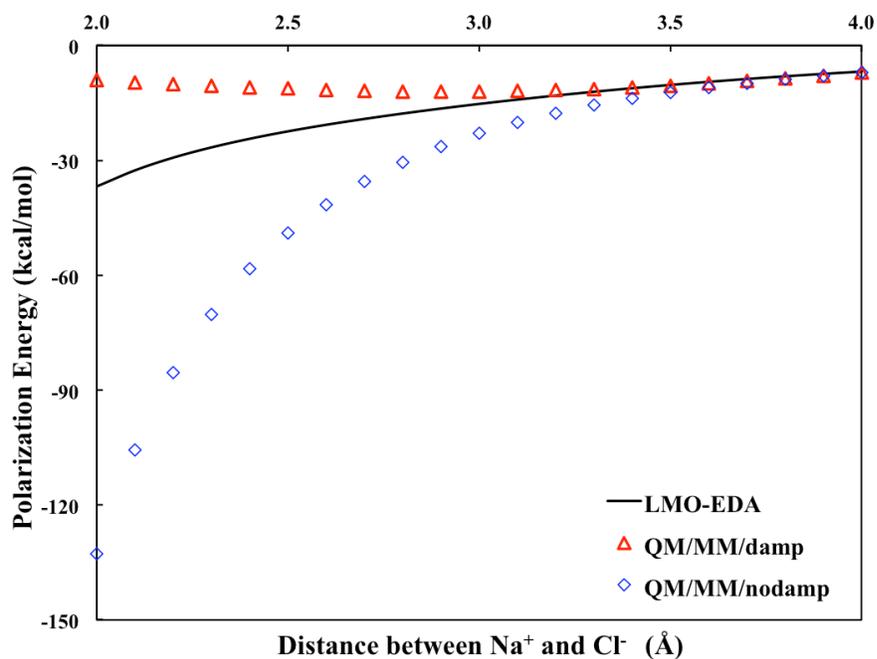
Figure 2-11. Polarization energy of Cl$^-$ calculated using LMO-EDA and QM/MM (with and without QM-MM damping function) for Na$^+$-Cl$^-$-Na$^+$ pair.

although problems may still exist in current work. The results show that the damping function works well in QM/MM simulations with polarizable force fields.

2.5   Conclusion

A damping function is introduced for QM-MM interaction when polarizable force field is used in the MM subsystem. This QM-MM damping function scales down the polarizability of MM atoms when they are close to QM atoms, and works well with the MM-MM damping functions that scale down the dipole-dipole interactions between MM atoms. The analytical gradients of the polarization energy with these QM-MM and MM-MM damping functions are also derived and implemented so geometry optimization and energy conserved MD simulation can be performed. The Gaussian-type QM-MM damping function introduced for QM/MM interactions performs very well according to the comparison of the polarization energies computed by using QM/MM methods and LMO-EDA method for Na^+-H_2O, Mg^{2+}-H_2O, Ca^{2+}-H_2O, Na^+-Cl^- and K^+-Cl^-. The results suggest that the damping constant d and the cutoff distance r_d in the Gaussian-type QM-MM damping function can be easily parameterized for different system.

CHAPTER 3 QM/MM Study on P450cam

3.1   Introduction

Cytochrome P450 enzymes (P450s) exist in a wide variety of biological systems. They function as catalysts in a number of biological oxidation reactions, such as the metabolism of drugs, xenobiotics and carcinogenesis[59-61]. One of the important reactions is the specific hydroxylation of non-active C-H to form C-O-H catalyzed by P450s at physiological conditions (Figure 3-1). A large amount of experimental mechanistic studies have been performed to understand how P450s catalyze this kind of reactions. Among them, a generally accepted mechanism called rebound mechanism is proposed[60, 62, 63]. In the rebound mechanism, a hydrogen abstraction step is considered as the rate-determining step (Figure 3-2). In the catalysis process, a hydroxylation transition state (TS$_H$) is formed from an oxo-ferric active compound, which is named as P450 compound I (Cpd I, Figure 3-3, RS in Figure 3-2). Cpd I is considered as the key oxidant in most of the important reactions catalyzed by P450s. Cpd I is predicted first by theoretical studies, then confirmed by experiments[64-67]. Thereafter, an iron-hydroxo intermediate radical pair (HYD in Figure 3-2) is yielded via the TS$_H$ transition state.
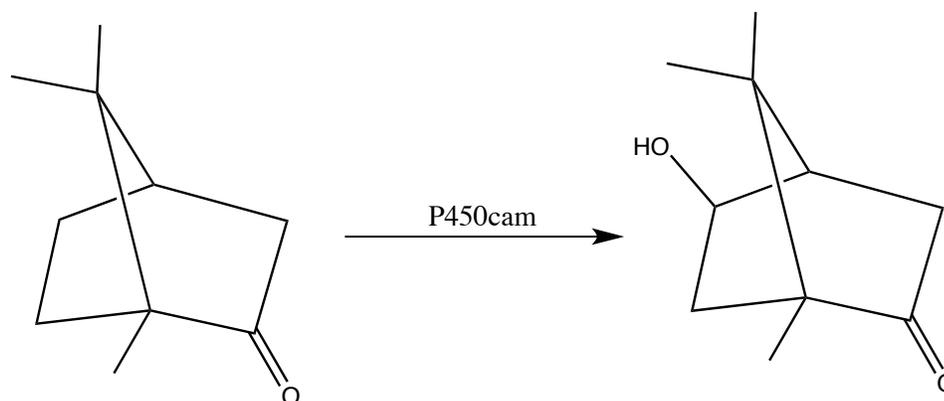
Figure 3-1. Hydroxylation of camphor catalyzed by P450cam
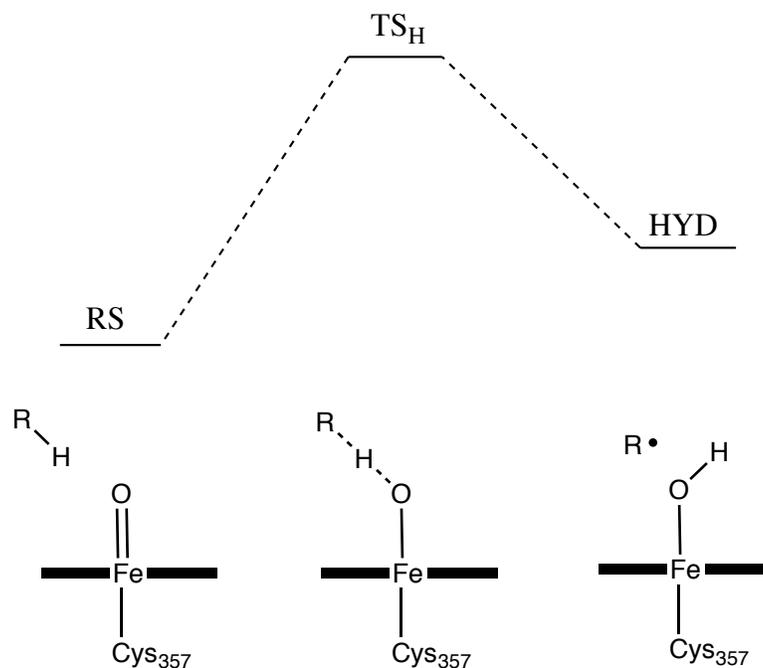
Figure 3-2. The rate-determining step (hydrogen abstraction step) in the rebound mechanism of C-H hydroxylation

Many QM/MM calculations have been performed to derive the activation free energy of the H-abstraction reaction. However, the results are controversial with values between 7 kcal/mol to 18 kcal/mol[68-72]. Because defects exist in most of the reported QM/MM studies, in this chapter, a general protocol for applying QM/MM methods to study the activation free energy is established. This protocol is applied to compute the activation free energy of the camphor hydrogen abstraction reaction catalyzed by P450cam. P450cam is a bacterial (*Pseudomonas Putida*) enzyme that catalyzes 5-exo-hydroxylation of camphor. It is the first crystallized cytochrome P450 enzyme[73] and has been used in a large amount of experimental and computational studies to illustrate the P450s catalytic mechanisms[74]. Since the hydrogen abstraction process conducted by Cpd I is generally considered as the rate-determining step[74], only the activation free energy of
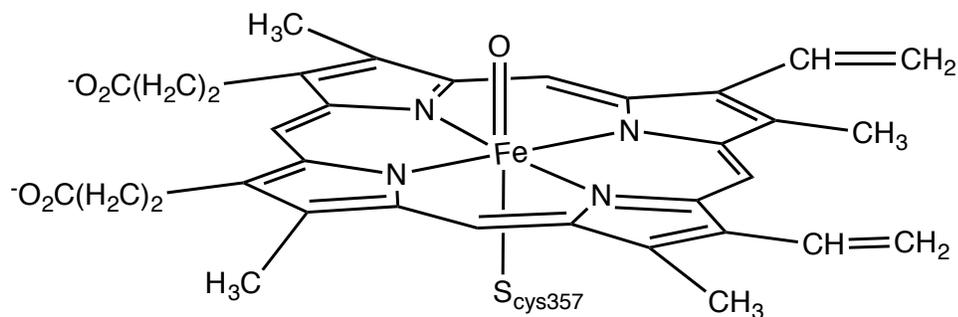
Figure 3-3. The structure of Cytochrome P450 Cpd I

the hydrogen abstraction step is studied in this work. By using the QM/MM protocol in this chapter, the calculated activation free energy is in good agreement with the experimental results and the results obtained from other QM/MM methods.

3.2 Computational Methods

3.2.1 Preparation

The X-ray structure 1DZ9[75] for P450cam (Figure 3-4; Graph generated by MacPyMOL[76]) from *Pseudomonas Putida* was obtained from the Protein Data Bank (PDB)[77]. All water molecules in 1DZ9 were kept except WAT2206, which is believed to be the water molecule formed by one of the two atoms in O_2. When Cpd I is formed, this water molecule should not present at the active site. Hydrogen atoms were added to the X-ray structure by using the WHAT IF web interface[78]. The file was manually edited so His355 is positively charged and Asp297 is neutral. The protein was described with the AMBER12 protein force field[79]. The water molecules in the PDB file were described with a three-point flexible and non-polarizable water model (QP301)[38]. Heme, Fe, O and camphor were described using a simplified universal force field implemented in QuanPol with the keywords LOUT=1 and NFFTYP=0. Lennard-Jones (LJ) parameters in this

force field are very similar to those in the AMBER force field. The atomic charges were determined by using accumulated bond polarization charges, and were similar to those parameterized in protein force field for common atoms such as H, C, N, O, F, P, S, and Cl. The atomic charges for metal ions, their ligand atoms, and ionized functional groups, such as ammonium and carboxylate, were modified manually. Covalent terms, including bond stretching, bond angle bending, dihedral angle rotation and dihedral angle bending (i.e., out of plane bending) parameters, were generated in a way that they favor the given conformation of the molecule. In this case, they were the covalent terms in camphor and heme, with the Fe and O ions. The AMBER12 force field file for the protein and the simplified universal force field file for the substrates were prepared separately and then combined together by using the keyword ICOMBIN=1 in QuanPol. After the combination, an additional bond stretching potential for Fe-S_{Cys357} with the force constant k=300 kcal/mol/Å^2 for $E=k(r-r_0)^2$ and r_0=2.271 Å, and an additional bond angle bending potential for Fe-S_{Cys357}-C_{Cys357} with the force constant k=50 kcal/mol/rad^2 and θ_0=111.5° for $E=k(\theta-\theta_0)^2$ were added manually. The standard AMBER12 charge for free thiolate S_{Cys357} was -0.8844 e. It was reduced manually by 0.5 e to -0.3844 e together with an equal increase of the Fe ion charge (finally Fe charge is +3.3 e). These partial atomic charge assignments, as force field parameters, were consistent with the following formal charges (or oxidation states): O is -2, Fe is +5, S is -1, each heme N is -0.5 (total -2). The net charge of the heme catalytic site is zero. The heme has two carboxylate groups. Each of the groups has -1 e charge. So the total charge of the heme/camphor is -2 e. These two negative charges are stabilized by three nearby positively charged residues: Arg112, Arg299, and His355. Arg112 is at the protein surface so its positive charge is partially
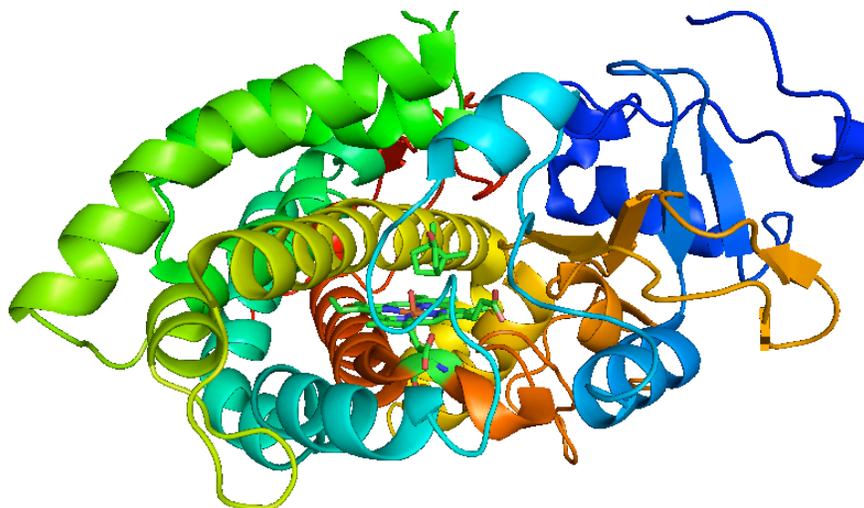
Figure 3-4. P450cam (PDB file 1DZ9 visualized with MacPyMOL)

stabilized by the bulk solvent. As mentioned above, the nearby Asp297 is neutral. The total charge of the protein and substrates is -14 e.

The protein/substrates were then solvated in an $80\times70\times90$ Å^3 periodic boundary condition water box filled with 13600 water molecules, and with 30 Na$^+$ and 16 Cl^{-1} ions randomly added. The AMBER12 LJ parameters were used for Na$^+$ and Cl$^-$. The total number of atoms is 48268, with a zero net charge. All water molecules (the crystalline waters in the PDB file and the 13600 added water molecules) were described with the three-point flexible and non-polarizable water model (QP301)[38].

3.2.2 Molecular Dynamics Simulation

Molecular dynamics (MD) simulation was run for the system by using pure MM method. The system was pre-equilibrated for 20,000 steps (20 ps). This process went smoothly. The volume and pressure of the system were stabilized. Then the system was equilibrated for one million steps (1 ns). Periodic boundary condition (PBC) was used in the MD simulation by using a shifting function in QuanPol with a cutoff distance of 12.0

Å for the charge-charge interactions, and a switching function in QuanPol with the switching range from 10.0 to 12.0 Å for the LJ interactions. The constant particle number, pressure and temperature (NPT, with P_{bath}=1.00 bar and T_{bath}=298.15 K) ensemble was used in the MD simulation with Berendsen barostat and thermostat[80]. The Beeman MD integrator[81] was used. After the equilibration, the average temperature and pressure were 298.13 K and 1.00 bar, respectively. The volume stabilized at $78.1 \times 68.3 \times 87.8$ Å^3. After the MD simulation, the added 30 Na$^+$ and 16 Cl$^-$ ions stayed in the bulk water, and did not penetrate into the protein. Using the one million MD configurations, the dielectric constant of the whole protein/water/ions system was simulated to be 77.6, a very reasonable value. After one million MD steps, the overall geometry of the protein is similar to that in the PDB file.

3.2.3 Geometry Optimization

Based on the geometry of the system at the last step of the MD simulation, QM/MM style UB3LYP (unrestricted B3LYP: Becke, three-parameter, Lee-Yang-Parr exchange-correlation functional[82]) geometry optimizations were performed for 1869 atoms around the reactive H atom of the camphor. In this study, we considered both S=1/2 (doublet state) and S=3/2 (quartet state) in the UB3LYP calculation to estimate the activation free energy. The 1869 atoms were selected by drawing a 16.0-Å-radius sphere around the 5-exo-H atom as in the last step of the MD simulation, including solvent water molecules. For comparability, the same 1869 atoms were consistently used in all subsequent geometry optimization and Hessian calculations. In the QM/MM calculation, the QM region had 106 atoms: the heme, Fe, O, camphor, and the side chain of Cys357

as SCH$_3$ (Figure 3-5). Of course, all of these 106 QM atoms are within the 1869 atoms to

be optimized. Periodic boundary condition (PBC) was used in the QM/MM system with

the volume fixed at the value of the last step of the MD simulation. The same shifting and

switching functions were used for MM charge-charge and LJ interactions as in the MD

simulation. The QM-MM interaction uses a special switching function[83] in the range

from 22.0 to 32.0 Å. The capping H atom method implemented in QuanPol was used to

treat the covalent bonds between QM and MM atoms [38] (here only the Cys357 lies across

QM and MM). The forces on all QM and MM atoms are evaluated analytically and

rigorously (including the effects of the shifting and switching functions) with an accuracy

near 1.0×10^{-6} hartree/bohr. In the QM/MM calculations a modified Wachters's triple zeta

valence (TZV[84], as implemented in GAMESS) basis set was used for Fe, the aug-cc-
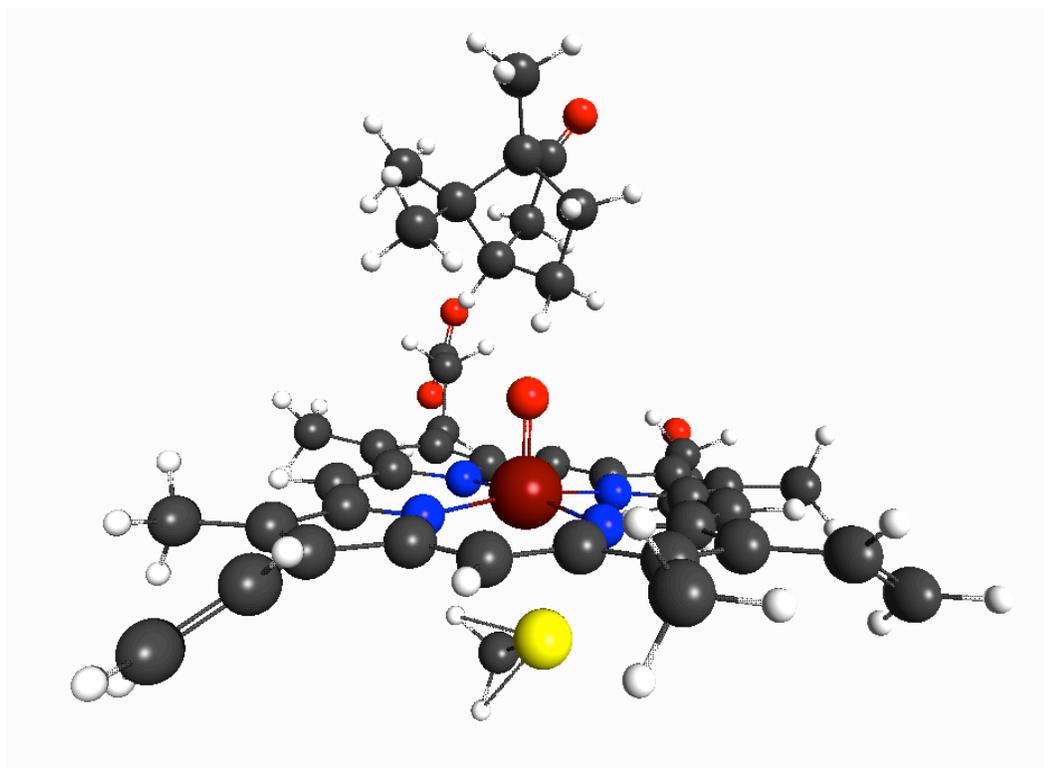
Figure 3-5. The atoms in QM region for reaction state (Cpd I and camphor)

pVDZ[85] basis set was used for the Fe ligand atoms (S, N, N, N, N, O) and the two reacting atoms (5-exo-H and C5) of camphor, and the 6-31G* basis set[86] was used for all other QM atoms. The total number of Cartesian Gaussian basis functions is 1077. The number of basis functions is larger than most of the calculations in the literature. In the current QM/MM calculation, atomic charges of the QM atoms were not used because they interact with the MM atomic charges using electrons and nuclear charges; the LJ interactions between QM and MM atoms were used.

The QM/MM geometry optimization of 1869 atoms (106 QM and 1763 MM) requires a significant computing time. In general, geometry optimization processes can be accelerated with Hessian methods. For the reactant state (RS), we used MM method to complete a geometry optimization of the 1869 atoms with a steepest-descent method. This typically requires tens of thousands of optimization steps and a few hours on a 32-core parallel computer. We found that a tight gradient criterion of 1.0×10^{-5} hartree/bohr can guarantee a good optimized minimum on the potential energy surface (PES). Then, a MM Hessian calculation was performed for the 1869 atoms via double displacement (total 11215 energy and gradient evaluations) method and a 0.01 bohr step size. Diagonalization of the mass-weighted force constant matrix (FCM) yielded no imaginary frequencies. The FCM obtained at the MM level was used to guide the QM/MM geometry optimization of the same 1869 atoms. Using a 32-core parallel computer, the QM/MM optimization took ~2.6 hours to finish one step, and took around 100 steps (260 hours) to reach the gradient criterion (maximum unsigned gradient 1.0×10^{-4} hartree/bohr, average unsigned gradient 0.333×10^{-4} hartree/bohr). The FCM was updated using the Broyden-Fletcher-Goldfarb-Shanno formula[87] in the QM/MM optimization process.

For the QM/MM transition state (TS) search, however, no MM force field is readily available to generate a FCM for the TS. We used the following procedure to generate a FCM for the TS. We implemented in QuanPol a keyword NQMVIB to input the specific QM atoms so they vibrate in Hessian calculation to update part of the FCM. The FCM is started from an existing reactant state. Here for P450, the FCM was obtained from the previous RS geometry optimization. The single displacement step size is set to be 0.01 bohr in this case. The MM charge is internally turned off so the QM-MM interactions are simply LJ interactions. This will not significantly affect the force constants between QM atoms, but can shorten the QM/MM computing time. The general way to use this function is to run a QM/MM geometry optimization for some steps and obtain a good approximate FCM for the reactant or product state, then reposition a few QM atoms to form the transition state. Here for P450, we assigned the camphor C-H distance to be 1.30 Å, the O-H distance to be 1.20 Å, and the O-C distance to be 2.50 Å, these values are known from the literature[71, 72, 88]. Then a QM/MM Hessian calculation is run with the specifically assigned NQMVIB atoms. Only the force constant matrix elements that belong to a pair of NQMVIB atoms are updated. This may lead to one or more imaginary frequencies depending on the quality of the guessed TS geometry. Visualization of the vibration modes can help one select the correct imaginary mode to follow by using the keyword IFOLOW=I (typically I is 1, the most negative mode) in the subsequent TS search for the QM/MM system, with the FCM supplied. This TS search may not necessarily lead to the anticipated TS geometry. It may be necessary to take the geometry after ~10 optimization steps, and run Hessian calculation with NQMVIB again to obtain a better FCM. In general, TS search is tricky so a few rounds may be required.

After several rounds, normal mode visualization using the MacMolPlt graphic software[89] shows that the imaginary frequency mode indeed describes the vibration of the H atom between the C and O atoms. This scheme is critical for accelerating QM/MM TS search. It is impractical to compute the FCM using the QM/MM method for all the 1869 atoms, which would take ~3.2 years using our parallel computers. This "Partial Vibrating Scheme" works very well, and is very efficient. For the purpose of making the final QM/MM energies comparable between the RS and the TS, the geometries of the entire QM/MM system must be very similar. Otherwise the energy difference cannot be taken as the activation energy of the TS. Therefore, we used the QM/MM optimized RS geometry and positioned the H atom to the middle of the O and C atoms to start the TS search. The FCM was updated using the Davidon-Fletcher-Powell formula[90, 91] in the QM/MM TS search process.

3.3   Results and Discussions

3.3.1   Activation Free Energy

For the doublet state S=1/2, the $<S^2>$ should be 3/4. However, the computed value from UB3LYP calculation is 1.80, implying that there is a large spin contamination. This large spin contamination is not sensitive to the geometry of the QM region, and is similar in both RS and TS. It suggests that the spin quantum number S should not be 1/2. Recent experimental results[65] suggest that the electronic spin quantum number of the heme/Fe/O system (the P450 Cpd I without substrate) can be effectively represented by S=1/2. However, our results show that the quartet state is better in representing the electronic structure. Therefore, in the following discussion we only present the results of S=3/2.

The P450cam Cpd I activation electronic energy computed using the UB3LYP/AMBER geometry optimization method is 22.7 kcal/mol for S=3/2. To find out the activation free energy, the zero point energy (ZPE) correction should be considered. The ZPE change is mainly due to the disappearance of the camphor C-H stretching mode, which has a frequency of ~2900 cm^{-1}, corresponding to a ZPE around 4.0 kcal/mol. A value is found in literature, which is 3.6 kcal/mol for quartet state[71]. With this correction, the activation free energy ΔG_a can be estimated as 19.1 kcal/mol for S=3/2. Lonsdale et al.[92] found that the inclusion of dispersion correction in DFT calculation lowers the activation electronic energy by ~3.6 kcal/mol for 5-exo-Hydroxylation of camphor and significantly improves the accuracy of activation energy. In our work, single point energy calculations by using the optimized geometry of RS and TS show that, the activation electronic energy is lowered by 3.9 kcal/mol with Grimme's empirical dispersion
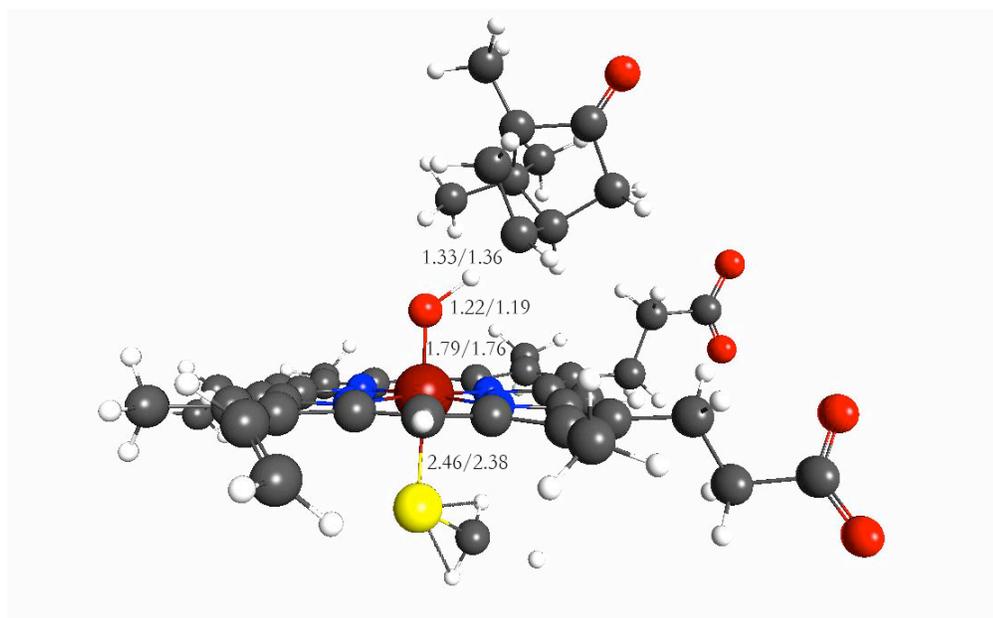
Figure 3-6. The optimized QM region of transition state geometry (distances are in Å, doublet/quartet).
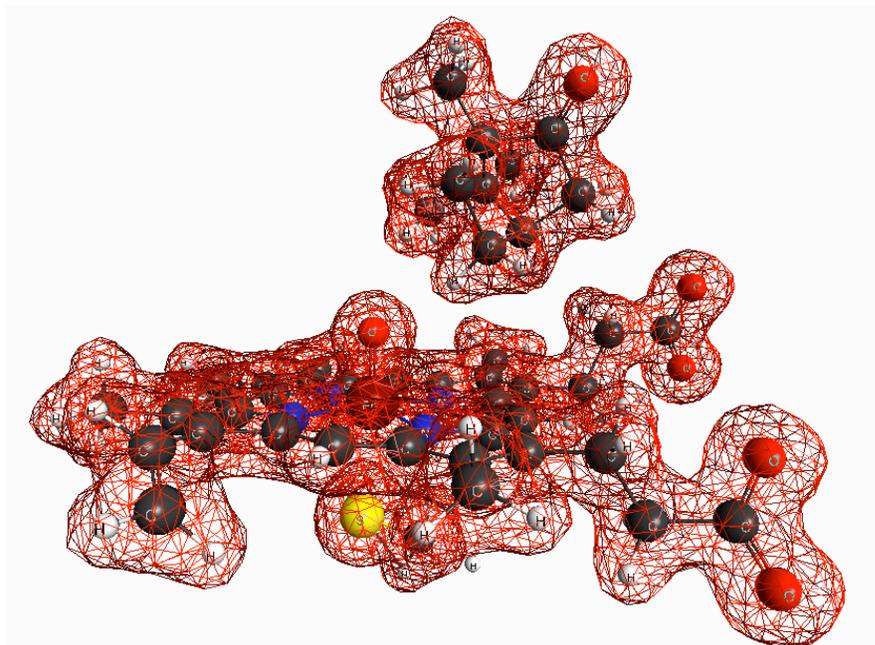
Figure 3-7. Visualization of spin densities in RS

correction[93, 94]. Therefore, with the empirical dispersion correction, the activation free energy should be lowered to 15.2 kcal/mol for the quartet state. This is the best estimation from the current work.

Experimental activation free energy can be estimated using the first-order rate constant reported by Rittle and Green[65], who obtained a lower limit of 1400 s[-1] for the thermophilic P450 from *Sulfolobus Acidocaldarius* and substrate lauric acid. The actual rate constant may be significantly higher. The higher limit of the activation free energy ΔG_a can be estimated as 12.2 kcal/mol at T=277 K with the unimolecular transition state theory formula:

$$k = \frac{k_B T}{h} \exp\left(\frac{-\Delta G_a}{RT}\right)$$

(3-1)
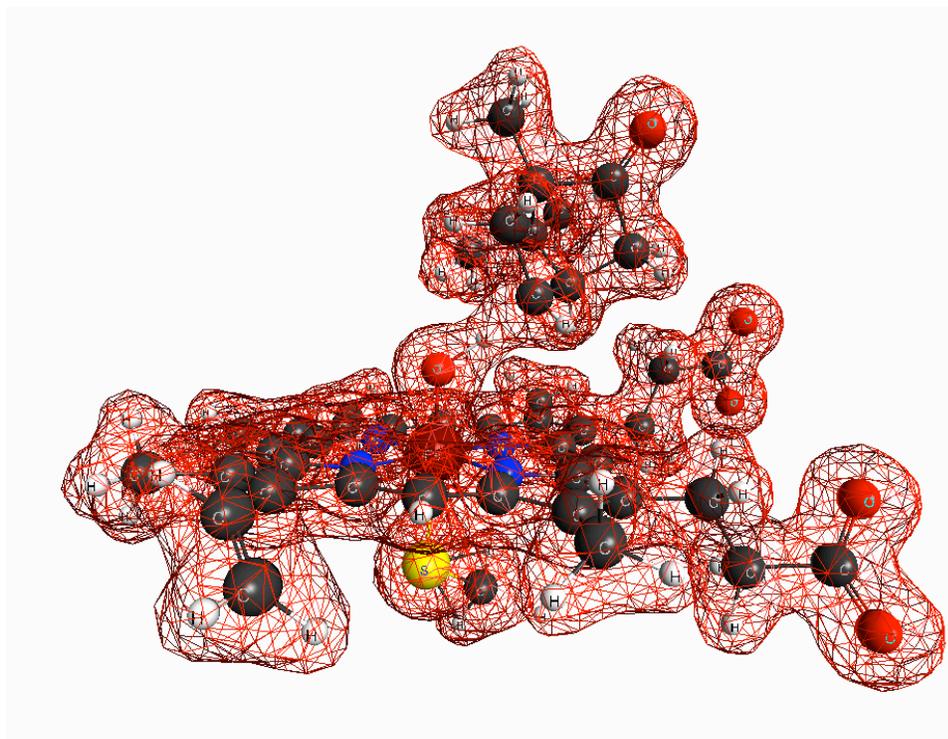
Figure 3-8. Visualization of spin densities in TS

Here k is the first order rate constant, k_B is Boltzmann constant, T is the temperature, h is Planck's constant, and R is the gas constant. It is very likely that the ΔG_a for P450 and camphor is similar to P450 and lauric acid. So the activation free energy given by the quartet state is higher than the experimental value by 3 kcal/mol. This difference is within the error of UB3LYP calculations.

The transition state geometry for quartet state show that the active atoms in QM region are Fe-S: 2.38 Å, Fe-O: 1.76 Å, O-H: 1.19 Å, C-H: 1.36 Å (shown in Figure 3-6). These distances are quite similar to most of current literature works. In general, the TS should be closer to the product side. As we see in Figure 3-2, the product of the H-abstraction reaction is a hydroxo intermediate and a radical. Therefore, the obtained TS geometry is reasonable. In this view, the H atom at the active site should be closer to O atom than to the C atom. Inspection of the spin density of the RS and TS shows that there
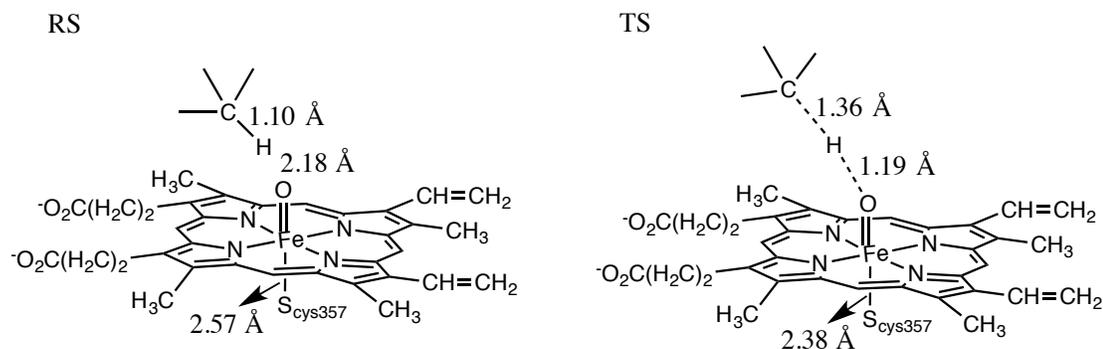
Figure 3-9. Selected structural data for reactant state and transition state

is a shift of unpaired spin density from the O atom to the C atom in camphor (Figure 3-7 and Figure 3-8). It is also observed that the unpaired spin density in Fe atom and S atom is decreased. The reduction of spin density on Fe and S can be used to explain why the distance between Fe and S is shortened from 2.57 Å in RS to 2.38 Å in TS (shown in Figure 3-9 and Table 3-1).

3.3.2 Comparison to Other Calculations

Various QM/MM methods have been used to study the catalytic mechanisms of P450s. Here we compare our methods and results with those in the literature (Table 3-2).

Guallar et al[64] obtained 11.7 kcal/mol for the quartet spin state, but with the water molecule Wat903 included near the oxo ligand of Cpd I. They used UB3LYP with the

Table 3-1. Spin densities (e/bohr^3) of the atoms in the active site

	Fe	O	C	H	S
RS	-0.19244	0.08643	0.00243	-0.00001	0.08833
TS	-0.17755	0.07182	0.11919	-0.01809	0.07360

force field OPLS-AA[95]. The system studied was treated as neutral and with 7448 atoms without consideration of adding solvent molecules. The QM/MM boundary was treated with frozen orbitals. With the ZPE contribution, the activation energy calculated by them was lowered to be 8.2 kcal/mol. In another paper, Guallar et al[70] figured out that the important intermediate distances in the transition state as O-H: 1.20 Å, C-H: 1.38 Å, Fe-O: 1.81 Å, Fe-S: 2.39 Å. Their calculations are questionable since the solvent effect is not considered in the QM/MM simulation. And the Wat903 should not be included in the active site since it is formed by one of the oxygen atom during the process of forming Cpd I. In addition, it has been shown that the inclusion of Wat903 will lower the calculated activation energy significantly[68]. In our calculation, the water molecule WAT2206 (the same water molecule with Wat903) was excluded. Furthermore, the protein system is solvated with water molecules and neutralized by adding 30 Na^+ and 16 Cl^- ions. This system is more similar to a real experimental system. Therefore, we believe our results are more reliable.

Schöneboom et al[88], studied the full mechanism of rebound scheme with the use of four different snapshots of MD simulations. They included a 16 Å water layer and the total number of atoms was 24394 with the charge of -10 e. They treated the QM/MM boundary by link atoms with charge shifting model. The force filed used for MM region was CHARMM[15]. By using the above QM/MM setting and optimizing 442 atoms in the QM/MM system, they obtained the activation electronic energy as 21 kcal/mol. The activation electronic energy obtained by them is similar to ours (22.7 kcal/mol). However, they did not neutralize the system, so the simulating system is not in consistent with a

real protein/solvent system. They only optimized 442 atoms, which is quite a small number compared to ours (1869 atoms). So our QM/MM methods are more reliable.

Altun et al[68] revisited the above cited QM/MM strategies with different treatment of basis set and different sizes of QM region. By using UB3LYP/CHARMM single point calculation on Guallar's model[64], they obtained the activation electronic energy range from 7.7 to 14.2 kcal/mol. Due to the different treatment of QM/MM boundary, they could not reproduce the exactly same results. So the treatment of the QM/MM boundary is very important in order to obtain reliable activation energy. In our work, we use a H-capping scheme to treat the QM/MM boundary. Our treatment is more general and simple compared to others' method. Guallar *et al*[64] handled the QM/MM boundary by frozen orbitals. In the frozen orbitals method, some localized orbitals are put in the boundary atoms but some of them are frozen without taking active in SCF iteration. This scheme is questionable for impractical frozen orbitals, the insufficiency in parameters and complexity in use. Schöneboom et al[88] used a link atom scheme with the charge shifting model. In this scheme, additional degrees of freedom will be introduced and sophisticated pseudopotentials need to be used. The H-capping method in QuanPol is similar to the "link atom" scheme, but no additional degrees of freedom are introduced and no carefully assigned pseudopotentials are needed. Furthermore, the H-capping method is generally applicable in all QM/MM methods without the need to be specifically parameterized for different MM force fields or QM methods. Altun et al[68] also presented that the protonation states of Asp297 almost do not affect the activation electronic energy. And the protonation of His355 would result in ~1.5 kcal/mol energy barrier difference as compared to the deprotonated one. They also showed that the presence of Wat903 would

result in a significant barrier lowering (4 kcal/mol). In our study, all of these issues are specifically considered (Asp297 protonated and His355 positively charged). More importantly, we excluded WAT2206 (reported as Wat903 in others' work) from the active site.

In another paper, Altun et al[69] studied the H-abstraction reaction of P450cam on the quartet potential energy surface with different basis sets and different treatment of QM regions with a similar treatment compared with their previous studies[68]. For the one with a QM region similar to ours, they included 120 atoms (with Cys357, CO group of Leu356 and NH-C$^\alpha$H unit of Leu358). When the Wat903 was excluded and Asp297 was protonated, with the number of basis functions being 329, 374 and 396, they obtained the H-abstraction activation electronic energy as 18.1, 21.0 and 19.1 kcal/mol respectively. Based on their results, different basis sets may result in a different (~3 kcal/mol) activation electronic energy. In our study, the QM region is smaller (106 atoms), but is sufficient. The number of basis functions is 1077, which is nearly 3 times larger than most of other authors' work.

Zurek et al[72] discussed how the activation energy is associated with the protonated Asp297 residue. They suggested that a better model would be obtained with the use of protonated Asp297. With a preliminary MD equilibration and QM/MM optimization, they obtained the activation electronic energy as 15.3/18.3 kcal/mol for the deprotonated/protonated models, respectively. With a larger QM region (106 atoms, similar to ours) and larger basis set (~400), they obtained the electronic energy barrier ~20 kcal/mol. They also found out that the key reaction distances for the transition state (TS) are Fe-S: 2.56 Å, Fe-O: 1.78 Å, O-H: 1.24 Å, C-H: 1.38 Å. These distances are kind

Table 3-2. Comparison of computed P450cam Cpd I activation free energy

	QM method	QM atoms	MM method	opt atoms	ΔE^{ele} (kcal/mol)	ΔG (kcal/mol)
Experiment						12.2
This work	B3LYP/[TZV/ACCD/6-31G*]	106	AMBER12	1869	22.7	15.2
Guallar et al	B3LYP/[LACVP/6-31G*]	126	OPLS-AA	N/A	11.7	8.2
Schöneboom et al	B3LYP/[LACVP/6-31G*]	84	CHARMM22	442	21.8	17.7
Altun et al	B3LYP/[LACVP/6-31G*]	120	CHARMM22	~400	19.1	15.1
Zurek et al	B3LYP/[LACVP/6-31G*]	106	CHARMM27	~1500	20	16
Lai et al	B3LYP/[LACVP/6-31G*]	N/A	CHARMM	N/A	17.4	7.8

of similar to our optimized TS geometries (Fe-S: 2.38 Å, Fe-O: 1.76 Å, O-H: 1.19 Å, C-H: 1.36 Å). In our QM/MM model, the system was solvated and neutralized with Na^+ and Cl^-. All these treatments make the system more close to a real experimental system. Most importantly, with a similar QM region, the number of basis functions is larger than theirs. From this point of view, our results are more convincible.

Lai et al[71] obtained a H-abstraction barrier as 17.4 kcal/mol for quartet state. They discussed that with the dispersion correction, the activation electronic energy should be lowered by ~6 kcal/mol. With the ZPE correction, they estimated the activation free energy as 7.8 kcal/mol. This value is quite smaller than experimental value (12.2 kcal/mol). They have excluded Wat903 in the calculation and no MM minimization was conducted before QM/MM calculation. The number of basis functions in their study is ~400. This number is quite smaller than ours (1077). Their optimized distances for reactive atoms in the TS are r_{OH}=1.30 Å, r_{CH}=1.28 Å. The geometry of the optimized transition state is quite different from ours (r_{OH}=1.19 Å, r_{CH}=1.36 Å). The hydrogen atom should be closer to the oxygen atom than to the carbon atom because the product of the transition state is iron-hydroxo and an intermediate radical (HYD in Figure 3-2). According to the above description, our results are more reliable.

In summary, our QM/MM calculation is more reasonable and reliable since we have used more convincible QM/MM settings and the QM/MM system is more similar to a real experimental system. The electronic energy difference between reaction state and transition state are computed as 22.7 kcal/mol for S=3/2. With the empirical dispersion correction, the activation electronic energy is estimated as 18.8 kcal/mol. With the ZPE correction, the activation free energy is estimated as 15.2 kcal/mol, close to the value (12.2 kcal/mol) estimated from the experimentally measured first order rate constant (1400 s^{-1})[66].

3.4   Conclusion

In this study, a general protocol for applying QM/MM methods is used to estimate the activation free energy of the hydrogen abstraction reaction in the hydroxylation process of camphor. A general and simple way to do the activation free energy calculation is introduced. The details of the protocol are shown below:

1) QM/MM model: protein system is solvated with PBC box and neutralized by adding Na$^+$ and Cl$^-$ into the solvent (a large molecular system with nearly 50000 atoms);

2) Perform MD equilibrium with PBC;

3) Reactant state geometry optimization: large number of optimization atoms (around 2000 atoms); simple and general treatment for QM/MM boundary (H-capping scheme);

4) Transition state search: FCM is updated with an affordable scheme (Partial Vibrating Scheme).

This QM/MM protocol can be generally used in QM/MM study of enzymes.

In this work, the QM/MM protocol is applied to calculate the activation free energy of the H-abstraction reaction catalyzed by P450cam. The P450cam protein system is solvated with water molecules and neutralized by adding Na^+ and Cl^- ions. Therefore, the modeled system is more similar to a real enzyme/solvent system. The QM/MM boundaries are treated with H-capping scheme, which is more general and simple compared to other schemes. A "Partial Vibrating Scheme" is used to generate the force constant matrix for transition state geometry search. By using this scheme, the computational cost in transition state geometry search is reduced. The geometry optimization of 1869 QM/MM atoms can be easily performed by using this scheme. With all these treatments, the electronic energy barrier between reaction state and transition state are obtained as 22.7 kcal/mol for S=3/2. With DFT empirical dispersion correction and the ZPE correction, the activation free energy can be estimated as 15.2 kcal/mol. The obtained distances for active atoms in QM region of transition state are Fe-S: 2.38 Å, Fe-O: 1.76 Å, O-H: 1.19 Å, C-H: 1.36 Å. The calculated activation free energy using our QM/MM protocol is in good agreement with experiments.

REFERENCES

1. E. Lewars, *Computational chemistry: introduction to the theory and applications of molecular and quantum mechanics*. (Springer, 2010).

2. I. N. Levine, *Quantum chemistry*. (Pearson Prentice Hall Upper Saddle River, NJ, 2009).

3. E. Schrödinger, Physical Review **28** (6), 1049-1070 (1926).

4. F. Jensen, *Introduction to computational chemistry*. (John Wiley & Sons, 2007).

5. A. Warshel and M. Levitt, Journal of molecular biology **103** (2), 227-249 (1976).

6. C. J. Cramer, *Essentials of computational chemistry: theories and models*. (John Wiley & Sons, 2013).

7. M. Born and R. Oppenheimer, Annalen der Physik **389** (20), 457-484 (1927).

8. D. R. Hartree, Mathematical Proceedings of the Cambridge Philosophical Society **24** (01), 89-110 (1928).

9. V. Fock, Phys. Rev **35**, 210 (1930).

10. J. C. Slater, Physical Review **35** (2), 210 (1930).

11. C. C. J. Roothaan, Reviews of modern physics **23** (2), 69 (1951).

12. G. G. Hall, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **205** (1083), 541-552 (1951).

13. A. R. Leach, *Molecular modelling: principles and applications*. (Pearson Education, 2001).

14. D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang and R. J. Woods, Journal of Computational Chemistry **26** (16), 1668-1688 (2005).

15. B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, Journal of Computational Chemistry **30** (10), 1545-1614 (2009).

16. M. Christen, P. H. Hünenberger, D. Bakowies, R. Baron, R. Bürgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholz, V. Kräutler, C. Oostenbrink, C. Peter, D. Trzesniak and

W. F. van Gunsteren, Journal of Computational Chemistry **26** (16), 1719-1751 (2005).

17. D. Bakowies and W. Thiel, The Journal of Physical Chemistry **100** (25), 10580-10594 (1996).

18. F. J. Vesely, Journal of Computational Physics **24** (4), 361-371 (1977).

19. F. H. Stillinger and C. W. David, Journal of Chemical Physics **69**, 1473-1484 (1978).

20. F. Stillinger, Journal of Chemical Physics **71**, 1647-1651 (1979).

21. P. Barnes, J. Finney, J. Nicholas and J. Quinn, Nature **282**, 459-464 (1979).

22. A. Warshel, Journal of Physical Chemistry **83** (12), 1640-1652 (1979).

23. P. Ren and J. W. Ponder, Journal of Computational Chemistry **23** (16), 1497-1506 (2002).

24. P. Ren and J. W. Ponder, The Journal of Physical Chemistry B **107** (24), 5933-5947 (2003).

25. J. W. Ponder and D. A. Case, in *Advances in Protein Chemistry*, edited by D. Valerie (Academic Press, 2003), Vol. Volume 66, pp. 27-85.

26. P. Cieplak, J. Caldwell and P. Kollman, Journal of Computational Chemistry **22** (10), 1048-1057 (2001).

27. Z.-X. Wang, W. Zhang, C. Wu, H. Lei, P. Cieplak and Y. Duan, Journal of Computational Chemistry **27** (6), 781-790 (2006).

28. G. A. Kaminski, H. A. Stern, B. J. Berne, R. A. Friesner, Y. X. Cao, R. B. Murphy, R. Zhou and T. A. Halgren, Journal of Computational Chemistry **23** (16), 1515-1531 (2002).

29. R. A. Friesner, in *Advances in Protein Chemistry*, edited by L. B. Robert and B. David (Academic Press, 2005), Vol. Volume 72, pp. 79-104.

30. B. T. Thole, Chemical Physics **59** (3), 341-350 (1981).

31. M. Masia, M. Probst and R. Rey, Chemical Physics Letters **420** (1–3), 267-270 (2006).

32. J.-P. Piquemal, L. Perera, G. A. Cisneros, P. Ren, L. G. Pedersen and T. A. Darden, The Journal of chemical physics **125** (5), 054511 (2006).

33. K. J. Miller, Journal of the American Chemical Society **112** (23), 8543-8551 (1990).

34. M. A. Thompson and G. K. Schenter, The Journal of Physical Chemistry **99** (17), 6374-6386 (1995).

35. K. Sneskov, T. Schwabe, O. Christiansen and J. Kongsted, Physical Chemistry Chemical Physics **13** (41), 18551-18560 (2011).

36. M. A. Thompson, The Journal of Physical Chemistry **100** (34), 14492-14507 (1996).

37. P. T. van Duijnen and M. Swart, The Journal of Physical Chemistry A **102** (14), 2399-2407 (1998).

38. N. M. Thellamurege, D. Si, F. Cui, H. Zhu, R. Lai and H. Li, Journal of Computational Chemistry **34** (32), 2816-2833 (2013).

39. H. Li, H. M. Netzloff and M. S. Gordon, The Journal of chemical physics **125** (19), 194103 (2006).

40. M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen and S. Su, Journal of Computational Chemistry **14** (11), 1347-1363 (1993).

41. M. S. Gordon and M. W. Schmidt, Theory and Applications of Computational Chemistry: the first forty years, 1167-1189 (2005).

42. M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. DeFrees and J. A. Pople, The Journal of Chemical Physics **77** (7), 3654-3665 (1982).

43. J. W. Caldwell and P. A. Kollman, The Journal of Physical Chemistry **99** (16), 6208-6219 (1995).

44. J. Wang, P. Cieplak, Q. Cai, M.-J. Hsieh, J. Wang, Y. Duan and R. Luo, The Journal of Physical Chemistry B **116** (28), 7999-8008 (2012).

45. H. C. Andersen, Journal of Computational Physics **52** (1), 24-34 (1983).

46. T. H. Dunning, The Journal of Chemical Physics **90** (2), 1007-1023 (1989).

47. D. E. Woon and T. H. Dunning, The Journal of Chemical Physics **100** (4), 2975-2988 (1994).

48. A. Schäfer, C. Huber and R. Ahlrichs, The Journal of Chemical Physics **100** (8), 5829-5835 (1994).

49. A. J. H. Wachters, The Journal of Chemical Physics **52** (3), 1033-1036 (1970).

50. B. Schropp and P. Tavan, The Journal of Physical Chemistry B **112** (19), 6233-6240 (2008).

51. G. Lamoureux, A. D. MacKerell Jr and B. Roux, The Journal of chemical physics **119** (10), 5185-5197 (2003).

52. C. Hättig and B. A. Heß, Journal of Chemical Physics **108**, 3863-3870 (1998).

53. A. Grossfield, P. Ren and J. W. Ponder, Journal of the American Chemical Society **125** (50), 15671-15682 (2003).

54. P. Su and H. Li, The Journal of Chemical Physics **131** (1), - (2009).

55. D. Feller, Journal of computational chemistry **17** (13), 1571-1586 (1996).

56. K. L. Schuchardt, B. T. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li and T. L. Windus, Journal of chemical information and modeling **47** (3), 1045-1052 (2007).

57. F. Jensen and T. Helgaker, The Journal of chemical physics **121** (8), 3463-3470 (2004).

58. F. Jensen, The Journal of Chemical Physics **136** (11), - (2012).

59. F. P. Guengerich, Chemical research in toxicology **21** (1), 70-83 (2007).

60. P. R. O. De Montellano, *Cytochrome P450: structure, mechanism, and biochemistry*. (Springer, 2005).

61. B. Meunier, S. P. de Visser and S. Shaik, Chemical Reviews **104** (9), 3947-3980 (2004).

62. J. T. Groves and G. A. McClusky, Journal of the American Chemical Society **98** (3), 859-861 (1976).

63. J. T. Groves, Journal of Chemical Education **62** (11), 928 (1985).

64. V. Guallar, M.-H. Baik, S. J. Lippard and R. A. Friesner, Proceedings of the National Academy of Sciences **100** (12), 6998-7002 (2003).

65. P. Bandyopadhyay, M. S. Gordon, B. Mennucci and J. Tomasi, Journal of Chemical Physics **116** (12), 5023-5032 (2002).

66. J. Rittle and M. T. Green, Science **330** (6006), 933-937 (2010).

67. I. Schlichting, J. Berendzen, K. Chu, A. M. Stock, S. A. Maves, D. E. Benson, R. M. Sweet, D. Ringe, G. A. Petsko and S. G. Sligar, Science **287** (5458), 1615-1622 (2000).

68. A. Altun, V. Guallar, R. A. Friesner, S. Shaik and W. Thiel, Journal of the American Chemical Society **128** (12), 3924-3925 (2006).

69. A. Altun, S. Shaik and W. Thiel, Journal of Computational Chemistry **27** (12), 1324-1337 (2006).

70. V. Guallar and R. A. Friesner, Journal of the American Chemical Society **126** (27), 8501-8508 (2004).

71. W. Lai and S. Shaik, Journal of the American Chemical Society **133** (14), 5444-5452 (2011).

72. J. Zurek, N. Foloppe, J. N. Harvey and A. J. Mulholland, Organic & Biomolecular Chemistry **4** (21), 3931-3937 (2006).

73. T. L. Poulos, B. C. Finzel and A. J. Howard, Journal of molecular biology **195** (3), 687-700 (1987).

74. S. Shaik, S. Cohen, Y. Wang, H. Chen, D. Kumar and W. Thiel, Chemical Reviews **110** (2), 949-1017 (2009).

75. S. Chalmet, D. Rinaldi and M. F. Ruiz-López, International Journal of Quantum Chemistry **84** (5), 559-564 (2001).

76. W. DeLano, DeLano Scientific LLC, Palo Alto, CA (2007).

77. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, Nucl. Acids Res. **28** (1), 235-242 (2000).

78. R. Rodriguez, G. Chinea, N. Lopez, T. Pons and G. Vriend, Bioinformatics **14** (6), 523-528 (1998).

79. D. Case, T. Darden, T. Cheatham III, C. Simmerling, J. Wang, R. Duke, R. Luo, R. Walker, W. Zhang and K. Merz, University of California, San Francisco **1** (2), 3 (2012).

80. A. H. Devries, P. T. van Duijnen, A. H. Juffer, J. A. C. Rullmann, J. P. Dijkman, H. Merenga and B. T. Thole, Journal of Computational Chemistry **16** (1), 37-55 (1995).

81. P. Bandyopadhyay and M. S. Gordon, Journal of Chemical Physics **113** (3), 1104-1109 (2000).

82. A. D. Becke, The Journal of Chemical Physics **98** (7), 5648-5652 (1993).

83. H. Li, Journal of Physical Chemistry A **115** (42), 11824-11831 (2011).

84. A. H. Steindal, K. Ruud, L. Frediani, K. Aidas and J. Kongsted, The Journal of Physical Chemistry B **115** (12), 3027-3037 (2011).

85. T. H. Dunning, Journal of Chemical Physics **90** (2), 1007-1023 (1989).

86. M. M. Francl, W. J. Pietro, W. J. Hehre, J. S. Binkley, M. S. Gordon, D. J. Defrees and J. A. Pople, Journal of Chemical Physics **77** (7), 3654-3665 (1982).

87. A. Ohrn and G. Karlstrom, Molecular Physics: An International Journal at the Interface Between Chemistry and Physics **104** (19), 3087 - 3099 (2006).

88. J. C. Schöneboom, S. Cohen, H. Lin, S. Shaik and W. Thiel, Journal of the American Chemical Society **126** (12), 4017-4034 (2004).

89. B. M. Bode and M. S. Gordon, Journal of Molecular Graphics and Modelling **16** (3), 133-138 (1998).

90. W. C. Davidon, SIAM Journal on Optimization **1** (1), 1-17 (1991).

91. R. Fletcher and M. J. Powell, The Computer Journal **6** (2), 163-168 (1963).

92. R. Lonsdale, J. N. Harvey and A. J. Mulholland, The Journal of Physical Chemistry Letters **1** (21), 3232-3237 (2010).

93. S. Grimme, Journal of Computational Chemistry **25** (12), 1463-1473 (2004).

94. R. Peverati and K. K. Baldridge, Journal of Chemical Theory and Computation **4** (12), 2030-2048 (2008).

95. G. A. Kaminski, R. A. Friesner, J. Tirado-Rives and W. L. Jorgensen, The Journal of Physical Chemistry B **105** (28), 6474-6487 (2001).