UNIVERSITY OF CALGARY

Multiscale Simulation of mRNA Synthesis by RNA Polymerase II

by

Rui Zhang

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

Department of Chemistry

CALGARY, ALBERTA

December, 2013

Abstract

RNA polyermase II, a crucial enzyme for gene expression in eukaryotes, synthesizes messenger RNAs with high selectivity. Despite its importance, its selection and catalysis mechanism is not well understood. We first investigated, by a stochastic simulation algorithm, the entire nucleotide addition cycle based on an event-driven model. The results suggest that the discrimination of unmatched nucleotide mainly lies in its thermodynamic instability in the addition site, and the selectivity for the 2'-OH is from the catalytic reaction. To understand the stability of different nucleotides in the addition site on the atomistic level, we performed MD and free energy perturbation simulations, and found that mutating a cognate GTP to a non-cognate UTP in the active site costs ~16.8kcal/mol while mutating a cognate GTP to a 2'-deoxyGTP costs ~2kcal/mol. Since two binding sites exist in the enzyme, we conducted molecular dynamics and umbrella sampling calculations to simulate the entry of a cognate GTP from the entry site to the addition site. The results demonstrate that two key motifs, the trigger loop and the bridge helix, play important roles in this process. Facilitated by these two motifs, the NTP entry is a spontaneous process with an energy decrease of ~6kcal/mol. Simulation of the catalytic reaction requires a quantum mechanical/molecular mechanical (QM/MM) method to adequately describe the reaction centre and enzyme surroundings. Therefore, we reviewed QM/MM methods in the literature and implemented our own version using CHARMM and deMon2k. With this QM/MM implementation, we performed geometry optimization and MD simulations on the system at the level of DFT/MM. To speed up the calculations and cover more possible reaction pathways, we employed a specifically parametrized semiempirical method – AM1/d-PhoT for the reaction pathway search. The results reveal a proton-transfer-facilitated mechanism. While the acceptor of the initial proton transfer may vary depending on the particular conformation of the active site,

all possible routes converge to the same destination. Comparison between different models shows that the role of Mg^{2+} (A) is more structural than catalytic.

Preface

In this thesis, I provide five chapters of original work, consisting of both methodology development and applications, towards a common theme – understanding the mRNA synthesis by RNA polymerase II (RNAP II) on multiple levels. Each research chapter is based on either a previously published paper or a paper submitted for publication, with complete abstract, introduction, methods, results, discussion, conclusions and bibliography.

I will first provide, in Chapter 1, a short background on crystal structures of RNAP II, previous relevant experimental and computational studies, and an introduction to the theories of Kinetic Monte Carlo (KMC), Molecular Dynamics (MD), Free Energy Perturbation (FEP), Umbrella Sampling, Quantum Mechanics/Molecular Mechanics (QM/MM), and Relaxed Surface Scans (RSS). Chapter 2 presents work published in the journal Interdisciplinary Sciences: Computational Life Sciences in 2009, on kinetic simulations of the mRNA synthesis cycle. Chapter 3 is work on the binding and selection mechanism of RNA polymerase II using molecular dynamics techniques, which has been submitted for publication. To further study the catalytic mechanism, a hybrid QM/MM approach is required. Chapter 4 presents a comprehensive review of the state-of-the-art developments in the QM/MM field from a book chapter in Advances in Quantum Chemistry, 2010. Chapter 5 presents our own implementation between CHARMM and deMon2k, and the test cases for this implementation. This work was published in the Journal of Computational Chemistry, 2010. Chapter 6 is our work on the reaction pathway search for the catalytic reaction in the RNAP II system using the CHARMM-deMon2k interface and another QM/MM method – AM1-dPhoT/MM. This manuscript has been submitted for publication.

For Chapter 2, I helped build the model, ran all the calculations, summarized the results and wrote part of the manuscript. For Chapter 4, I modified the deMon2k source code, coded part of the interface in CHARMM, helped run the tests and wrote roughly half of the manuscript. I wrote the remainder of the chapters, although acknowledging tremendous help and guidance from coworkers. I helped conceive and design all the simulations, ran all the simulations and analyzed most of the results.

The following papers have been reproduced with permissions as Chapters 2, 4 and 5:

Zhu, R., de la Lande, A., **Zhang, R.**, and Salahub, D.R., *Exploring the Molecular Origin of the High Selectivity of Multisubunit RNA Polymerases by Stochastic Kinetic Models.* Interdisciplinary Sciences-Computational Life Sciences, 2009. **1**(2): p. 91-98.

Zhang, R., Lev, B., Cuervo, J.E., Noskov, S.Y., and Salahub, D.R., *A Guide to QM/MM Methodology and Applications.* Advances in Quantum Chemistry, Vol 59, 2010. **59**: p. 353-400.

Lev, B.[*], **Zhang, R.**[*], De la Lande, A., Salahub, D., and Noskov, S.Y., *The QM-MM Interface for CHARMM-deMon.* Journal of Computational Chemistry, 2010. **31**(5): p. 1015-1023. [*]*These authors contributed equally.*

Chapters 3 and 6 are based on the following submitted manuscripts:

Zhang, R., Silburt, J. and Salahub, D., *Bridge Helix and Trigger Loop In Action – How RNA Polymerase II Binds And Selects NTPs.* Submitted, 2013

Zhang, R., Bhattacharjee, A., Salahub, D., and Field, M., *Reaction mechanism in RNAP II -- Proton relay via competing routes.* Submitted, 2013.

The following paper was also completed during my Ph.D:

Alvarez-Ibarra, A., Koster, A.M., **Zhang, R.**, and Salahub, D.R., *Asymptotic Expansion for Electrostatic Embedding Integrals in QM/MM Calculations.* Journal of Chemical Theory and Computation, 2012. 8(11): p. 4232-4238.

Acknowledgements

Many thanks to my supervisor, Dr. Dennis Salahub, for all of the support throughout my Ph.D. and for providing a terrific learning environment. I have learned a great deal, and have had all possible opportunities for travel, computer resources, and career development. My success as a graduate student is mostly due to Dennis' mentoring.

I would like to thank former and current Salahub group members, collaborators from around the world, and specifically, my collaborators Joey Silburt and Dr. Anirban Bhattacharjee. Without Joey's help, I would not have been able to try out so many ideas on simulating this difficult system and obtain convincing results. Without Anirban's collaboration, I would not have been to have access to the pDynamo package and complete all the computations by myself. Special thanks to former group members, Dr. Aurelien De La Lande, Dr. Rui Zhu, Dr. Yue Zhang and Dean Lang, for their tremendous help in my research.

Thanks to my supervisory committee and the support staff at the University of Calgary. Dr. Noskov and Dr. Kusalik have both been excellent mentors. I appreciate all the advice and knowledge you have provided throughout my Ph.D. Thank-you to the administrator Joyce Simoes for all her help in my work and life.

Thanks to Compute Canada for computer resources and to their support staff.

Dedication

To my parents.

Table of Contents

List of Tables

List of Figures and Illustrations

List of Symbols, Abbreviations and Nomenclature

Symbol	Definition
RNA	Ribonucleic acid
NTP	Nucleotide triphosphate
DNA	Deoxyribonucleic acid
RNAP	Ribonucleic acid polymerase
GTP	Guanosine triphosphate
CTP	Cytidine triphosphate
UTP	Uridine triphosphate
2’-dGTP	2’-deoxy guanosine triphosphate
A site	Addition site
E site	Entry site
BH	Bridge helix
TL	Trigger loop
PDB	Protein data bank
MD	Molecular dynamics
MM	Molecular mechanics
QM	Quantum mechanics
FEP	Free energy perturbation
KMC	Kinetic Monte Carlo
DFT	Density functional theory
AM1	Austin Model 1
RSS	Relaxed surface scans

CHAPTER ONE: INTRODUCTION

1.1 RNA Polymerase II

As the first step in gene expression, transcription is an end point of a great many signal transduction pathways. Extensive studies have been focused on transcription in eukaryotic cells, which is highly instructive in the understanding of human gene expression [1]. Three types of DNA-dependent RNA polymerases have been discovered to be responsible for the gene transcription of eukaryotic cells. Among them, RNA polymerase II, which catalyzes the synthesis of messenger RNAs, is crucial to gene transcription and is of primary interest among other RNA polymerases [2-6]. RNAP II has a total molecular weight of 514 kDa and comprises 12 subunits, ten of which form a structurally conserved core. Transcription catalyzed by RNAP II can be divided into three mechanistically distinct stages: initiation, elongation and termination. During initiation, RNAP II recognizes a promoter, unwinds DNA near the start site and begins RNA synthesis. The following elongation of the RNA transcript proceeds with uninterrupted synthesis of RNA chains thousands of nucleotides long. This elongation process is not terminated until RNAP II recognizes terminator nucleotides of the DNA template.

Messenger RNAs are synthesized during the elongation process by RNA Pol II, where the high selectivity of the polymerase is fully exerted. The transcription elongation complex is composed of RNAP II, the unwound double-stranded DNA and the RNA transcript. In each elongation cycle, a nucleoside triphosphate (NTP), whose base can pair with the corresponding DNA template nucleotide, first binds into the active site of RNAP II. It is then added to the growing RNA 3’ end forming a phosphodiester bond with the help of RNAP II. Upon the completion of each elongation cycle, the RNAP II translocates along the DNA and proceeds to elongate the RNA transcript.

In fact, a cell contains a pool of various kinds of NTPs such that RNAP II must select matched NTPs to the DNA template nucleotide over the unmatched NTPs (wrong nucleotides) and 2'-deoxyNTPs (wrong sugar ring). This means that if the DNA template nucleotide is a CTP, RNAP II has to select a GTP over ATPs, UTPs, CTPs, TTPs and all 2'-deoxyNTPs. As an example, chemical structures of CTP, GTP and 2'-deoxyGTP are shown in Figure 1-1.

CTP GTP 2'-deoxyGTP

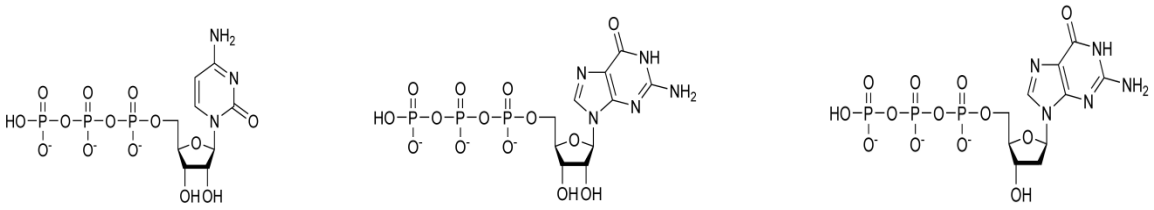


Figure 1-1: Chemical structures of CTP, GTP and 2'-deoxyGTP

Information about the binding mechanism of RNAP II has been revealed in crystal structures of the transcription complex. As shown in the crystal structure with a matched NTP (PDB: 2E2H), the catalytic site is composed of an α-helical structure denoted as the bridge helix (BH), a flexible loop motif termed the trigger loop (TL), and two magnesium ions – Mg(A) and Mg(B) in addition to surrounding protein residues of the RNAP II domains Rpb1 and Rpb2 [7]. This catalytic site (Figure 1-2A) is termed the addition site (A site) where a matched NTP binds the DNA template and is added to the RNA primer. Intriguingly though, another binding site exists, as determined from a protein crystal using a 2'-deoxyGTP [5, 7]. This site (Figure 1-2B) is termed the entry site (E site) which serves as an entrance for the passage leading to the A site. Distinctly, all nucleotides bind to the E site whereas only a nucleotide that is complementary to the template can further bind to the A site. Importantly, both the BH and the TL appear in distinctly different states between the E site and A site (Figure 1-2A and 1-2B). In the A site, the BH bends in the presence of a matched nucleotide, while it is straight upon nucleotide binding at

the E site. The TL was found to open the A site when a nucleotide is in the E site and close the A site when a matched nucleotide arrives in the A site.

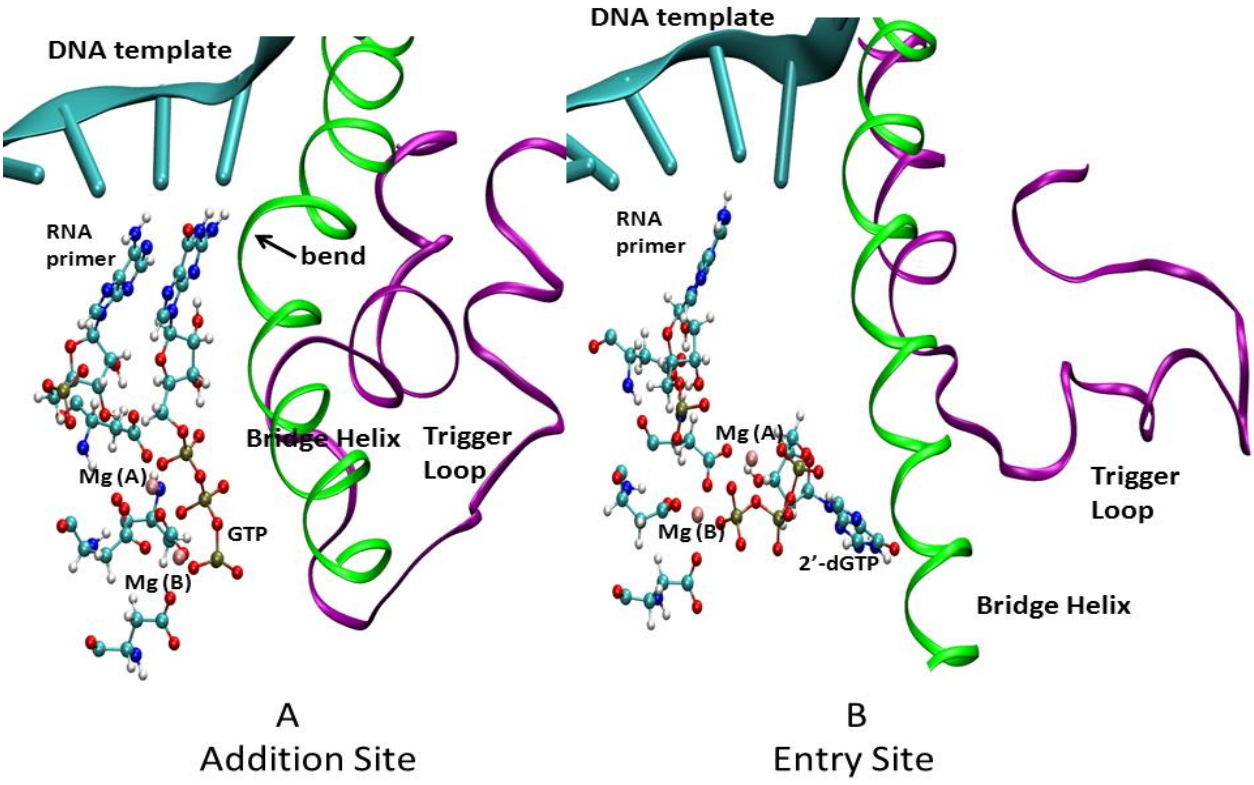


Figure 1-2: Crystal structures of a GTP in the addition and entry site

A) Crystal structure of a GTP in the addition site B) Crystal structure of a 2'-dGTP in the entry site where the trigger loop is colored in purple, the bridge helix in green, Mg ions in pink, carbon atoms in cyan, hydrogen atoms in white, oxygen atoms in red, nitrogen atoms in blue and phosphorus atoms in tan.

When a cognate NTP enters the A site and matches with the DNA template, the catalytic reaction takes place. The mechanism of this catalytic reaction – nucleotidyl transfer, is illustrated in Figure 1-3. Crystal structures of the system in this stage have also been resolved: 2E2H [5] and 2E2J [5] in PDB code. However, to obtain a complex with the substrate bound in the active site, chemical modifications were made in both structures. In 2E2H, the 3'-OH of the RNA primer was removed and in 2E2J, the $O_{\alpha\beta}$ (the oxygen between the α- and β-phosphate) was

replaced by a methylene group. The modification in 2E2H results in no coordination between the 3'O and the Mg^{2+} (A) (Figure 1-4A) while the NTP is in good coordination with both Mg^{2+}. The modification in 2E2J leads to a large gap between the RNA primer and the NTP as a result of the weak interaction between the Mg^{2+} and the triphosphate of the NTP. Unlike in 2E2H though, the 3'O coordinates well with Mg^{2+} (A) (Figure 1-4B).

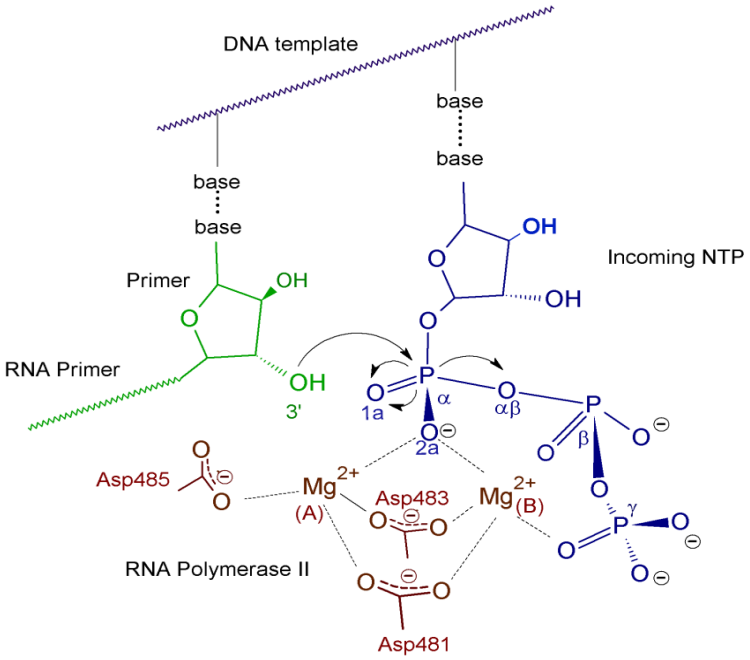
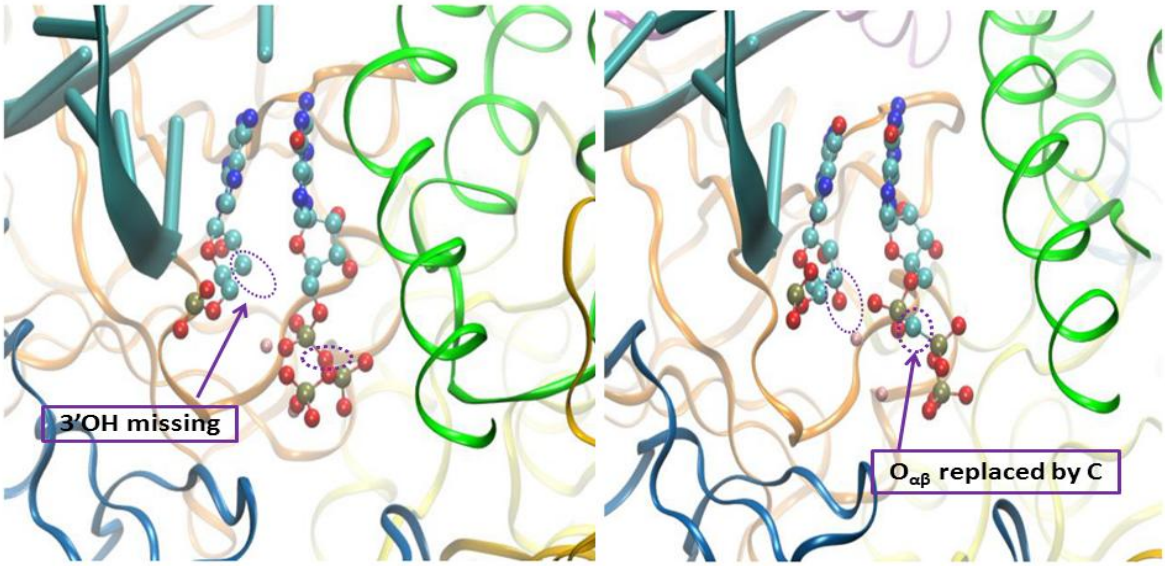


Figure 1-3: The nucleotidyl transfer reaction mechanism in RNAP II



A. Active site of 2E2H B. Active site of 2E2J

Figure 1-4: Active sites in the crystal structures

A) Active site in 2E2H B) Active site in 2E2J where the 3'-O and $O_{\alpha\beta}$ positions are circled in purple dashes

When models were built based on these crystal structures with chemical modifications to the substrates, the original structures of the substrates were properly restored. Details of model-building are presented in Chapters 2 and 6.

1.2 Methods and theory

My thesis is based on kinetic, molecular dynamics and quantum mechanical/molecular mechanical simulations. The kinetic simulations are based on the kinetic Monte Carlo method previously coded with MATLAB by our group [8]. MD simulations are performed using the NAMD software package [9]. QM/MM calculations are performed with the CHARMM-deMon interface [10] and the pDynamo package [11].

1.2.1 Kinetic Monte Carlo simulation

The time evolution of some processes with known given rates can be simulated numerically. One method to simulate the kinetics of these processes is kinetic Monte Carlo, since randomly generated numbers are adopted. An efficient algorithm implementing the kinetic Monte Carlo method is the Gillespie algorithm [12]. A succinct overview of the Gillespie algorithm is given below.

*Step **0.*** (Initialization). Input the desired values for the M reaction constants c_l, ... ,c_M and the N initial molecular population numbers X_I, ... ,X_N. Set the time variable t and the reaction counter n both to zero. Initialize the unit interval uniform random number generator.

*Step **1***. Calculate and store the M quantities $a_l = h_l\, c_l$, ... ,$a_M = h_M\, c_M$ for the current molecular population numbers, where h, is a function of X_I, ... ,X_N defined as above, and $h_\mu \equiv$ number of distinct R_μ molecular reactant combinations. Also calculate and store as a_0 the sum of the M a_ν values. As an example, for the reaction A + 2B $\rightarrow$ D, $h = X_A \binom{2}{X_B}$, where $\binom{2}{X_B}$ is the combination for choosing any 2 molecules from X_B molecules.

*Step **2.*** Generate two random numbers r_1 and r_2 using a unit-interval uniform random number generator, and calculate τ and μ according to

$$\tau = (1/a_0)\ln(1/r_1) \quad \text{and} \quad \sum_{\nu=1}^{\mu-1} a_\nu < r_2 a_0 \le \sum_{\nu=1}^{\mu} a_\nu$$

*Step **3.*** Using the values of τ and μ obtained in step 2, increase t by τ**,** and adjust the molecular population levels to reflect the occurrence of one R_μ reaction. Then increase the reaction counter n by 1 and return to step 1.

The Gillespie algorithm has been coded with MATLAB. Previous work by our group [8], as well as its applications by others [13], have proven that it is a powerful tool to study interrelated stochastic processes.

1.2.2 Molecular Dynamics

Much of the material and equations throughout this section are adapted from the NAMD user manual [14].

1.2.2.1 Potential energy function

MD simulations use an empirical force field for the potential energy. A force field defines all of the parameters in the potential function for each atom and molecule.

$$E = \sum K_r \left(r - r_{eq}\right)^2 + \sum K_\theta \left(\theta - \theta_{eq}\right)^2 + \sum \frac{V_n}{2}\left[1 + \cos(n\varphi - \gamma)\right] + \sum \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon r_{ij}}\right]$$

Eq. 1

The above equation is a generic example for typical molecular dynamics simulations. The first three terms are intra-molecular terms. The first term is for bonded interactions, where K_r is the bond force constant, and r_{eq} is the equilibrium bond distance. The second term is the angle potential, between three bonded atoms in a molecule. As with bonds, K_θ is the force constant, and θ_{eq} is the equilibrium angle. The third term is the dihedral potential for the angle between the planes formed by four consecutive atoms in a molecule. Improper torsions are also used to enforce specific conformations for out-of-plane bending in molecules.

The fourth term is the non-bonded interactions between all i and j atoms in the system. The first term in the double summation is the Lennard-Jones 12-6 potential to model the Van der Waals interactions. The second term is Coulomb's law for the electrostatic interactions between

the atoms' partial charges. Within molecules, non-bonded interactions are neglected between atoms separated by 1 or 2 bonds. Due to computational complexity when N gets large, calculating all the interactions is impractical. Cut-offs and long-range electrostatic schemes are used to reduce the number of interactions.

I have used a cut-off of 12Å and a switch function beyond 8 Å for Van der Waals interactions throughout my MD simulations. Short range non-bonded interactions are calculated every step during MD simulations. Pair lists are maintained, and updated every 20 steps, with only interactions between atoms within the cut-off calculated. The force field used throughout this thesis is the CHARMM force field [15].

For all of my MD simulations, I have used the particle mesh Ewald method [16, 17] for long-range electrostatics. The Ewald method was developed for calculating the electrostatic potential of ionic crystal systems. In an infinite periodic system, the electrostatic potential can be determined more efficiently in reciprocal space. The particle mesh Ewald method is a variant of this method, where the charges are placed on a grid for computational efficiency. Using the Fourier transformation of the grid, the reciprocal space contribution can be summed more efficiently. The force on each atom is then obtained by interpolation of the grid. PME is by far the most common method for electrostatic interactions in biomolecular dynamics simulations.

1.2.2.2 Periodic boundary conditions

All of my simulations have used periodic boundary conditions, which are artificial, but computationally efficient. Boundaries or walls can create long-range order and artifacts, relative to the size of the simulation. It is important to have a box that is big enough to avoid artifacts, and have enough water to keep the protein properly solvated, but this has to be weighed against the additional computational cost of more atoms.

1.2.2.3 Langevin dynamics

In general, molecular dynamics simulations solve Newton's equations of motion numerically for each atom in the simulation, using an empirical potential energy function. The force on each atom can be determined by taking the negative derivative of the potential energy. Once the force is calculated, Newton's equations of motion are used to update the positions of each atom. There is a number of methods available for numerical integration of the equations of motion. As an example, the Verlet integration method [18] is described in equations 2 and 3.

$$\vec{x}(t+\Delta t) = 2\vec{x}(t) - \vec{x}(t-\Delta t) + \vec{a}(t)\Delta t^2 \qquad \text{Eq. 2}$$

$$\vec{v}(t) = \frac{\vec{x}(t+\Delta t) - \vec{x}(t-\Delta t)}{2\Delta t} \qquad \text{Eq. 3}$$

where $\vec{x}$ is the position, the velocity $\vec{v} = \dot{\vec{x}}$, the acceleration $\vec{a} = \ddot{\vec{x}}$, and t the time. The last term of both equations represent the neglected higher order terms.

In my MD simulations, I have employed Langevin dynamics -- a variant of Newton's equation, with a stochastic force introduced. For a system of N particles with masses M, with coordinates $\vec{x} = \vec{x}(t)$, the resulting Langevin equation is

$$M\ddot{\vec{x}} = -\nabla U(\vec{x}) - \gamma M\dot{\vec{x}} + \sqrt{2\gamma k_B}TMR(t) \qquad \text{Eq. 4}$$

where $U(\vec{x})$ is the particle interaction potential; ∇ is the gradient operator such that $-\nabla U(\vec{x})$ is the force calculated from the particle interaction potentials; the dot is a time derivative such that $\dot{\vec{x}}$ is the velocity and $\ddot{\vec{x}}$ is the acceleration; T is the temperature, k_B is Boltzmann's constant; and $R(t)$ is a delta-correlated stationary Gaussian process with zero-mean. γ is the friction coefficient (or damping constant) and $-\gamma M\dot{\vec{x}}$ is the friction force by the solvent. I have used a friction coefficient of 10ps-1 for all my MD simulations. To integrate the Langevin equation,

NAMD uses the Brünger-Brooks-Karplus (BBK) method [19], a natural extension of the Verlet integration method [18] for the Langevin equation. In the framework of Langevin dynamics, a constant temperature is maintained by the friction force and the stochastic force, ensuring a thermostat. In my simulations, the temperature was maintained at 300K.

As all my simulations were run with a constant number of atoms, constant pressure, and constant temperature (NPT ensemble), a barostat was also required. Pressure is controlled by dynamically adjusting the size of the unit cell and rescaling all atomic coordinates (other than those of fixed atoms) during the simulation. I used the Langevin piston Nose-Hoover method in NAMD as the barostat. It is a combination of the Nose-Hoover constant pressure method [20], with piston fluctuation control implemented using Langevin dynamics [21]. With this barostat, the pressure in my simulations was maintained at 1atm when the oscillation time of the piston was set to 100fs and the decay time 50fs.

1.2.3 Free energies: Free energy perturbation

1.2.3.1 Theoretical background

In MD simulations, the system is evolved with changing positions of atoms and energies. For a system at a constant temperature and pressure, the free energy can be calculated using the energies from the MD simulation through the following equation

$$A = -\frac{1}{\beta}\ln[\sum_i exp(-\beta H_i)] \qquad \text{Eq. 5}$$

where $\beta=1/k_BT$, k_B is the Boltzmann constant, T the temperature and H_i the Hamiltonian of each configuration of the system.

Free energy perturbation (FEP) is a method to calculate the free energy difference between two states [22]. Perturbing from state a to state b, the free energy difference is expressed by

$$\Delta A_{a\to b} = -\frac{1}{\beta}\ln\langle exp\{-\beta[H_b(x,p_x) - H_a(x,p_x)]\}\rangle_a \quad \text{Eq. 6}$$

Here, $H_a(x,p_x)$ and $H_b(x,p_x)$ are the Hamiltonians characteristic of states a and b, respectively. $\langle ... \rangle_a$ denotes an ensemble average over configurations representative of the initial, reference state, a. Convergence of Equation 6 implies that low-energy configurations of the target state, b, are also configurations of the reference state, a, thus resulting in an appropriate overlap of the corresponding ensembles. In practice, transformation between the two thermodynamic states is replaced by a series of transformations between non-physical, intermediate states along a well-delineated pathway that connects a to b. This pathway is characterized by a general extent parameter, often referred to as ``coupling parameter", λ, that makes the Hamiltonian and, hence, the free energy, a continuous function of this parameter between a and b:

$$\Delta A_{a\to b} = -\frac{1}{\beta}\sum_{i=1}^{N}\ln\langle exp\{-\beta[H(x,p_x;\lambda_{i+1}) - H(x,p_x;\lambda_i)]\}\rangle_i \quad \text{Eq. 7}$$

Here, *N* stands for the number of intermediate stages, or ``windows" between the initial and the final states. In my calculations, I have used 22 windows. And each window spanned over 500ps (step size=1fs) with an equilibration for the first 50ps. The energies were saved every 150fs for ensemble averaging at the end of each window.

1.2.3.2 <u>The dual-topology paradigm</u>

In a typical FEP setup involving the transformation of one chemical species into an alternate one in the course of the simulation, the atoms in the molecular topology can be

classified into three groups, (i) a group of atoms that do not change during the simulation -- *e.g.* the environment, (ii) the atoms describing the reference state, a, of the system, and (iii) the atoms that correspond to the target state, b, at the end of the alchemical transformation. The atoms representative of state a should never interact with those of state b throughout the MD simulation. Such a setup, in which atoms of both the initial and the final states of the system are present in the molecular topology file, is characteristic of the so-called ``dual topology'' paradigm [23]. The hybrid Hamiltonian of the system, which is a function of the general extent parameter, λ, that connects smoothly state a to state b , is calculated as a linear combination of the corresponding Hamiltonians:

$$H(x, p_x; \lambda) = H_0(x, p_x) + \lambda H_b(x, p_x) + (1 - \lambda) H_a(x, p_x) \quad \text{Eq. 8}$$

where $H_a(x, p_x)$ describes the interaction of the group of atoms representative of the reference state, a, with the rest of the system. $H_b(x, p_x)$ characterizes the interaction of the target topology, b, with the rest of the system. $H_0(x, p_x)$ is the Hamiltonian describing those atoms that do not undergo any transformation during the MD simulation.

During my simulations, in the case of the perturbation from GTP to UTP, only the base is perturbed while the ribose and the triphosphate groups remain the same. The perturbation was performed similarly for the case of GTP to 2'-dGTP.

1.2.3.3 Special treatments

In order to avoid the so-called ``end-point catastrophes'', it is crucial to avoid situations where growing particles overlap with existing particles with an unbounded interaction potential, which would approach infinity as the interaction distance approaches zero [24]. One possible route for avoiding overlap of unbounded electrostatic potentials consists of allowing a bounded (soft-core) vdW potential to repel first all overlapping particles at low values of λ. As λ

increases, once the particles are repelled, it becomes safe to turn on FEP electrostatics. For the soft-core vdW potential in my simulations, the radius-shifting coefficient was set to 5.

During my FEP calculations, electrostatic interactions of the annihilated particles were linearly decoupled from the simulation between $\lambda = 0$ and $\lambda = 0.5$, and electrostatic interactions of the appearing particles were decoupled from the simulation between $\lambda = 0.5$ and $\lambda = 1$. Van der Waals interactions of the annihilated particles were linearly decoupled from the simulation as λ increases from 0 to 1while vdW interactions of the appearing particles were coupled to the simulation as λ increases from 0 to 1.

1.2.4 Free energies: Umbrella sampling

While free energies can be obtained for systems of different compositions as described above, they can also be calculated for different states of the same system. In most cases, these states we are interested in are results of rare transitions, and therefore, not accessible by conventional MD methods. Biasing potentials are often required to sample high-energy states of the system. When a biasing potential **V** is applied to a system along the reaction coordinate **z**, the new probability distribution **P** becomes

$$\mathbf{P}_b(z) \propto \mathbf{P}(z) \exp[-\beta \mathbf{V}(z)] ,$$

where $\mathbf{P}_b$ is the probability distribution of the biased system at **z** and $\beta = 1/k_B T$, k_B is the Boltzmann constant, T the temperature.

Taking the logarithm of this relationship,

$$\mathbf{A}(z) = -\frac{1}{\beta} \ln[\mathbf{P}(z)] - \mathbf{V}(z) + \text{const} , \qquad \text{Eq. 9}$$

where **A** is the free energy and the last term is a constant.

When the biasing potential takes the form of a harmonic potential, it is then called umbrella sampling [25]. In my work, the coefficient of the harmonic potential was set to 10kcal/Å^2. There were 32 and 22 windows along the reaction coordinate for the 2 umbrella sampling calculations. The starting structure of each window was selected from MD simulations (details in Chapter 3) and the umbrella sampling for each window was then run in parallel. The neighboring windows were adequately overlapped. In both umbrella sampling calculations, each window spanned over 2ns with a step size of 1fs. This included an equilibration of 100ps in the beginning. The reaction coordinate value was collected every 0.1ps.

After data collection, a weighted-histogram analysis method (WHAM) [26] was used to post-process the results from umbrella sampling. The version of WHAM used in this work was a stand-alone code implemented by Dr. Grossfield [27]. The convergence tolerance of the WHAM analysis was 0.001kcal/mol.

1.2.5 QM/MM methods

We used QM/MM methods to calculate the reaction pathway of the nucleotidyl transfer reaction catalyzed by RNAP II. A comprehensive review of QM/MM techniques is presented in Chapter 4. Methods pertinent to our work are summarized here.

The molecular mechanical (MM) description of the system is by the CHARMM force field [15]. The quantum mechanical (QM) part of the system was described by density functional theory (DFT) and by a re-parametrized Austin model 1 (AM1). DFT is the state-of-the-art computational tool for biomolecular systems. Its formalism is

$$\left[-\frac{h^2}{4\pi^2 m_0}\nabla^2 - e v_{nuc}(\vec{r}) + \frac{e^2}{4\pi\varepsilon_0}\int \frac{n(\vec{r}')}{|\vec{r}-\vec{r}'|} d^3r' + v_{xc}(\vec{r})\right]\Phi_i(\vec{r}) = \varepsilon_i \Phi_i(\vec{r}) \qquad \text{Eq. 10}$$

where the electron density $n(\vec{r}) = \sum_i |\Phi_i(\vec{r})|^2$, $\Phi_i(\vec{r})$ is the one-particle Kohn-Sham orbital, $v_{nuc}(\vec{r})$ is the potential of the nuclei and $v_{xc}(\vec{r})$ the exchange-correlation potential. The exchange-correlation potential is based on a pre-parametrized functional and hence the name density functional theory. Functionals based on generalized gradient approximations are widely used for DFT calculations. In my work, the PBE functional [28] has been used for most of my DFT calculations. Part of our benchmark calculations were performed with the B3LYP functional [29]. The basis set we used was DZVP.

The AM1 method is based on the modified neglect of differential diatomic overlap approximation (MNDO). In MNDO [30], the Hamiltonian of the system consists of the electronic energy and the nuclear repulsion energy. Electrons of an atom are divided into core electrons and valence electrons. Core electrons are combined with the nuclear charge and represented as the reduced nuclear charge Z'. Since MNDO only treats main group elements, valence electrons are only composed of *s* and *p* orbitals. Three types of interactions are considered in MNDO: core-electron, electron-electron and core-core interaction. The core-electron Hamiltonian

$$h_{\mu\nu} = \langle \mu_A | \boldsymbol{h} | \nu_B \rangle = h_{\mu\nu} U_\mu - \sum_{a \neq A}^{N_{nuclei}} Z'_a \langle \mu_A \mu_A | \nu_A \nu_A \rangle, \qquad \text{Eq. 11}$$

where μ and ν are valence orbitals, a runs over all atoms, A and B are different atoms, δ is a delta function and Z' is the reduced nuclear charge. U_μ takes the form of

$$U_\mu = \langle \mu_A \left| -\frac{1}{2}\nabla^2 - V_A \right| \nu_A \rangle \qquad \text{Eq. 12}$$

where $-\frac{1}{2}\nabla^2$ is the kinetic energy operator and V_A is the potential by the nucleus.

The electron-electron interaction energy is written in Eq. 13 as a product of the corresponding overlap integral multiplied by the average of two atomic "resonance" parameters, β. The overlap $S_{\mu\nu}$ is calculated explicitly.

$$\langle \mu_A | \boldsymbol{h} | \nu_B \rangle = \frac{1}{2} S_{\mu\nu} \left(\beta_\mu + \beta_\nu \right) \qquad \text{Eq. 13}$$

The integral in the second term of Eq. 11 are parametrized within a *sp*-basis set:

$$\langle ss|ss\rangle = G_{ss}$$

$$\langle sp|sp\rangle = G_{sp}$$

$$\langle ss|pp\rangle = H_{sp}$$

$$\langle pp|pp\rangle = G_{pp}$$

$$\langle pp'|pp'\rangle = G_{p2}$$

where the G-type parameters are Coulomb terms while the H parameter is an exchange integral. The G_{p2} integral involves two different types of *p* functions.

The core-core interaction should be $Z'_A Z_B' / R_{AB}$. However, due to the inherent approximations in the method, this term is not cancelled by electron-electron terms at long distances, resulting in a net repulsion between uncharged molecules or atoms. To overcome this artefact, MNDO calculates the core-core interaction between atoms A and B, $V_{nn}(A, B)$ as

$$V_{nn}^{MNDO}(A,B) = Z'_A Z'_B \langle s_A s_A | s_A s_A \rangle (1 + e^{-\alpha_A R_{AB}} + e^{-\alpha_B R_{AB}}), \qquad \text{Eq. 14}$$

where the α exponents are taken as fitting parameters.

In AM1 [31], the core-core interaction further improved as

$$V_{nn}^{AM1}(A,B) = V_{nn}^{MNDO}(A,B) \frac{Z'_A Z'_B}{R_{AB}} \sum_k \left(a_{kA} e^{-b_{kA}(R_{AB} - c_{kA})^2} + a_{kA} e^{-b_{kA}(R_{AB} - c_{kA})^2} \right), \qquad \text{Eq. 15}$$

where k is between 2 and 4, depending on the atom and a_k, b_k ,c_k are fitting parameters. A major drawback of AM1 is the inadequate description of bond formation involving phosphorus atoms due to the lack of *d* orbitals. Although AM1's core-core interaction is known to perform well for hydrogen bonds, it is nonetheless desirable to diminish it for phosphorus bonding. Therefore, a model needs to be parametrized to balance between the addition of *d* orbitals and the scaling of the core-core interaction.

Recently, a newly re-parametrized AM1 method with *d* orbitals added, AM1/d-PhoT, has been developed for phosphoryl transfer reactions [32]. It has re-fit the parameters of H, O and P atoms and been successfully applied to the phosphoryl transfer reaction by the hairpin ribozyme [33]. Since this reaction is quite similar to the nucleotidyl transfer reaction, AM1/d-PhoT is also suitable for our study of RNAP II. Benchmark calculations also support the application of this method to our system as shown in Figure 1-5. Details of the benchmark calculations are in 6.3.

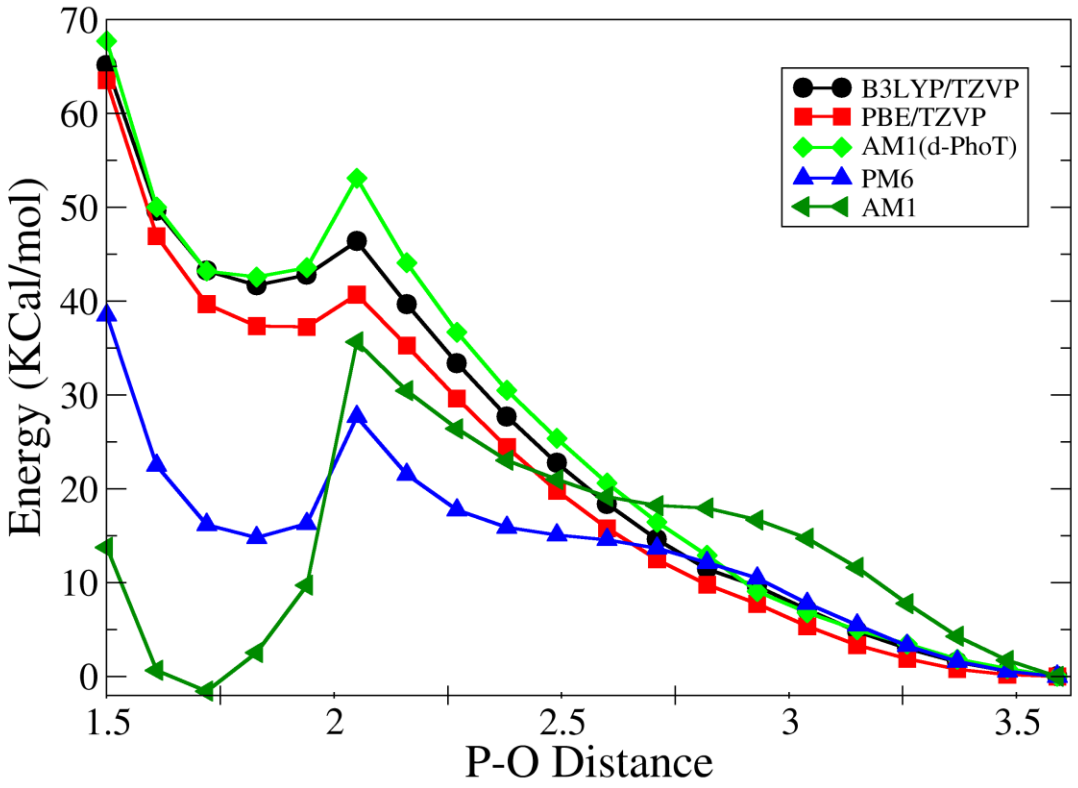


Figure 1-5: Benchmarking of AM1/d-PhoT

1.2.6 Relaxed surface scan

Relaxed surface scan is a method to explore the potential energy surface while scanning along certain reaction coordinates. In this method, a harmonic potential is applied along the reaction coordinates to move the system from the reactant state to the product state. At each step along the reaction coordinates, the rest of the system is allowed to relax by optimization. Therefore, the attained reaction pathway corresponds to a minimum energy pathway. In our work, we used a constraint with k=2500kJ/mol and divided each reaction coordinate into 21

steps. We adopted this method to ensure the transition of the system along certain reaction coordinates so as to compare different possible pathways.

1.3 Main conclusions of the thesis

We build a kinetic model and successfully recover the rate of the nucleotide addition cycle based on empirical reaction parameters. We find that the selection for the matched base is achieved in the binding process while the discrimination of the 2'-OH should be fulfilled in the catalytic reaction. We identify from MD trajectories the important residues in the NTP transfer and binding process. The free energy profile of the cognate NTP transfer from the entry site to the addition site suggests that the trigger loop and the bridge helix actively participate, and that this process is not rate-limiting with the participation of the trigger loop and the bridge helix. The free energy difference between different types of NTPs demonstrates that the cognate NTP is the most stable NTP in the addition site. The 2'-dNTP is slightly less favored, by 1.91 kcal/mol, and the unmatched NTP is the least stable by 16.80 kcal/mol. The MD simulation results indicate that the instability of the 2'-dNTP is due to its twisted ribose, a direct result of the absent 2'-OH, which results in less interaction with surrounding residues. The instability of the unmatched NTP lies in the lack of contacts between its misplaced base and surrounding residues, stemming from mismatching between the base and the DNA template. The results of thermodynamic stability and kinetic transfer calculations suggest that the NTP is mainly discriminated in the addition site. In the case of unmatched NTPs, this is directly the result of thermodynamic instability. 2'-dNTPs, however, are likely discriminated through catalytic inefficiency.

With respect to methodology development, we first present a comprehensive review of the QM/MM methodology as a backdrop for our own QM/MM implementation. We then describe an implementation of a QM/MM interface between deMon2k and CHARMM. Finally,

we build three models to correct the defects in the RNAP II crystal structures and conduct MD simulations and QM/MM relaxed surface scans on each of them. Regarding the difference in the 3'O-Mg coordination among the three models, our results show that this coordination is not required for the reaction to proceed as it is evidently broken or weak in most of the scans that produce low energy barriers. Therefore, the role of Mg(A) in RNAP II appears to be more structural than catalytic. Regarding the nucleotidyl transfer reaction by RNAP II, the results show that the 3'-H is transferred to the α-phosphate either directly or indirectly, facilitating the formation of the 3'O-P_α bond and the weakening of the P_α-$O_{\alpha\beta}$ bond. Following this, the 3'-H migrates to the $O_{\alpha\beta}$, resulting in the pyrophosphate leaving. The quintessential part of this mechanism is that the proton so efficiently mediates among different parties engaged in the reaction to facilitate the P-O bond forming and breaking. Although the acceptor of the initial proton transfer may vary depending on the particular conformation of the active site, all possible routes converge to the same destination.

1.4 Bibliography

1. 2013. *http://en.wikipedia.org/wiki/RNA_polymerase_II* Wikipedia
2. Cheung, ACM, Cramer, P. 2012. *Cell* 149: 1431
3. Cramer, P, Armache, KJ, Baumli, S, Benkert, S, Brueckner, E, Buchen, C, Damsma, GE, Dengl, S, Geiger, SR, Jaslak, AJ, Jawhari, A, Jennebach, S, Kamenski, T, Kettenberger, H, Kuhn, CD, Lehmann, E, Leike, K, Sydow, JE, Vannini, A. 2008. *Annu Rev Biophys* 37: 337
4. Kornberg, RD. 2007. *P Natl Acad Sci USA* 104: 12955
5. Wang, D, Bushnell, DA, Westover, KD, Kaplan, CD, Kornberg, RD. 2006. *Cell* 127: 941
6. Yuzenkova, Y, Bochkareva, A, Tadigotla, VR, Roghanian, M, Zorov, S, Severinov, K, Zenkin, N. 2010. *Bmc Biol* 8
7. Steitz, TA. 1998. *Nature* 391: 231
8. Zhu, R, Ribeiro, AS, Salahub, D, Kauffman, SA. 2007. *J Theor Biol* 246: 725
9. Phillips, JC, Braun, R, Wang, W, Gumbart, J, Tajkhorshid, E, Villa, E, Chipot, C, Skeel, RD, Kale, L, Schulten, K. 2005. *J Comput Chem* 26: 1781

10. Lev, B, Zhang, R, De la Lande, A, Salahub, D, Noskov, SY. 2010. *J Comput Chem* 31: 1015
11. Field, MJ. 2008. *J Chem Theory Comput* 4: 1151
12. Gillespie, DT. 1977. *J Phys Chem-Us* 81: 2340
13. Gillespie, DT. 2007. *Annu Rev Phys Chem* 58: 35
14. M. Bhandarkar, AB, E. Bohm, R. Brunner, F. Buelens, C. Chipot, A. Dalke, S. Dixit, G. Fiorin, P. Freddolino, P. Grayson, J. Gullingsrud, A. Gursoy, D. Hardy, C. Harrison, J. Hénin, W. Humphrey, D. Hurwitz, N. Krawetz, S. Kumar, D. Kunzman, J. Lai, C. Lee, R. McGreevy, C. Mei, M. Nelson, J. Phillips, O. Sarood, A. Shinozaki, D. Tanner, D. Wells, G. Zheng, F. Zhu. 2012.
15. MacKerell, AD, Bashford, D, Bellott, M, Dunbrack, RL, Evanseck, JD, Field, MJ, Fischer, S, Gao, J, Guo, H, Ha, S, Joseph-McCarthy, D, Kuchnir, L, Kuczera, K, Lau, FTK, Mattos, C, Michnick, S, Ngo, T, Nguyen, DT, Prodhom, B, Reiher, WE, Roux, B, Schlenkrich, M, Smith, JC, Stote, R, Straub, J, Watanabe, M, Wiorkiewicz-Kuczera, J, Yin, D, Karplus, M. 1998. *J Phys Chem B* 102: 3586
16. Darden, T, York, D, Pedersen, L. 1993. *J Chem Phys* 98: 10089
17. Essmann, U, Perera, L, Berkowitz, ML, Darden, T, Lee, H, Pedersen, LG. 1995. *J Chem Phys* 103: 8577
18. Verlet, L. 1967. *Phys Rev* 159: 98
19. Brunger, A, Brooks, CL, Karplus, M. 1984. *Chem Phys Lett* 105: 495
20. Martyna, GJ, Tobias, DJ, Klein, ML. 1994. *J Chem Phys* 101: 4177
21. Feller, SE, Zhang, YH, Pastor, RW, Brooks, BR. 1995. *J Chem Phys* 103: 4613
22. Chipot, C, Pearlman, DA. 2002. *Mol Simulat* 28: 1
23. Pearlman, DA. 1994. *J Phys Chem-Us* 98: 1487
24. Beutler, TC, Mark, AE, Vanschaik, RC, Gerber, PR, Vangunsteren, WF. 1994. *Chem Phys Lett* 222: 529
25. Torrie, GM, Valleau, JP. 1977. *J Chem Phys* 66: 1402
26. Kumar, S, Bouzida, D, Swendsen, RH, Kollman, PA, Rosenberg, JM. 1992. *J Comput Chem* 13: 1011
27. Grossfield, A. *http://membrane.urmc.rochester.edu/content/wham/*, *version 2.0.6*
28. Perdew, JP, Burke, K, Ernzerhof, M. 1996. *Phys Rev Lett* 77: 3865
29. Becke, AD. 1993. *J Chem Phys* 98: 1372
30. Dewar, MJS, Thiel, W. 1977. *J Am Chem Soc* 99: 4899
31. Dewar, MJS, Zoebisch, EG, Healy, EF, Stewart, JJP. 1985. *J Am Chem Soc* 107: 3902
32. Nam, K, Cui, Q, Gao, JL, York, DM. 2007. *J Chem Theory Comput* 3: 486
33. Nam, KH, Gaot, JL, York, DM. 2008. *J Am Chem Soc* 130: 4680

CHAPTER TWO: EXPLORING THE MOLECULAR ORIGIN OF THE HIGH SELECTIVITY OF MULTISUBUNIT RNA POLYMERASE II BY STOCHASTIC KINETIC MODELS

2.1 Abstract

RNA polymerases are molecular machines of great fidelity, which can recognize matched NTPs from unmatched NTPs and 2'-dNTPs. We investigated by a stochastic simulation algorithm the whole nucleotide addition cycle based on an event-driven model. This model allows us to examine possible molecular origins of the high fidelity of RNA polymerases. For unmatched NTP selectivity, the conclusions drawn from simulated elongation rates corroborate those derived from structural analysis. The presence of two conformations (E site and pre-insertion site) for the incoming nucleotide before the polymerization reaction is sufficient to allow selectivity. Concerning sugar selectivity, our results indicate that selectivity is only achievable if slow chemical reactions occur for 2'-dNTP. These results can be used to understand recent experimental observations.

2.2 Introduction

One critical process in gene expression is the transcription step during which an RNA transcript is produced from a DNA strand. The molecular machinery of transcription is extraordinarily complex and involves a large number of proteins that all contribute to the regulation of gene expression. At the heart of this machinery, RNA polymerases (RNAPs) are recruited to catalyze the polymerization of messenger RNAs (mRNAs) from nucleoside triphosphate molecules (NTP) and the DNA template. One of the remarkable features of RNAP is its high degree of selectivity. This selectivity plays both on the base pairing (T-A and G-C) and on the sugar (NTP *vs*. 2'-dNTP). The exact molecular origin of such high selectivity is however still a matter of discussion. It is generally admitted that one elongation cycle of the

RNA transcript involves several sub-processes, each of them being subject to a stochastic kinetic behavior. The problem is thus to understand how the overall selectivity of the transcription process emerges from the microscopic (molecular or atomic) characteristics of RNA polymerase given the multitude of chemical or physical-chemical events. To address this question, we have elaborated during the last years a multiscale methodology relying on tools of quantum chemistry (to get atomic level properties) and on stochastic simulations. In this paper we follow our initial investigations and integrate into our stochastic kinetic scheme some recent biochemical evidence, especially concerning the role of molecular fragments at the active center of RNA Polymerase II (RNA Pol II).

Recent extensive genetic, crystal X-ray and biochemical studies demonstrate several possible pictures of the high-substrate-selectivity mechanism by multisubunit RNAPs. Multisubunit RNAPs exist both in eukaryotes and prokaryotes, being closely related to each other [1]. Genetic data show that eukaryotic RNA Pol II (yeast) and prokaryotic RNAPs (*E. coli* and *T. Thermophilius*) have a high degree of genetic conservation, particularly within their active sites. Their crystal X-ray data further highlight their structural conservation. They have for instance a common NADFDGD motif within the active site [2, 3] with an embedded triad of aspartic acid residues. This triad always holds an Mg2+ ion which is proposed to assist in the formation of a nucleophile in the catalysis (shown in Fig. 2-1) [4]. Recently particular attention has been devoted to an element situated close to the entry site of the NTP and called the Trigger Loop. [5, 6] It can change its loop conformation to a helix conformation and seems to play the role of a door, closing and opening or eventually assisting the entry of the active site for the substrate, and hence probably playing an important role in NTP selectivity.

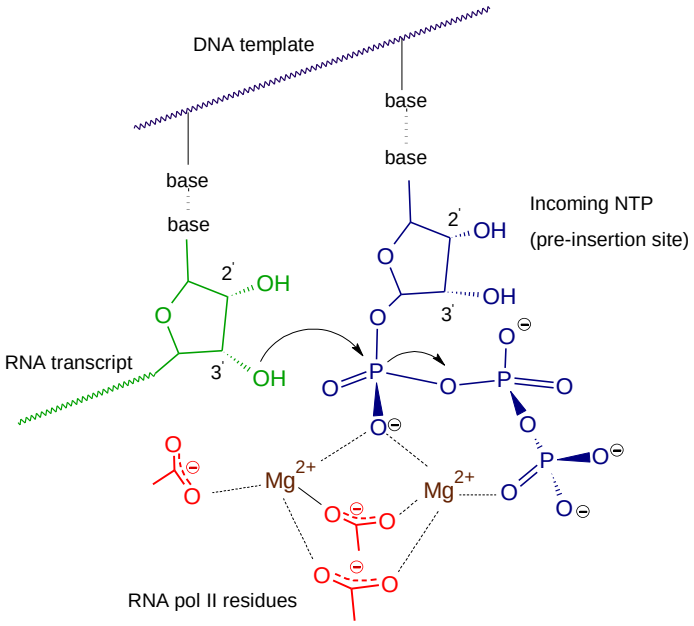


Figure 2-1: The two steps of the polymerization reaction catalyzed by RNAPol II

The last RNA transcript is represented in green, the incoming NTP in blue. The deprotonated aspartic triad that holds the magnesium cations (in brown) are represented in red.

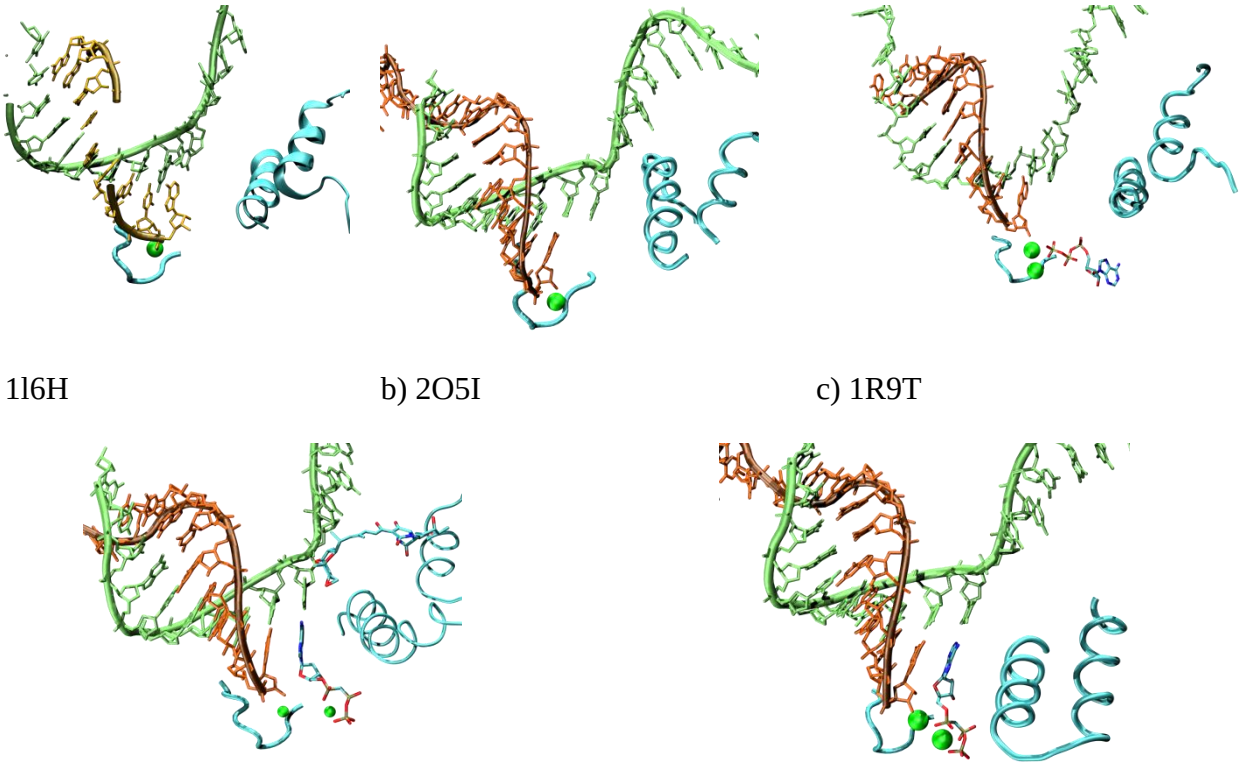
The NTP selectivity may originate from thermodynamic (stabilization or destabilization) or kinetic ("catalysis") considerations. Of course both aspects could play complementary roles. Substrate binding could be the process that mainly determines substrate selection since the base pairing/packing interactions and the residues' − 2'-OH interactions are all heavily involved in this process. Biochemical studies of mutants did indicate that the RNAPs' high substrate selectivity may partly come from residues that make contacts with the 2'-OH/3'-OH groups of the incoming NTP. One proposed residue is the Asn residue in the NADFDGD motif [7]. All those elements argue in favor of a thermodynamic control of NTP selectivity. However, another recent biochemical study [6] showed that the principle defect of elongation caused by mutations in E. Coli's TL was in catalysis rather than in substrate binding, implying that pyrophosphate release may be the process that most determines substrate selection. Very recently, two mutation

studies showed that the TL in yeast also functioned strongly in substrate selection. One study showed that one residue, His, which actually exists in both eukaryotic and prokaryotic TLs, can increase incorporation rates for correct substrates much more than those for incorrect substrates [8]. The other study indicated that the long closing period of the TL improved incorporation efficiency for incorrect substrates much more than for correct substrates [9]. Thus, the TL also seems to control the fidelity of transcription elongation. This accumulation of experimental evidence thus suggests that NTP selectivity is more controlled by kinetics than by thermodynamics.

Our goal is to integrate the previous biochemical evidence into a reasonable kinetic model of the polymerization process. One key feature of the transcription elongation process is that it involves a series of sub-processes, such as RNAP moving back and forth along the DNA template, the substrate diffusing to an entry site, the substrate binding to a pre-insertion site, the TL closing and opening the active site, phosphodiester bond formation, and pyrophosphate release. Notice that some of them might be concerted in the real system, for instance the rotation of the nucleotide and the closing of the TL. At a single enzyme level, all these events occur in a stochastic manner and the methodology used should take account of this aspect. We thus built a stochastic kinetic model of the elongation cycle based on experimentally observed/inferred events. Using this model, we obtained kinetic features which may give insights into previous biochemical experiments regarding substrate selection. Our major finding is that only differences in substrate binding affinity cannot explain the high fidelity of RNAPs. Differences in the chemical reaction rate for different substrates appear to play a dominant role.

2.3 Models and methods

Following the previous discussion and based on several pieces of biochemical evidence we built the following elongation catalytic cycle which encompasses five steps. This kinetic scheme is mainly based on available X-Ray structures (Fig. 2-2) to determine the nature of the chemical intermediates. When relevant, in view of experimental data, the reverse reactions are also considered. We finally remark that we merged together the three sub-steps of the chemical catalytic reaction (deprotonation of the 3'OH group, phosphodiester bond formation, and pyrophosphate release) into one step (eq. 9).



a) 1l6H b) 2O5I c) 1R9T

d) 2PPB e) 2O5J

Figure 2-2: Five molecular events caught by X-ray crystallography involved in a putative elongation cycle

(a) catalytic reaction (PDB: 1I6H), (b) RNAP moving forward to the next base along the DNA template (PDB: 2O5I), (c) substrate diffusing to an entry site (PDB: 1R9T), (d) substrate binding to a pre-insertion site (PDB: 2PPB), and (e) TL closing the active site

(PDB: 2O5J). The loop with a triad of aspartate residues is denoted in blue tube, the TL in blue ribbon, DNA template in green tube, RNA transcript in light red tube, and Mg^{2+} ions in green spheres. Base pairs are also shown between the DNA template and the RNA transcript. Note that the DNA template does not show well in (b) and that the TLs in (a-d) are not completely shown in the crystal structures.

RNAP moving forward to the next base along the DNA template [10, 11, 12, 13]

$$\mathrm{RNAPpdt} \xrightarrow{k_1} \mathrm{RNAPfwd} \quad (1)$$

$$\mathrm{RNAPfwd} \xrightarrow{k_2} \mathrm{RNAPpdt} \quad (2)$$

Substrate diffusing to an entry site [14]

$$\mathrm{NTPfree} + \mathrm{Esite} \xrightarrow{k_3} \mathrm{NTPesite} \quad (3)$$

Substrate binding to a pre-insertion site [6]

$$\mathrm{NTPesite} + \mathrm{RNAPfwd} \xrightarrow{k_5} \mathrm{NTPinsert} \quad (5)$$

$$\mathrm{NTPinsert} \xrightarrow{k_6} \mathrm{NTPesite} + \mathrm{RNAPfwd} \quad (6)$$

TL closing the active site [5, 6]

$$\mathrm{NTPinsert} \xrightarrow{k_7} \mathrm{NTPasite} \quad (7)$$

$$\mathrm{NTPasite} \xrightarrow{k_8} \mathrm{NTPinsert} \quad (8)$$

Catalytic reaction [15]

$$\mathrm{NTPasite} \xrightarrow{k_9} \mathrm{RNAPpdt} + \mathrm{Esite}. \quad (9)$$

RNAPpdt and RNAPfwd mean that RNAP reaches the product (Fig. 2-2a) and forward (Fig. 2-2b) states, respectively. Note that RNAPpdt corresponds to the pre-translocational state and RNAPfwd to the post-translocational state. NTPfree, NTPesite, NTPinsert and NTPasite are four species of NTP: in free diffusion, at the entry site, at the pre-insertion site, and at the addition

site. Esite denotes the state of the entry site. k_1 to k_9 are probability rate constants for these nine molecular events.

The above equations actually give evolutionary rules for the molecular species. These probability rate constants specify how often the stochastic events occur. Given initial molecular states, using a Monte Carlo method called the Gillespie algorithm or the Kinetic Monte Carlo method [16] we can perform a stochastic simulation. In the simulation, the algorithm uses two random numbers to determine at what time which event occurs next, which can be used to determine the next states of the system by following the evolutionary rule corresponding to the event. The algorithm can repeat this process to produce a temporal evolution of molecular species. This temporal evolution is a possible realization of the underlying stochastic process for the model. More details about the algorithm can be found in Ref. [16].

The initial conditions for the model we are using in this study are: NTPfree = 1000, RNAPpdt = 1, RNAPfwd = 0, NTPesite = 0, NTPinsert = 0, NTPasite = 0, and Esite = 1. That is, the number of NTPs in free diffusion is 1000, RNAP is in the pre-translocational state, and no substrates occupy any site. Note that Esite = 1 denotes that the entry site is available. These settings can be used to mimic the situation that the elongation just gets started or that the elongation recovers from a pause.

Estimations of the individual rate constants ($k_{1\text{-}9}$) of the various steps were determined in view of the following considerations. The average rate of RNA synthesis by *E. coli* in vivo is about 50 nucleotides per s. Therefore, any reaction rate involved in the elongation cycle should be larger than 50 s^{-1}. For the back-and-forth motion of RNAP (k_1 and k_2), 60-70 percent of the transcription elongation complexes are known to stay in the pre-translocational state when the elongation complex is stalled due to a lack of matched substrates [9]. To match this ratio k_1 was

set to $100s^{-1}$ (RNAP moving from pre- to post-translocation) and k_2 to 200 s^{-1} (reverse process). Concerning the diffusion of the incoming NTP to the entry site (Esite) we used a theoretical evaluation of 200 s^{-1} [17], leading to $k_3 = 0.2$ s^{-1} ($1000 \times k_3 = 200 s^{-1}$).The reasonable rate for substrates to leave the entry site should be between 50 and 200 s^{-1}. We thus set $k_4 = 100 s^{-1}$ (substrate leaving E_{site}). The following rate constants (k_5 to k_9) are harder to evaluate as they are associated with a large number of atoms and with complex physical-chemical events. The determination of adequate reaction coordinates, activation energies and finally estimates of their rate constants are challenging. In the present study, we bypass these modeling difficulties for the moment and aim at determining the weights of the individual chemical events in the selectivity of the nucleotide by performing a sensitivity study. Three physical-chemical or chemical steps are analyzed, the rotation (in and out) of the nucleotide within the active site (k_5 and k_6 respectively), the closure of the TL (k_7 and k_8) and the polymerization reaction (k_9). In a first kinetic model (model 1) we explicitly hypothesize that steps 5-6 and 7-8 are uncoupled. In fact, it is hard to confirm this hypothesis from experimental data. Consequently we also consider the possibility of a kinetic scheme where the rotation of the nucleotide and the closure of the trigger loop are concerted (model 2).

Biochemical data (X-Ray) indicate that the nature of the incoming substrate (correct, unmatched or 2'deoxy) has to be taken into account in the rotation step. In our kinetic model, this is done through the respective values of k_5 and k_6 the ratio of which defines the equilibrium constant K_{rot} of the rotation step. For a regular elongation step, no matched NTP are observed in the E_{site}, but only in the pre-insertion site, suggesting that the rotation equilibrium is totally displaced toward the pre-insertion site. Based on this experimental evidence, we have assigned a

value of 100 to K_{rot}. On the other hand, for unmatched NTP, all the nucleotides are observed in the E_{site}, and not in the pre-insertion site, leading in our protocol to a value of 0.01 for K_{rot}. Finally, the 2'-deoxy NTP is an interesting intermediate case with a value of 0.4, reflecting an equilibrium constant allowing the nucleotide to be substantially stabilized in both sites. This ratio was derived from X-Ray experiments showing the presence of 2'-deoxy NTP in both the E_{site} and the pre-insertion site.

2.4 Results

As described in the previous part, stochastic simulations were performed for various sets of rate constants to get averaged values of the polymerization rate. Specifically, 100 simulations were carried out for each averaged value. In the following tables, we also report the ratio of the elongation rates (denoted by E) between the correct nucleotide and the unmatched or 2'-deoxy nucleotides. These values are discussed in the next paragraphs and will be compared to the available experimental values. For the unmatched nucleotides, various kinetic experiments indicate a ratio E_{unmat}/E_{mat} of about 1.25×10^{-4} [8]. Concerning the sugar selectivity, the sensitivity seems to be less pronounced with a reported ratio $E_{2'deoxy}/E_{mat}$ of about 2.1×10^{-3} [8].

We first examine the hypothesis that there is no coupling between the $k_{5/6}$ and $k_{7/8}$ reactions (model 1). Thus, all the reactions listed above were used. Table 2-1 summarizes the various elongation rates for various $k_{5/6}$ keeping $k_{7/8}$ and k_9 constant. The obtained elongation rates indicate different selectivity depending on the nucleotide that is considered. For unmatched NTP, a ratio of 0.02 is found for the highest values of k_5 and k_6 (fast rotation), the selectivity being partly lost for slow rotation. For the deoxy case, the selectivity is significantly lower (ratio around 0.5), and is nearly lost for a slow rotation process (ratio=0.9).

Table 2-1: Elongation rates of different NTPs

The rates are in nucleotides per second ($nt \cdot s^{-1}$) with fixed k_7 ($500\ s^{-1}$), k_8($1000\ s^{-1}$), and k_9($1000 s^{-1}$)

	Matched NTP		Unmatched NTP			2'-deoxy NTP		
	K_{rot} = 100		K_{rot} = 0.01			K_{rot} = 0.4		
k_5 (s^{-1})	k_6 (s^{-1})	E_{mat} ($nt \cdot s^{-1}$)	k_6 (s^{-1})	E_{unm} ($nt \cdot s^{-1}$)	E_{unm}/ E_{mat}	k_6 (s^{-1})	$E_{2'deoxy}$ ($nt \cdot s^{-1}$)	$E_{2'deoxy}/ E_{mat}$
500	5	48.16	50 000	1.14	0.024	1250	23.54	0.489
400	4	45.81	40 000	1.10	0.024	1000	22.88	0.499
300	3	42.37	30 000	1.09	0.026	750	21.83	0.515
100	1	26.17	10 000	1.12	0.043	250	16.40	0.627
10	0.1	4.32	100	0.97	0.225	25	3.95	0.914

In comparison with the experimental data those values are not sufficiently small to discriminate different nucleotides and, hence, to account for RNA Pol II's selectivity. Actually, the better selectivity observed for the unmatched NTP case is obviously related to the fact that the incoming NTP is not stabilized in the pre-insertion site. Few monomer molecules have access on average to the pre-insertion site. This point is related to the rotation equilibrium constant $K_{rot}=k_5/k_6$. Nevertheless, the value proposed in the present study (0.01) has only to be seen as an upper limit for the rotation step, a value that accounts for the non-observation of unmatched NTP in the pre-insertion site. Lower values for K_{rot} might not be excluded from the present computational modeling. For instance with $K_{rot}=10^{-3}$, a selective ratio of 0.00036 is found. The proper evaluation of this equilibrium constant with an accurate computational method is

currently under way in our group. Finally, as proposed by Ref. [14], the presence of a rotation step associated with two conformational positions of the nucleotide is probably sufficient to discriminate between matched and unmatched NTPs. At this stage the question of selectivity of the correct sugar remains open.

In the next computational experiment, we set k_5 to 500 s^{-1}, a value which renders the best selectivity, but we decrease the chemical reaction rate (k_9). An improvement of the selectivity is obtained for the slowest rate constants (Table 2-2). The selectivity remains however very far from the experimental data. Moreover, the small gain in selectivity seems to have to be paid for by a serious loss of enzymatic activity since the elongation rate for the correct NTP amounts only to 3.2 nt.s^{-1} ($k_9 = 10s^{-1}$). Finally, even when adjusting the different values for k_7 and k_8, no satisfactory selectivity is obtained for the 2'-deoxy monomer (Table 2-3).

Table 2-2: Influence of the chemical reaction rate constant k_9 on the elongation rates Values obtained with fixed k_5 (500 s^{-1}), k_7 (500 s^{-1}) and k_8 (1000s^{-1})

	Matched NTP	Unmatched NTP		2'-deoxy NTP	
k_9 (s^{-1})	E_{mat} ($nt \cdot s^{-1}$)	E_{unm} ($nt \cdot s^{-1}$)	E_{unm}/ E_{mat}	$E_{2'deoxy}$ ($nt \cdot s^{-1}$)	$E_{2'deoxy}/ E_{mat}$
1000	48.16	1.14	0.024	23.54	0.489
700	45.83	0.96	0.021	20.66	0.451
600	44.39	0.92	0.021	19.40	0.437
500	42.81	0.87	0.020	17.70	0.413
400	40.48	0.69	0.020	15.48	0.383
100	21.36	0.28	0.013	5.85	0.274

10	3.19	0.098	0.031	0.75	0.236
1	0.44	0.07	0.161	0.15	0.337

Table 2-3: Influence of the rate constants ($k_7$and k_8) on the elongation rates k_5 and k_9 are set to 500 s^{-1} and 1000 s^{-1} for all three cases. k_6 is set to 5 s^{-1}, 50000 s^{-1}, and 1250 s^{-1}, respectively, for the matched, unmatched and 2-deoxy NTP.

	Matched NTP	Unmatched NTP		2'-deoxy NTP	
	E_{mat} ($nt \cdot s^{-1}$)	E_{unm} ($nt \cdot s^{-1}$)	$E_{unm/}$ E_{mat}	$E_{2'deoxy}$ ($nt \cdot s^{-1}$)	$E_{2'deoxy/}$ E_{mat}
k_7=50s^{-1} / k_8=100s^{-1}	24.41	0.09	0.0037	6.28	0.2573
k_7=1000s^{-1} / k_8=2000s^{-1}	49.99	0.23	0.0046	27.56	0.5513

We now turn our attention to the second hypothesis which is related to coupling of the rotation step and the TL closing/opening movement (Model 2). In other words we assume that the rotation of the NTP is catalyzed by the closure of the TL. In practice, this is achieved here by deleting the seventh and eighth reactions (k_7 and k_8) and by considering that reactions five and six (k_5 and k_6) are now relevant to a merged concerted reaction. To connect reactions 5 and 6 to reaction 9, we changed NTPinsert in reactions 5 and 6 to be NTPasite. Similar simulations have been performed with the same sets of rate constants. The results are gathered in Table 2-4. Clearly the modified kinetic scheme does not lead to any improvement of the selectivity for any kind of NTP. Obviously, it is even less pronounced than before, particularly for the cases with fixed k_9.

Nevertheless a very interesting element appears for the unmatched NTP case. When compared with the previous kinetic model 1, the set of elongation rates for model 2 implies some loss of selectivity. Assuming the ability of the present kinetic model to describe the main RNA Pol II kinetic features, these results mean that a supplementary step between the rotation of the nucleotide and the chemical reaction seems useful to discard unmatched NTP if other conditions are the same. A comparison of greatest interest listed in Table 2-4 is the last one with fixed k_9, where there are slowest NTP rotation rates. In such a situation, in the absence of the separating step, the rotation of the nature of the sugar ring may be expected as long as similar values of k_9 are taken for all the kinds of nucleotides. However, when one considers the possibility of individual chemical reaction rates for every nucleotide, good selectivity is achievable. For example assigning a value of $k_9 = 10^3\ s^{-1}$ for the matched nucleotide and the relative stabilization of the nucleotide within the E and pre-insertion site are almost of no use for selectivity. In view of the recent literature we have proposed here that this supplementary step might be the closing/opening of a sub-chain called the Trigger Loop. Indeed, yeast RNA Pol II adds nucleotides and translocates slowly and with reduced accuracy in the presence of α-amanitin [8]. The inhibitor, α-amanitin, seems to stop the Trigger Loop from performing the closing [8, 13], which may affect the fidelity of RNAPs [18].

Table 2-4: Elongation rates of different NTP when k_5 and k_9 are varied

	Matched NTP		Unmatched NTP			2'-deoxy NTP		
	$K_{rot} = 100$		$K_{rot} = 0.01$			$K_{rot} = 0.4$		
fixed $k_9 = 1000\ s^{-1}$								
k_5	k_6	E_{mat}	k_6	E_{unm}	E_{unm}/ E_{mat}	k_6	$E_{2'deoxy}$	$E_{2'deoxy}/ E_{mat}$

(s^{-1})	(s^{-1})	($nt \cdot s^{-1}$)	(s^{-1})	($nt \cdot s^{-1}$)		(s^{-1})	($nt \cdot s^{-1}$)	
500	5	57.35	50 000	57.13	1.0	1250	57.39	1.0
400	4	54.30	40 000	54.20	1.0	1000	54.29	1.0
300	3	49.54	30 000	49.47	1.0	750	49.56	1.0
100	1	28.87	10 000	28.99	1.0	250	28.87	1.0
10	0.1	4.30	100	4.31	1.0	25	4.30	1.0
fixed $k_5 = 500\ s^{-1}$								
k_5 (s^{-1})	k_6 (s^{-1})	E_{mat} ($nt \cdot s^{-1}$)	k_6 (s^{-1})	E_{unm} ($nt \cdot s^{-1}$)	E_{unm}/ E_{mat}	k_6 (s^{-1})	$E_{2'deoxy}$ ($nt \cdot s^{-1}$)	$E_{2'deoxy}/ E_{mat}$
1000	5	57.35	1.0	57.13	1.0		57.39	1.0
700	5	56.12	1.0	56.05	1.0		56.10	1.0
600	5	55.53	1.0	55.36	1.0		55.55	1.0
500	5	54.80	1.0	54.57	1.0		54.79	1.0
400	5	53.66	1.0	53.66	1.0		53.68	1.0
100	5	40.01	1.0	40.01	1.0		40.00	1.0
10	5	8.83	1.0	8.81	1.0		8.84	1.0
1	5	1.11	1.0	1.11	1.0		1.11	1.0

Altogether, our results indicate that no selectivity on the nature of the sugar ring can be expected as long as similar values of k_9 are taken for all kinds of nucleotides. However, when

one considers the possibility of individual chemical reaction rates for every nucleotide, good sensitivity is achievable. For example assigning a value of $k_9 = 10^{-3}$ s^{-1} for the matching case and 1.0 s^{-1} for the 2'-deoxy case (Tables 2-1 and 2-2), the experimental selectivity of c.a. 2.0×10^{-3} can be recovered. Applying an Arrhenius law, such a difference for the rates constants of 1000 s^{-1} is equivalent to a difference of 17.1 $kJ.mol^{-1}$ for the activation energies (for T=298K). This value is certainly attainable thanks to a few electrostatic interactions within the catalytic core of the enzyme. Indeed it roughly corresponds to the energy of a single hydrogen bond. This result suggests that selectivity against the nature of the sugar might not be so difficult to reach during the polymerization. This may be used to explain why RNAPs are very sensitive machines in terms of hydrogen-bonded interactions between the NTP and RNAP. Nevertheless it brings a puzzling question about the corresponding molecular basis as X-Ray structures show that the sugar moiety of the incoming NTP is a couple of Angstroms away from the reactive sphere *i.e.* the P_α atom of the monomer and the O3' atoms of the last transcript nucleotide (Fig. 2-1). Hence more mechanistic investigations will have to be carried out to understand how the influence of the O2' atoms (or its absence for a 2'-deoxy NTP) can be dynamically transmitted to the reactive sphere. Such work is in progress in our lab and will be presented in due course.

2.5 Conclusions

The transcriptional elongation cycle is a complicated biochemical process involving many kinds of molecules and molecular events. Following our previous investigations, we have used in this study the Gillespie algorithm to investigate the whole nucleotide addition cycle based on an event-driven stochastic model. Two models were considered in view of the most recent findings from the biochemical literature devoted to these polymerases. The main objective

of the study was to understand how the selectivity of the incoming NTP (matched vs. unmatched and 2'-oxy vs. 2'-deoxy) could emerge from the complex kinetic network involved in the elongation process.

Concerning the unmatched NTP, the computed elongation rates indicate that selection of the nucleotide can be partly achieved thanks to the presence of two possible conformations of the monomer (E_{site} and pre-insertion site). This proposal is not new and had already been made on the basis of X-Ray experiments. Our kinetic study has brought nonetheless a complementary but essential point which is the possible help of a supplementary event between the rotation and the chemical reaction to obtain selectivity. Based on biochemical studies, we have proposed that such an event could be the closing/opening of the active site's entrance by the Trigger Loop. Concerning the 2'-deoxy NTPs, the computational results indicate that no conclusive selectivity should arise only from the relative stability of the nucleotide between the E and pre-insertion sites. We have shown that different rate constants for the chemical reaction depending on the nature of the incoming NTP, on the contrary, should allow selection of 2'-oxy substrates.

Considering this factor, the experimental selectivity of 1.25×10^{-4} for unmatched NTPs can also be easily obtained. This strongly indicates that differences in substrate binding affinity alone cannot contribute to the high selectivity of RNAPs. They must work together with differences in the chemical reaction rate for different kinds of substrates. As stated above, the latter can be easily achieved by the difference of one single hydrogen-bond energy. Recall that the catalytic reaction event consists of three sub-events: deprotonation of the 3’-OH group, phosphodiester bond formation, and pyrophosphate release. Our previous dynamics study [19] showed that pyrophosphate release appears to dominate the kinetics of the catalytic reaction. The

exact role of the polymerase residues on the catalysis like histidine 1085 remains elusive and more work is needed to address this issue.

Our modeling strategy needs further refining in several respects. With respect to the kinetic model, if molecular processes get too complicated, delay stochastic simulation techniques [20] can also be employed to speed up the modeling. In addition, multiscale modeling should be tightly combined with corresponding experiments. Future work will be devoted to obtaining reliable estimates of the examined rate constants. Due to the diversity of the encountered phenomena, a variety of computational tools will have to be used in conjunction with experimental data.

2.6 Bibliography

1. Werner F (2008) Trends in Microbiology 16:247-250
2. Sonntag K-C, Darai G (1996) Virus Genes 11:271-284
3. Zhu R, Janetzko F, Zhang Y, van Duin ACT, Goddard WA, Salahub DR (2008) Theoretical Chemistry Accounts 120:479-489
4. Sosunov V, Zorov S, Sosunova E, Nikolaev A, Zakeyeva I, Bass I, Goldfarb A, Nikiforov V, Severinov K, Mustaev A (2005) Nucleic Acids Res. 33:4202-4211
5. Wang D, Bushnell DA, Westover KD, Kaplan CD, Kornberg RD (2006) Cell 127:941-954
6. Vassylyev DG, Vassylyeva MN, Zhang JW, Palangat M, Artsimovitch I, Landick R (2007) Nature 448:163-U164
7. Svetlov V, Vassylyev DG, Artsimovitch I (2004) Journal of Biological Chemistry 279:38087-38090
8. Kaplan CD, Larsson KM, Kornberg RD (2008) Mol. Cell 30:547-556
9. Kireeva ML, Nedialkov YA, Cremona GH, Purtov YA, Lubkowska L, Malagon F, Burton ZF, Strathern JN, Kashlev M (2008) Mol. Cell 30:557-566
10. Vassylyev DG, Vassylyeva MN, Perederina A, Tahirov TH, Artsimovitch I (2007) Nature 448:157-U153
11. Kettenberger H, Armache KJ, Cramer P (2004) Mol. Cell 16:955-965
12. Abbondanzieri EA, Greenleaf WJ, Shaevitz JW, Landick R, Block SM (2005) Nature 438:460-465
13. Brueckner F, Cramer P (2008) NSMB 15:811-818
14. Westover KD, Bushnell DA, Kornberg RD (2004) Cell 119:481-489
15. Gnatt AL, Cramer P, Fu JH, Bushnell DA, Kornberg RD (2001) Science 292:1876-1882
16. Gillespie DT (1977) J. Phys. Chem. 81:2340-2361

17. Batada NN, Westover KD, Bushnell DA, Levitt M, Kornberg RD (2004) Proc. Natl. Acad. Sci. USA 101:17361-17364
18. Svetlov V, Nudler E (2008) NSMB 15:777-779
19. Zhu R, Salahub DR (2007) AIP Conference Proceedings 963:104-110
20. Zhu R, Ribeiro AS, Salahub D, Kauffman SA (2007) J. Theor. Biol. 246:725-745

CHAPTER THREE: BRIDGE HELIX AND TRIGGER LOOP IN ACTION – HOW RNA POLYMERASE II BINDS AND SELECTS NTPS

3.1 Abstract

RNA polyermase II, a crucial enzyme for gene expression in eukaryotes, synthesizes messenger RNAs with high selectivity. Despite its importance, the mechanism for nucleotide binding and nucleotide discrimination is not well understood. To dissect the origin of high selectivity, we performed molecular dynamics (MD) calculations for cognate and non-cognate Nucleoside Tri-Phosphates (NTPs) in the active site and identified key residues important for stabilizing the cognate NTP. Our free energy perturbation calculations show that mutating a cognate GTP to a non-cognate UTP in the active site costs ~16.8kcal/mol while mutating a cognate GTP to a 2'-deoxyGTP costs ~2kcal/mol. Hence, the selectivity for cognate vs. non-cognate NTPs can be accounted for on a thermodynamic basis. 2'-dNTPs, however, are likely discriminated against through catalytic inefficiency. Since two binding sites exist in the enzyme, we conducted MD to simulate the entry of a cognate GTP from the entry site to the active site. The results demonstrate that two key motifs, the trigger loop and the bridge helix, play important roles in this process. Facilitated by these two motifs, the NTP entry is a spontaneous process with an energy decrease of ~6kcal/mol, as shown by our umbrella sampling calculations.

3.2 Introduction

RNA polymerases (RNAP) are ubiquitous cellular machines essential for converting DNA templates into RNA molecules, an important component of molecular biology's central dogma. RNA polymerase II (RNAP II), which is responsible for synthesizing messenger RNA (i.e. transcription) in eukaryotes, has been the focus of much research in recent years [1-9]. As shown *in vivo*, it is capable of selecting nucleotide triphosphates (NTP) complementary to a

DNA template with great efficacy. Typically, nucleotides (nt) are incorporated into an RNA transcript at a speed of 20-70 nt per second with an error rate of 1 per 10^5 [1, 7]. Underlying such high selectivity is a complex proof-reading mechanism.

Information about the binding mechanism of RNAP II has been revealed in crystal structures of the transcription complex. As shown in the crystal structure with a matched NTP (PDB: 2E2H), the catalytic site is composed of an α-helical structure denoted as the bridge helix (BH), a flexible loop motif termed the trigger loop (TL), and two magnesium ions – Mg(A) and Mg(B) in addition to surrounding protein residues of the RNAP II domains Rpb1 and Rpb2 [10]. This catalytic site (Figure 3-1A) is termed the addition site (A site) where a matched NTP binds the DNA template and is added to the RNA primer. Intriguingly though, another binding site exists, as determined from a protein crystal using a 2'-deoxyGTP [10, 11]. This site (Figure 3-1B) is termed the entry site (E site) which serves as an entrance for the passage leading to the A site. Distinctly, all nucleotides bind to the E site whereas only a nucleotide that is complementary to the template can further bind to the A site. Importantly, both the BH and the TL appear in distinctly different states between the E site and A site (Figure 3-1A and 3-1B). In the A site, the BH bends in the presence of a matched nucleotide, while it is straight upon nucleotide binding at the E site. The TL was found to open the A site when a nucleotide is in the E site and close the A site when a matched nucleotide arrives in the A site.

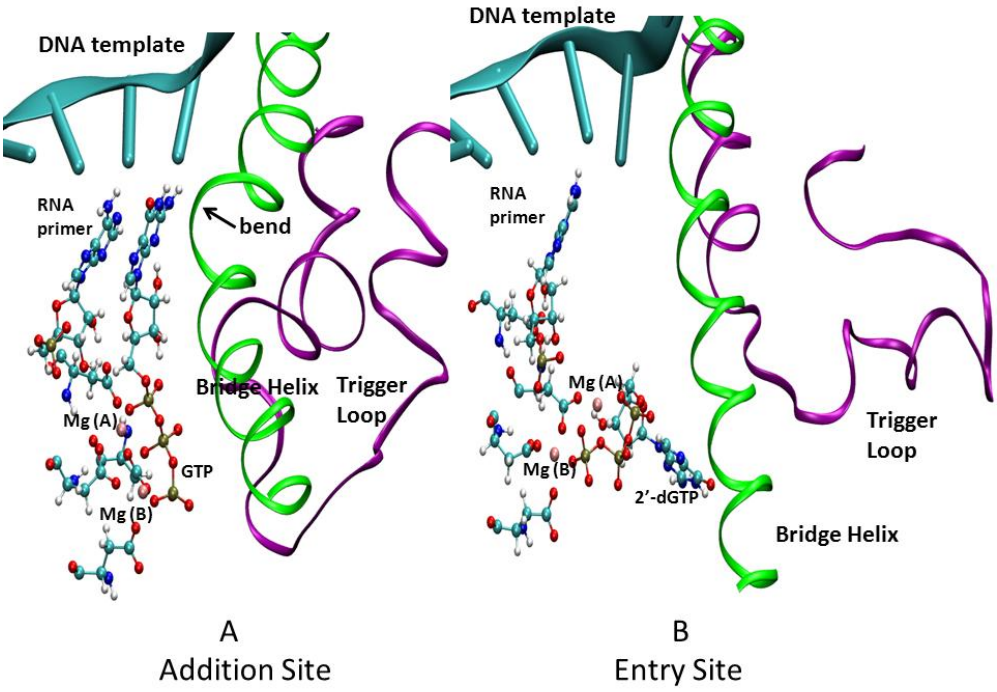


Figure 3-1: Crystal structures of the addition site and the entry site

A) Crystal structure of a GTP in the addition site B) Crystal structure of a 2'-dGTP in the entry site where the trigger loop is colored in purple, the bridge helix in green and Mg ions in pink.

Although the BH and the TL have been proven important through site-directed mutations [7-10], how they help transfer the NTP from the E site to the A site is still in question. Addressing this question requires dynamic information of this system in addition to the static structures determined by crystallography. A suitable technique to study the dynamics is computational simulation which has been adopted by many researchers in their investigations of RNAP II. In short, Kornberg and coworkers [2] demonstrated by course-grained simulation of NTP diffusion, that nucleotide binding to the E site greatly enhances its probability of binding to the A site. Feig and Burton [5] performed normal mode analysis and showed that an open trigger loop facilitates translocation of RNAP II along the DNA strand. Additionally, they conducted

full molecular dynamics (MD) simulation and concluded that fidelity control and catalysis require the trigger loop closing [12, 13] . Huang et al, through their MD simulations, explicitly showed the stabilizing effects of His1085 and Leu1081 of the trigger loop towards matched NTPs in the A site [14]. Although these simulation studies have filled in details in the connection between states of the elongation complex to various extents, one of the key steps, nucleotide transfer from the E site to the A site, still remains obscure. And the function of the BH and the TL in this step also needs to be identified.

Moreover, the selectivity mechanism for the matched NTP over the unmatched and 2'-deoxy NTPs is also unaddressed. Since the unmatched and 2'-deoxy NTPs are not stable in the A site, no crystal structure with them in the A site has been resolved. Although there have been numerous kinetic studies comparing the rates between cognate and non-cognate NTPs [10, 15], the structural basis should be better understood through computational approaches. Comparing the cognate and non-cognate NTPs in the A site on a structural basis could be crucial for illustrating the intricate proof-reading mechanism of RNAP II.

In this paper, we attempt to delineate the selectivity mechanism and identify the function of the BH and the TL during the NTP transfer from the E site to the A site. To this end, we have constructed all-atom models based on crystal structures and their combinations, and we have performed extensive MD simulations of each model. From this, we have identified important residues for selectivity, calculated free energy differences among different NTPs in the A site, and the energy profile of the NTP transfer between the two binding sites.

3.3 Methods

3.3.1 System setup

Models are constructed based on crystal structures of the ternary elongation complex with either a GTP present in the addition site (PDB ID: 2E2H) [10] or a 2'-dGTP present in the entry site (PDB ID: 2E2I) [10]. In both structures, a number of residues were not resolved as a result of structural disorder. In the subunit Rpb1 of 2E2H, missing residues 1446-1733 at the end of the chain were not inserted as they are not important to the core function of RNAP II. Additionally, modeling of large surface loops proves unreliable. Missing non-end residues, 156-160, 186-191, 315-318 and 1232-1235, which are not missing in 2E2I were added by adopting the same psi and phi angles as in 2E2I. Missing non-end residues, 192-198, 1177-1186 and 1244-1253, which are missing in both 2E2H and 2E2I, were inserted by manually entering the psi and phi angles in coordination with the adjacent, known residues. The same protocol was followed for other subunits of 2E2H and 2E2I. After all necessary missing residues were restored, a geometry optimization was performed with non-missing residues constrained. The missing 3'-O atom of the RNA primer was added, based on the topology in the CHARMM 27 force field [16]. The missing second Mg ion in 2E2I, Mg (B), was inserted at the midpoint between OD1 of Rpb2-Asp837and OG1 of the 2'-dGTP as it coordinates with both atoms in 2E2H.

Protonation states of titratable residues were determined by pKa calculation through the GBMV module [17] in CHARMM [18]. In the case of histidine, the site that has the lowest calculated pKa was protonated. Protonation states were held the same in all the models for consistency. Nucleotide triphosphates were deprotonated in all models and therefore carry a charge of -4 [2, 12].

To compare the stability of different nucleotides in the A site, the GTP in 2E2H was replaced by a 2'-dGTP or a UTP. This was done by changing the 2'-hydroxyl group and the base respectively. This was followed by geometry optimization and MD simulation. To understand the stability of GTP in the E site, the 2'-dGTP in the E site was replaced by a GTP by adding the 2'-hydroxyl group. Likewise, this was followed by geometry minimization and MD simulation. To simulate the nucleotide transport from the E site to the A site, 3 starting structures were interpolated by combining 30% of the 2E2I coordinates and 70% of the 2E2H coordinates, 50% each, and 70% 2E2I and 30% 2E2H, respectively. This was followed by geometry optimization and 7ns of regular MD simulation. Interpolation was performed before solvation. These structures are referred to as interpolated structures throughout this paper.

Each model was fully solvated in a cubic box of explicit water with a length of ~160 Å. To neutralize the system, a total of 88 Na^+ ions were added by randomly replacing the water molecules at the surface of the box. As a result, each model comprises a total of ~340,000 atoms.

3.3.2 Simulations

3.3.2.1 Molecular dynamics

The CHARMM 27 force field [16, 19] was used to describe the protein and nucleic acids. Explicit water was modeled with the TIP3P model [19]. All metal ions, except for Mg^{2+} were modeled with the CHARMM 27 force field. Mg^{2+} maintained a charge of 2+, however, its van der Waals parameters were modified to vdW radius R^* = 1.300 Å, and well depth was modified to ε = 0.06. These changes serve to avoid overestimation of Mg-O coordination in accordance with previous studies [20, 21]. Periodic boundary conditions were applied, and, the particle mesh Ewald summation was used to obtain accurate electrostatic interactions. Langevin-type thermostat and barostat were used to maintain the temperature at 300K and the pressure at 1bar,

respectively. All systems were subject to an optimization of 10000 steps and an equilibration of 200 ps before production runs with a time step of 1fs. All simulations were performed with NAMD 2.9 [22] and analyzed with VMD 1.9.1 [23].

3.3.2.2 Free energy perturbation

To compare the stability of different nucleotides in the addition site, we performed free energy perturbation (FEP) calculations to measure the free energy differences of changing one nucleotide to another. This was performed both in RNAP II ($\Delta G_{protein}$) and in water solution (ΔG_{water}). From this, the binding free energy difference (ΔG_{bind}) was calculated as

$$\Delta G_{bind} = \Delta G_{protein} - \Delta G_{water}.$$

A dual-topology paradigm as implemented in NAMD 2.9 was employed in all FEP calculations. In the case of the perturbation from GTP to UTP, only the base is perturbed while the ribose and the triphosphate groups remains the same. The perturbation was performed similarly for the case of GTP to 2'-dGTP. The soft-core van der Waal's potential with a radius-shifting coefficient of 5 was enabled. Electrostatic interactions of the annihilated particles were linearly decoupled from the simulation between $\lambda = 0$ and $\lambda = 0.5$, and electrostatic interactions of the appearing particles were decoupled from the simulation between $\lambda = 0.5$ and $\lambda = 1$. Van der Waals interactions of the annihilated particles were linearly decoupled from the simulation as λ increases from 0 to 1while vdW interactions of the appearing particles were coupled to the simulation as λ increases from 0 to 1. There were 22 λ windows in total. The first two windows increment lambda from $\lambda = 0$ to $\lambda = 0.005$ and from $\lambda = 0.005$ to $\lambda = 0.05$, while the last two windows increment lambda from $\lambda = 0.95$ to $\lambda = 0.995$ and $\lambda = 0.995$ to $\lambda = 1$. The remaining 18 windows, increment from $\lambda = 0.05$ to $\lambda = 0.95$ by a value of 0.05. Each window spanned over

500ps with time step of 1fs preceded by an equilibration of 50ps. The energies were saved every 150fs for ensemble averaging at the end of each window by the following equation:

$\Delta A_{a \to b} = -1/\beta \ln \langle \exp\{ -\beta [H_b(x, p_x) - H_a(x, p_x)]\} \rangle_a$,

where $1/\beta = k_B T$, k_B is the Boltzmann constant and T is the temperature. $H_b(x, p_x)$ and $H_a(x, p_x)$ are the Hamiltonians characteristic of states a and b, respectively. $\langle \dots \rangle_a$ denotes an ensemble average over configurations representative of the initial, reference state, a. Details about the theory can be found in [24, 25].

3.3.2.3 Umbrella Sampling

To calculate the potential of mean force of GTP transport from the E site to the A site, we performed umbrella sampling (US) calculations where the US potential was constructed as

$$V(x) = \frac{1}{2} k (x - x_0)^2,$$

in which k = 10 kcal/(mol·Å^2) and x_0 is the equilibrium center.

x was calculated as the center of mass (COM) between the GTP and a dummy atom. The dummy atom is positioned at the COM of the last residue of the RNA primer, determined from a snapshot of regular MD trajectories of GTP in the A site. In total, there were 32 windows, located between x = 5.1 and x = 14.1. The starting structures of each window were snapshots of MD trajectories of the GTP in the A site, the GTP in the E site, and the interpolated structures where COM of the GTP lies within the proximity of the x_0 for that window. In order to better investigate BH function, US calculations were also conducted on the distance between N7 of the GTP and O_γ of Rpb1-Thr827 (of the BH). , This simulation used 22 windows between the distances x = 3.3 and x = 9. Again, the starting structure of each window was a snapshot of MD trajectories, selected from the interpolated structure trajectories. In both US calculations, each window spanned over 2ns with a step size of 1fs. This included an equilibration of 100ps in the

beginning. X values were collected every 0.1ps. Post-processing of the US data was performed using Grosfield's version of the weighted histogram analysis method (WHAM) [26]. The convergence tolerance of the WHAM analysis was 0.001kcal/mol.

3.4 Results and discussion

3.4.1 Different NTPs in the addition site

To understand the specificity of RNAP II to different NTPs, a 7-ns-long MD simulation of the 2E2H structure with the correct NTP, a guanosine triphosphate (GTP), in the A site was performed. Concurrently, two additional MD simulations using the same structure were run with a 2'-deoxyGTP and a uridine triphosphate (UTP), respectively, instead of a GTP. These models represent incorrect NTPs. Snapshots from the trajectories of the above three systems are compared in Figure 3-2. In Figure 3-2A, the correct GTP forms stable hydrogen bonds with the template DNA base and interacts with the RNA primer through base stacking. In Figure 3-2B, these two types of interactions are also present for the 2'-dGTP. However the 2'-dGTP ribose ring lies in a plane perpendicular to the GTP ribose ring. Consequently, the 2'-C points away from the adjacent RNA nucleotide instead of towards it, indicating a 180-degree rotation in the 2'-dGTP ribose when compared with the GTP ribose. In Figure 3-2C, the unmatched UTP loses hydrogen bonds with the template DNA base, leading to its base detaching from the addition site and pressing the residues interacting with the triphosphate groups. It is noteworthy that the matched GTP moved slightly downstream along the template DNA strand whereas the non-cognate GTPs shifted slightly upstream. This is seemingly characteristic of a proofreading mechanism [10, 27]. The downstream translocation upon binding a correct NTP has also been found by Feig and Burton in their MD simulations [12]. Although unbinding of the non-cognate NTPs were not observed due to the relatively short simulation time, their instability in the A site

can nonetheless be inferred from their inadequate interactions with surrounding residues. Quantitative measures of the instability will be presented in 3.4.2.

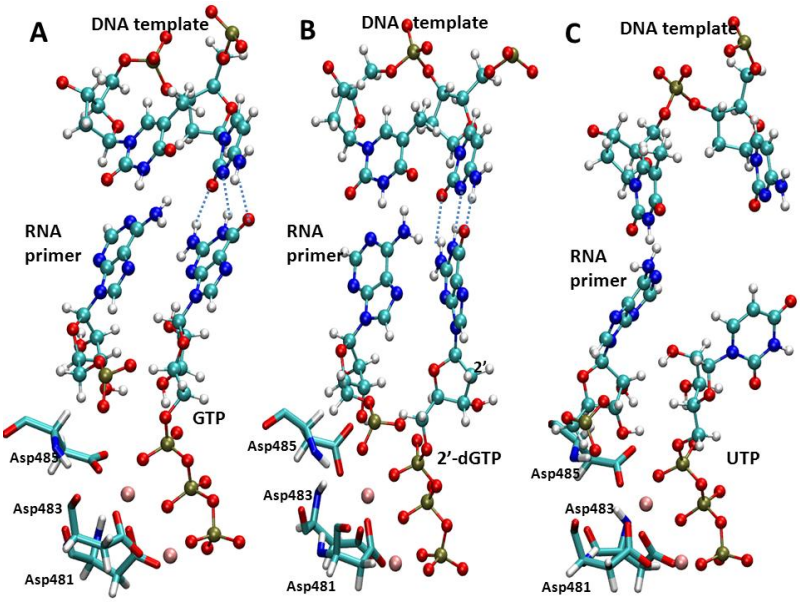


Figure 3-2: Different types of NTPs in the addition site

A) GTP, B) 2'-dGTP, C) UTP) in the addition site where hydrogen bonds between the NTP and the DNA template are shown in blue dots, Mg ions in pink balls and the aspartic triad in licorice.

The difference between the cognate GTP and unmatched UTP is primarily recognized by the DNA template through hydrogen bonds. Nevertheless, the enzyme residues surrounding the base contribute substantially to the orientation and stabilization of the base as shown in Figure 3-3. Notably, the Rpb1-Leu1081 of the trigger loop (TL) is believed to position the base with its iso-propyl side chain through hydrophobic interactions [14]. This is also found in our MD trajectories of the base-paired GTP and 2'-dGTP in the addition site while it is absent in the case of the unmatched UTP as its base regresses from the DNA template. In addition to Leu1081, other residues with hydrophobic side chains such as Thr827, Ala828 and Thr831 of Rpb1 also interact with the base through methyl groups. Of interest, all these three residues belong to

another important motif of the RNAP II, the bridge helix (BH), and are notably located at the bent region of the BH - believed to play an important role in the NTP binding process [10]. Throughout the simulations of all NTPs, the bend of the BH is never straightened and the hydrophobic side chains thereof are always positioned towards the base of the cognate NTP.

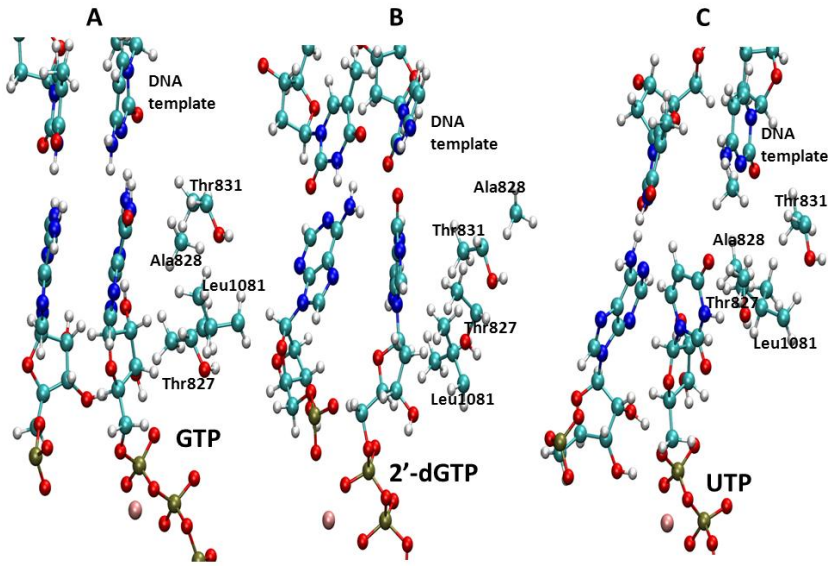


Figure 3-3: Interactions between the bases of NTPs

A) GTP, B) 2'-dGTP, C) UTP and the addition site residues Note the hydrogen-bond interactions in A) and B) and their absence in C). See text for a discussion of the hydrophobic interactions with various important residues of the BH and TL.

The difference between the GTP and 2'-dGTP lies in the presence of the 2'-OH group of the ribose ring. When it is present in the case of GTP (Figure 3-4A), Rpb1-Arg446 hydrogen-bonds to the 2'-OH as a hydrogen donor and positions the 2'-OH to form an intra-molecular hydrogen bond with the GTP base. It is found in the trajectory that the internal hydrogen bond is more stable than the one with Rpb1-Arg446 and promotes the GTP base to be close to the DNA template. In the case of 2'-dGTP (Figure 3-4B), both the internal hydrogen bond and the one with

Rpb1-Arg446 are missing, resulting in a 180-degree rotation of the ribose. This rotation is evident in the comparison of the C2'-C1'-N9-C4 dihedral angle between the GTP and 2'-dGTP trajectories (Supporting information S3-1). As a result of this ribose rotation, the 2'-dGTP base and the DNA template base are not so well aligned when compared with the GTP case, leading to weaker hydrogen bonds. In the case of UTP, the ribose is in the same orientation as in the case of GTP and a weak interaction exists between its 2'-OH and Rpb1-Arg446. Another important group of the ribose is the 3'-OH which reportedly interacts with Rpb1-Asn479 [10]. Although this interaction is present in the early part of the trajectories, it falters in the majority of the trajectories, and instead interacts with Rpb1-Arg446, orienting the guanidinium group towards the ribose 2'- and 3'-OH (distance between Arg446 and Asn479 shown in Supporting Information S3-2). Our observation that Rpb1-Asn479 indirectly interacts with the 3'-OH through Arg446 could be the reason why the Rpb1-N479S mutant caused loss of selectivity for 3'-OH [10].

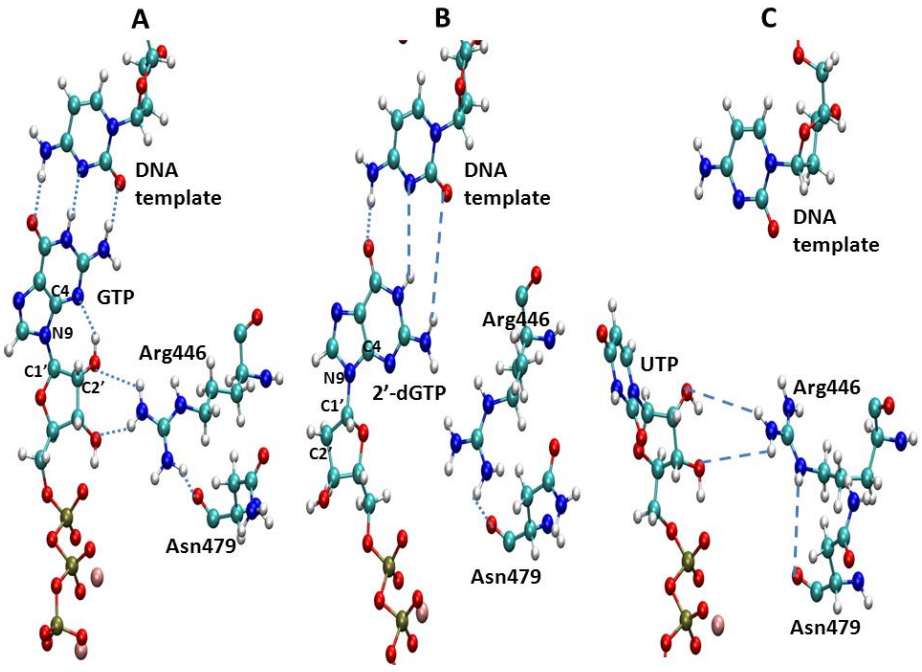


Figure 3-4: Interactions between riboses of NTPs

A) GTP, B) 2'-dGTP, C) UTP) and surrounding residues where strong hydrogen bonds are shown in blue dots and weak ones in blue dashes.

The triphosphate group is a common structure shared by all NTPs. Although not discriminatory for the different NTPs, the triphosphate is a major contributor of NTP binding to RNAP II. It anchors the NTP in the addition site by coordinating with two Mg ions, Mg(A) and Mg(B), which coordinate tightly with the well-known aspartic triad, Rpb1-Asp481, Asp483 and Asp485. In the trajectories of all NTPs, the aspartic triad remains in the same position with very little deviation. Additionally, the coordination spheres of the Mg ions appear quite stable. Mg(A) coordinates with the aspartic triad, α- and β-phosphate and a solvent water molecule, as has also been reported by Feig and Burton [12]. It is noteworthy that Mg(A) did not form any coordination with the 3'-OH of the RNA primer in all the trajectories. Mg(B) coordinates with Asp481, Asp483, β- and γ-phosphate, and Rpb2-Asp837. Besides the Mg ions, the negatively charged triphosphate also interacts with Rpb2-Arg766 and Rpb2-Arg1020.

Based on studies of its various protonation states, the trigger loop's His1085 has been reported to interact with the β-phosphate [14]. In our pKa calculations of the 2E2H structure, the pKa values of N_δ- and N_ε-protonated His1085 are similar, 5.08 and 5.24, respectively. Considering the neutral condition of cells *in vivo*, either of these two positions can be protonated, but not both simultaneously. Since the N_δ is closer to the β-phosphate, His1085 is protonated at the N_δ position. The trajectories show that in the case of GTP, His1085 is around 4Å from the β-phosphate early on and then retracts to a distance of about 6Å while in the other two cases, His1085 shifts from the proximity of the β-phosphate to the vicinity of the γ-phosphate (Figure 3-5). In place of His1085, another residue of the TL, Asn1082, interacts with the β-phosphate through its amide group. Moreover, interaction with the bridge oxygen ($O_{\alpha\beta}$) between the α- and β- phosphate is of particular interest, as the P_α- $O_{\alpha\beta}$ bond is broken after the nucleotidyl transfer

reaction. Rpb2-Tyr769 is found to hydrogen bond with $O_{\alpha\beta}$ in the trajectory of GTP. This interaction is weaker in the trajectory of 2'-dGTP, and disappears in the trajectory of UTP (Supporting Information S3-3). Since Rpb2-Tyr769 may help stabilize the pyrophosphate during the reaction, diminished interaction with it could be a source of selectivity for the cognate NTP. Another reported aide in catalysis, is a salt-bridged pair Rpb2-Glu529 and Rpb2-Lys987 which is believed to help position the P_{α} atom through a hydrogen bond between Rpb2-Lys987 and the 5'-O of the NTP. [12, 28] Correlation between the salt bridge and the hydrogen bond is also found in our trajectories (Supporting Information S3-4).

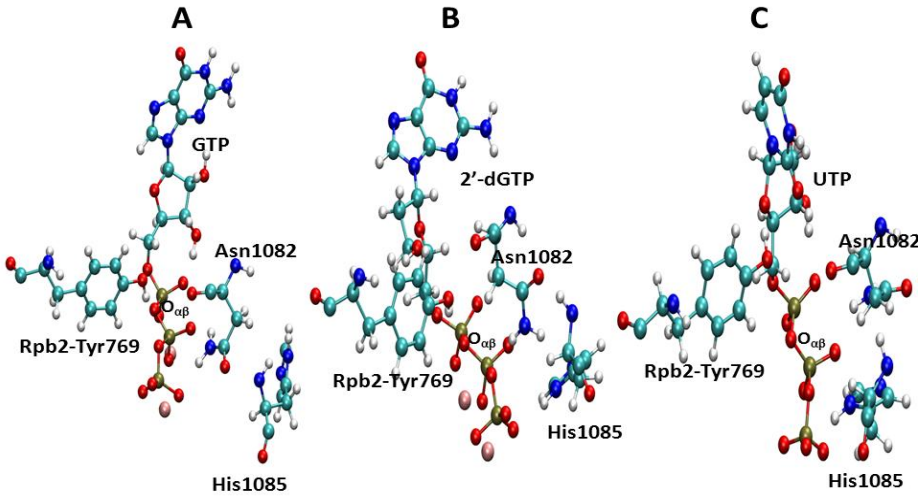


Figure 3-5: Interactions between the triphosphates of NTPs
A) GTP, B) 2'-dGTP, C) UTP) and surrounding residues. See text for discussion.

3.4.2 Thermodynamic stability of NTPs in the addition site

As mentioned in 3.4.1, stabilization of the correct NTP occurs through a network of cooperative interactions with the DNA template, the RNA primer, the surrounding enzyme

residues, and the co-factor Mg ions. As, individually, these components are insufficient for A-site stability, all relevant components have to be considered inclusively to quantitatively measure the stability. To this end, we performed free energy perturbation (FEP) calculations to compare the overall stability of the cognate GTP and the non-cognate 2'-dGTP and UTP. Binding free energy is calculated as the difference between the NTP in solution and in the addition site. Results are summarized in Table 3-1. As can been seen from Table 3-1, perturbation from GTP to 2'-dGTP results in a free energy increase of 1.91kcal/mol, while perturbation from GTP to UTP results in an increase of 16.80kcal/mol. The larger instability of UTP over that of 2'-dGTP is in agreement with the observation that RNAP II is more discriminatory against an unmatched NTP than a matched 2'-dNTP. The small difference between the stability of GTP and that of 2'-dGTP suggests that the selectivity for the 2'-OH is mainly due to reduced catalysis instead of A-site binding stability. This finding is in direct support of the hypothesis originally postulated by Kornberg and colleagues [10]. This is also compatible with the observation that only 2'-dNTPs have been found bound to the A-site in a crystal structure; unmatched NTPs never do. [10, 11].

Table 3-1: Binding free energy differences between GTP, 2'-dGTP and UTP

Free energy changes when mutating a cognate GTP in solution and in RNAP II, and the total binding free energy as a difference between the former two energies

	GTP → 2'-dGTP	GTP → UTP
In solution (kcal/mol)	-50.61	24.40
In RNAP II (kcal/mol)	-48.70	41.20
Binding free energy (kcal/mol)	1.91	16.80

3.4.3 NTP transfer from the entry site to the addition site

The above thermodynamic analysis of the RNAP II complex is based on the crystal structure (PDB:2E2H) with the NTP in the addition site (Figure 3-1A). Another crystal structure (2E2I) reveals the NTP's location in the entry site (Figure 3-1B). In this structure, notably, the bridge helix (BH) is straight and the trigger loop (TL) open. We subsequently analyzed a possible mechanism for how the NTP transfers from the entry (E) to the addition (A) site in association with the conformation change of the BH and TL. The 2E2I structure actually contains a 2'-dGTP, not a GTP in the E-site. As a result, to represent the cognate NTP, we replaced the 2'-dGTP with a GTP. One-ns-long MD simulations performed on both the 2'-dGTP and GTP systems show slight movement of both substrates towards the A site. These are largely characterized by interactions with Gln1078 of the TL. Neither the BH nor the TL moves significantly. It should be noted that the original crystal structure does not include the Mg(B) ion. Additionally, the Mg(A) ion is surrounded by the aspartic triad and not coordinated with the 2'-dGTP in the E site. Since Mg(B) is required in the two-metal-ion mechanism, it was added to coordinate with the 2'-dGTP and GTP for our simulations (details in Methodology). After equilibration, Mg(A) coordinates with the triphosphate, resulting in a position different from that in the crystal structure.

Since the NTP transfer involves large domain motions beyond the nanosecond scale, regular MD simulations are not adequate to sample the entire process. As a result, three transition structures were linearly interpolated between the structures of the A site and the E site (Supporting information S3-5), which were then followed by regular 7-ns MD to minimize artifacts (details in Methodology).

The trajectory of the interpolated structure with 70% E site and 30% A site shows significant movement by the TL towards the GTP. The ribose and the GTP base are oriented towards their positions in the A site (Figure 3-6A). In the trajectory, the ribose of the GTP is first drawn up by Gln1078 of the TL. This then orients the base to interact with Gln1078. This in turn props up the ribose. The side chain of glutamine is composed of both a hydrogen donor and a hydrogen acceptor which makes it a versatile residue to interact with different parts of the substrate. Other residues of the TL are not found to be in close contact with the GTP although the TL is moving towards the GTP throughout the trajectory.

Extensive interactions between the GTP and the TL are seen in the trajectory of the interpolated structure with 50% E site and 50% A site. In this trajectory, the TL is seen to move close to the GTP. As such, the side chains of Leu1081, Asn1082, Thr1083 and His1085 align and interact with the GTP (Figure 3-6B). Among the residues interacting with the GTP, Leu1081 interacts with the base, forming hydrophobic interactions. Meanwhile, the remaining residues align with the ribose and triphosphate, forming mainly hydrophilic interactions. There is no interaction found between the GTP and the BH at this stage.

Considerable interactions between the GTP and the TL are found in the trajectory of the interpolated structure with 30% E site and 70% A site. In this trajectory, the GTP is pushed up towards the BH, and in the process contacts the BH. This results in interactions between the BH and GTP, inducing a more pronounced bend within the middle of the BH (Figure 3-6C). Thr827, Ala828, Thr831 and Ala832 of the BH located within the bend region, align with the ribose and base of the GTP. While closely interacting with the BH, the GTP shifts away from the TL. Eventually, the TL will re-approach the GTP as is evident from the structure of the GTP in the A site. The intermittent contact of the TL with the GTP suggests that the TL movement is likely not

driven by any specific affinity for GTP, but instead is mostly thermally driven. This is in accordance with its flexible nature as a loop structure.

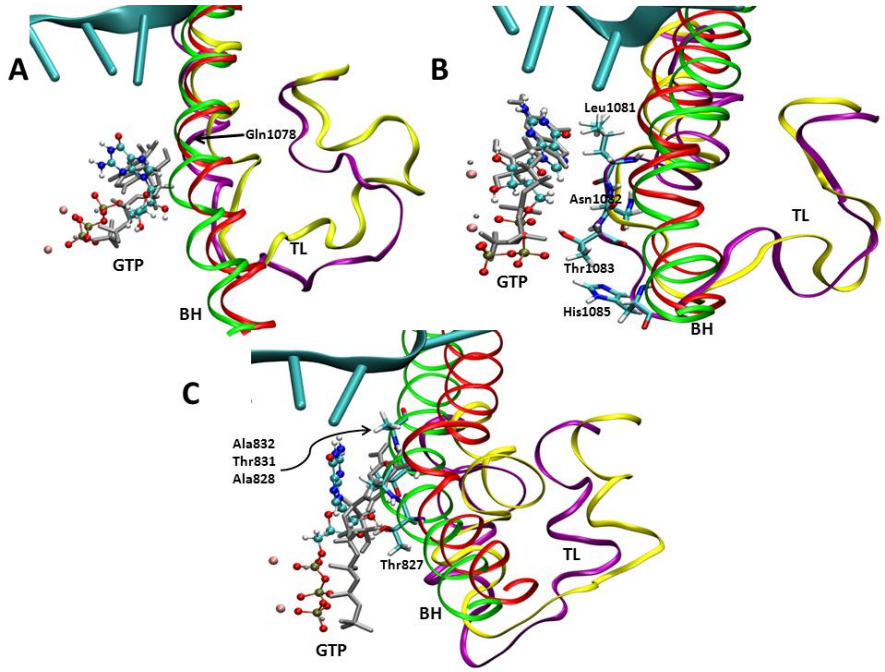


Figure 3-6: Comparison between initial and final positions of different domains

Initial positions of the TL (purple), the BH (green) and the GTP (gray licorice) and final positions of the TL (yellow), the BH (red) and the GTP (ball and stick). A) 70% E site + 30% A site B) 50% E site + 50% A site C) 30% E site + 70% A site

Piecing the above information together, a sequence of events in the NTP transfer can be outlined. In the E site, the TL comes in contact with the NTP through Gln1078 as driven by thermal movement. More residues of the TL, such as Leu1081, Asn1082, Thr1083 and His1085 then join in to push the NTP towards the A site. The NTP then meets the BH, a midway checkpoint, inducing a bend on the BH through interactions with Thr827, Ala828, Thr831 and Ala832. This pronounced bend helps push the NTP closer to the A site. When the NTP finally arrives in the A site, the BH bend remains and the TL locks in, along with the DNA template, and surrounding A-site residues stabilizing the NTP.

3.4.4 Free energy of the NTP transfer

Having delineated a path for the NTP transfer, it was pertinent to determine the energy associated with this path. This requires the explicit definition of a reaction coordinate. As the NTP interacts with different parts of the system on its way to the A-site, it is difficult to define a uniform reaction coordinate which incorporates all these interactions. In order to effectively include all interactions, a reaction coordinate should incorporate the ever-changing position of the GTP. In the A-site, since the NTP directly contacts the adjacent RNA primer through base-base stacking, the centre of mass (COM) of the last residue of the RNA primer serves as a good reference point for tracking GTP movement. To fix this reference point, a dummy atom is defined at the COM of the last residue of the RNA primer. The reaction coordinate is therefore defined as the distance between the COM of this fixed dummy atom and the COM of the incoming NTP, a GTP in this case. From the trajectories of the two crystal and three interpolated structures, 31 structures were selected to represent structural snapshots of the system throughout the GTP transfer. With this predefined path and the reaction coordinate above, an umbrella sampling calculation (details in Methodology) was then performed to recover the free energy of this process. The results are graphed in Figure 3-7A.

As shown in the graph, the free energy steadily decreases when the GTP transfers from the E site to the A site. This indicates that binding to the A site is more favorable than to the E site, which agrees with the experimental findings that a cognate NTP is always found in the A site. Moreover, the fact that the crystal structure (2E2I) with 2'-dGTP reveals electron density of a 2'-dGTP in both sites [11] can be explained by the relatively small free energy difference between these two sites coupled with the larger instability of 2’-dNTPs in the A site. Although there is an almost flat intermediate region when the GTP is 9Å away from the A site, this process

lacks any significant barrier, which suggests that it should be a rapid process and not rate-limiting, with active involvement of the BH and the TL. However, the shallow slope of the curve and the presence of plateaus, imply that this process is largely thermally driven. This would result in random back and forth movement by both the incoming NTP and the concerned protein domains, as is observed throughout the trajectories. The limited simulation time of our 7-ns MD of the interpolated structures is probably why we could not sample the entire process, since the TL motion is mostly thermally driven as mentioned above. Recently, the TL has been studied by Feig and co-workers [13] with targeted MD. In this study, the TL closing motion is simulated while the NTP is already bound in the A site. They found that the TL closing is spontaneous with barriers of around 5kcal/mol. As different from their study, our work involves both the TL and the NTP as the NTP moves from the E site to the A site, incorporating associated dynamic interactions which occur. This distinction could explain the difference between the free energy calculations, as the interactions between the TL and the NTP may be important for the TL closing (i.e that NTP entry, and TL closing are concurrent, and dependent events).

Since the choice of the reaction coordinate is a very important factor in umbrella sampling calculations, to better explore the specific energies associated with TL closure and BH bending, we conducted another umbrella sampling calculation using an interaction-based reaction coordinate. Compared with the thermally driven TL motion, the BH bend has a more solid structural basis as it is more rigid. Therefore, an interaction based reaction coordinate involving the BH bend should be more reliable than the pure COM distance. As we found in the trajectories of the interpolated structures, the N7 atom of the GTP base forms a hydrogen bond with the hydroxyl group of Thr827 when the bend is induced (Figure 3-7C). The distance between the N7 atom and the side chain oxygen of Thr827 is then defined as the reaction

coordinate. An umbrella sampling calculation of 22 windows was performed using one of the 31 selected structures as the start structure. This structure was selected because the GTP and Thr827 atoms are far apart. Results of this calculation are graphed in Figure 3-7B. The BH bending process approximately corresponds to the GTP transfer between 8.5 and 6.5 on the reaction coordinate axis. The free energy of the BH bending process follows a similar trend to that of the GTP transfer and their energy scales are consistent. This calculation also supports the idea that BH bending is an integral part of NTP entry and is not a defect in the crystal structure. This unique feature of the BH has been found in archaeal and bacterial RNA polymerases [8, 29] and suggested by researchers for eukaryotic RNAP II [30].

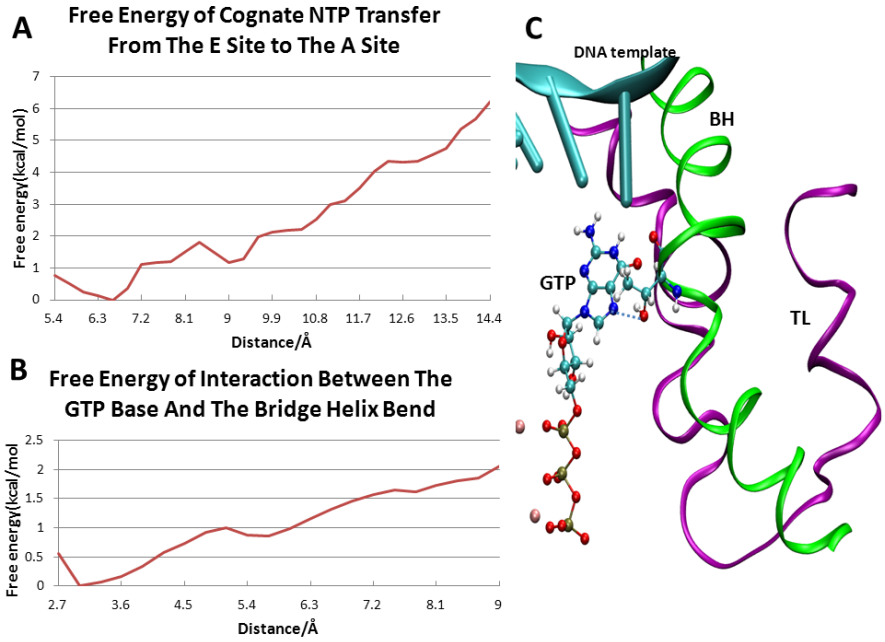


Figure 3-7: Free energy plots of the NTP transfer

A) Free Energy of cognate NTP transfer from the E site to the A site B) Free Energy of Interaction Between The GTP Base And The Bridge Helix Bend C) Interactions between the GTP and Thr827 of the BH where a hydrogen bond is shown in blue dots

In summary, both the TL and the BH are key to NTP transfer from the entry site to the addition site. For a cognate NTP, this transfer process should not be rate-limiting, with the participation of the BH and the TL. Since no constant specific interactions were found to cause selectivity, transfer of unmatched NTPs and 2'-dNTPs should be of the same nature as that of cognate NTPs. The selectivity mainly comes from the thermodynamic stability of different types of NTPs in the addition site. Likewise selectivity for the 2'-OH probably comes from catalytic efficiency, and not the kinetics of NTP entry.

3.5 Conclusion

In this work, we studied the thermodynamic stability of different types of NTPs, namely cognate NTP, 2'-dNTP and unmatched NTP, in the addition site (active site) of RNAP II. Additionally, we examined the kinetic transfer of a cognate NTP from a secondary binding site, the entry site, to the addition site. Functions of two important structural motifs, the bridge helix (BH) and the trigger loop (TL), are elucidated and quantified through molecular dynamics (MD) simulations and free energy calculations. Residues of importance in the NTP transfer and binding are summarized in Table 3-2.

Table 3-2: Summary of important residues in the NTP transfer and binding

	Trigger Loop	Bridge Helix	Other Residues of the A Site
Base of the NTP	Leu1081in the A site and during the NTP transfer Glu1078 during the NTP transfer	Thr827, Ala828, Thr831in the A site and during the NTP transfer	DNA template
Ribose of the NTP	Gln1078 during the NTP transfer		Arg446 (positioned by Asn479)

Triphosphate of the NTP	Asn1082 in the A site Asn1082, Thr1083 and His1085 during the NTP transfer	Rpb1-Asp481, Asp483, Asp485 and Rpb2-Asp837 through Mg coordination Rpb2-Arg766, Arg1020 with γ-phosphate Rpb2-Tyr769 with $O_{\alpha\beta}$ Rpb2-Glu529, Lys987 with O5’

The free energy difference between different types of NTPs demonstrates that the cognate NTP is the most stable NTP in the addition site. The 2’-dNTP is slightly less favored by 1.91kcal/mol, and the unmatched NTP is the least stable by 16.80kcal/mol. The MD simulation results indicate that the instability of the 2’-dNTP is due to its twisted ribose, a direct result of the absent 2’-OH, which results in less interaction with surrounding residues. The instability of the unmatched NTP lies in the lack of contacts between its misplaced base and surrounding residues, stemming from mismatching between the base and the DNA template. The free energy profile of the cognate NTP transfer from the entry site to the addition site suggests that the trigger loop and the bridge helix actively participate, and that this process is not rate-limiting with the participation of the trigger loop and the bridge helix. The results of thermodynamic stability and kinetic transfer calculations suggest that the NTP is mainly discriminated in the addition site. In the case of unmatched NTPs, this is directly the result of thermodynamic instability. 2’-dNTPs, however, are likely discriminated through catalytic inefficiency.

The simulations in this work cover up to 100ns timescales, which are substantial for a system of this size. However, compared to the actual transcription process, which occurs on the

order of milliseconds to seconds, these simulations may still not be statistically significant. Longer simulations or more advanced techniques will be necessary in the future. Although we employed a rather rudimentary biased simulation technique-linear interpolation between initial and final structures, more advanced techniques such as targeted MD or meta-dynamics should be adopted in future work. To make a biased simulation technique work efficiently, an effective reaction coordinate is important, especially for free energy calculations. Other types of reaction coordinate such as root mean square distance (RMSD) between structures should be tried. Moreover, more accurate potentials such as hybrid quantum mechanical/molecular mechanical potentials should be used to better examine the effect of A-site-bound 2'-dNTP on catalysis, as mentioned in 3.1. This approach has been adopted in our simulation of the catalysis for the cognate NTP [31].

It is our hope that these findings from theoretical simulations will complement previous experimental and computational studies and provide structural insights for future work on RNA polymerase II that will lead to a more complete understanding of the way it functions.

3.6 Supporting information

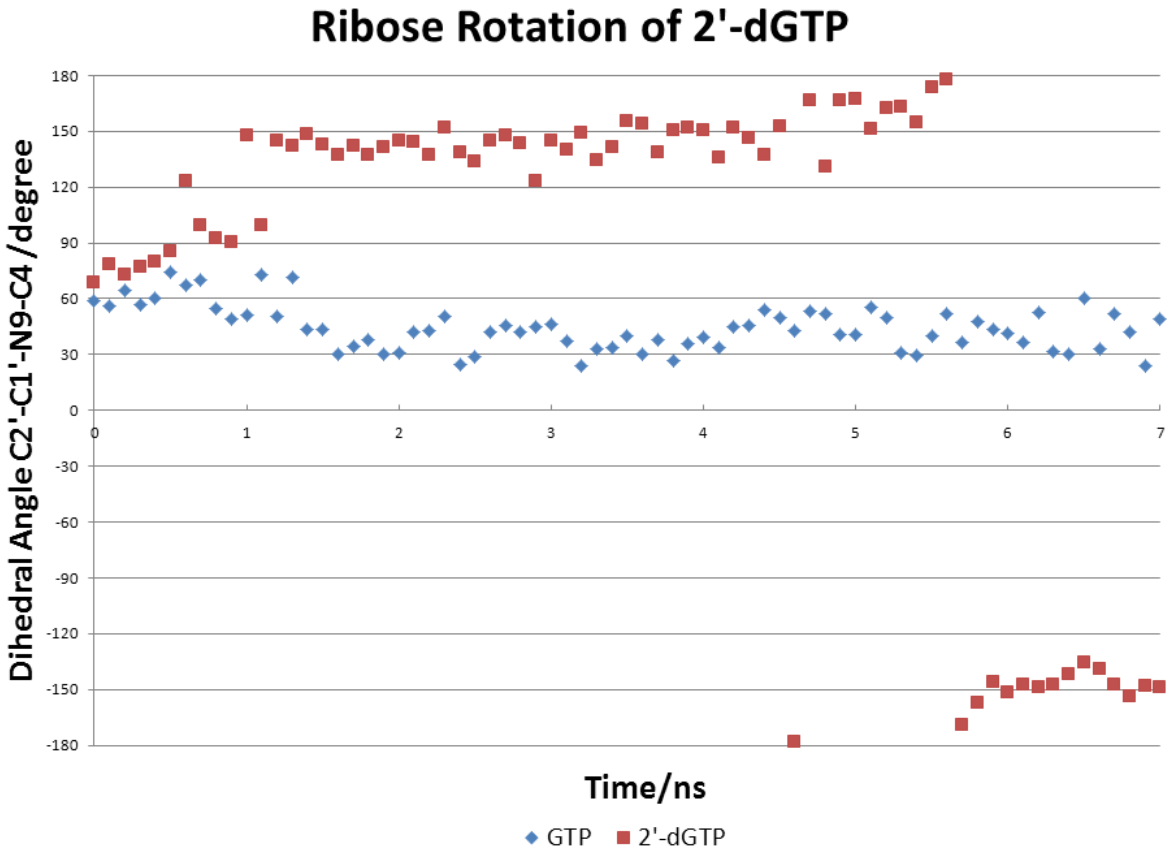


Figure S3-1: Ribose rotation of 2’-dGTP

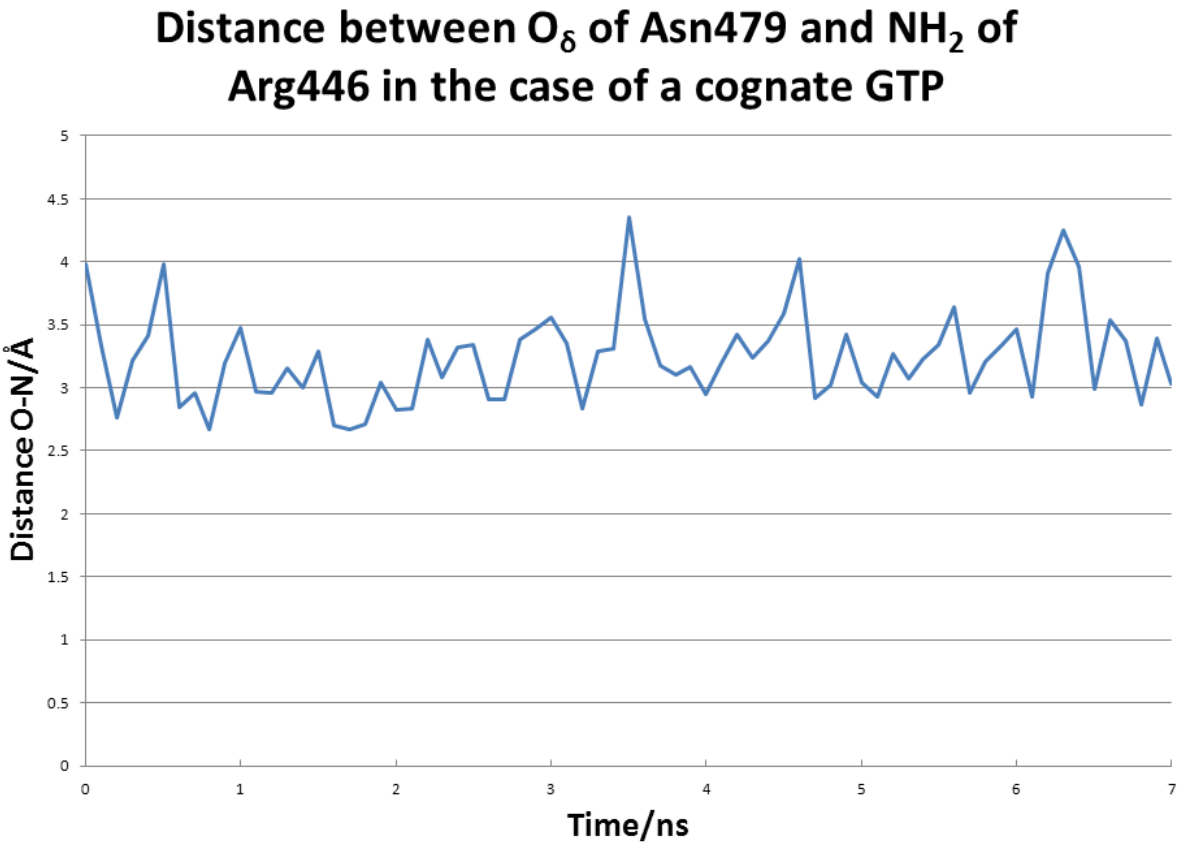


Figure S3-2: Distance between O_δ of Asn479 and NH_2 of Arg446 in the case of a cognate GTP

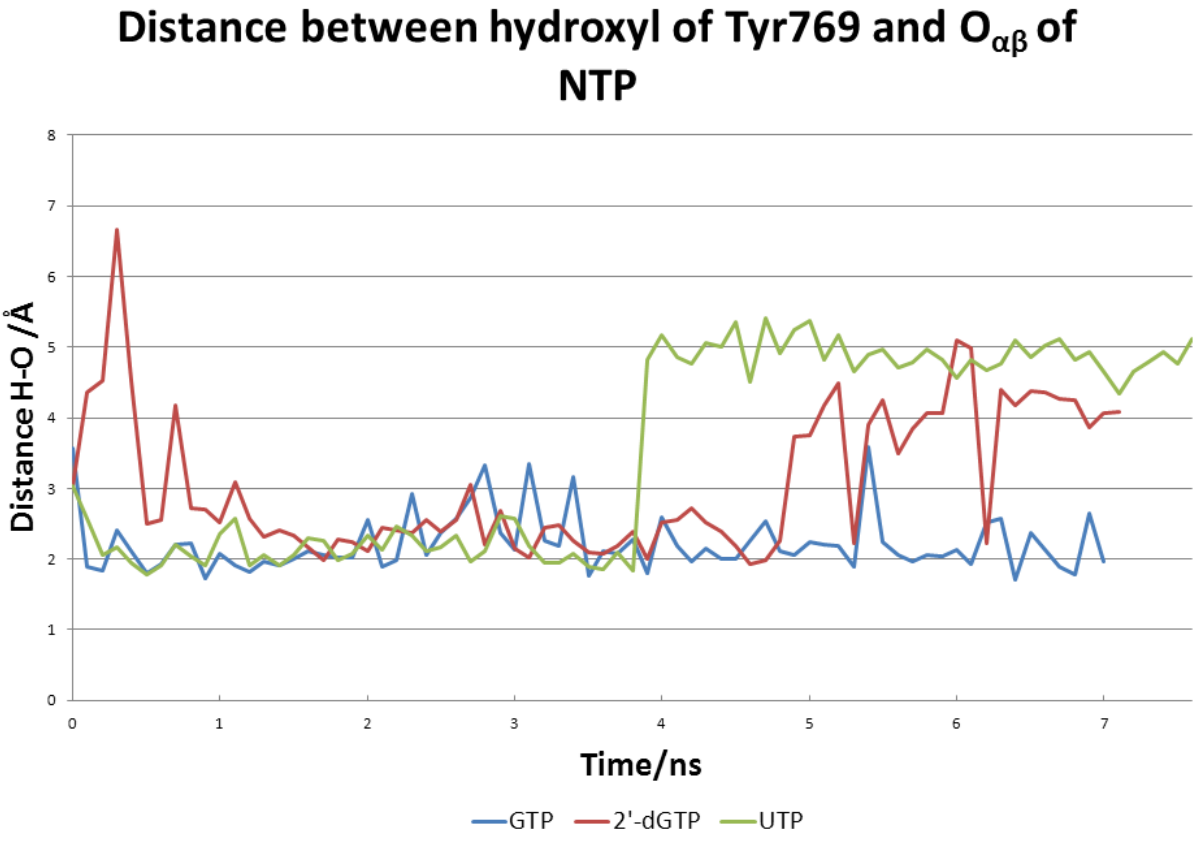


Figure S3-3: Distance between hydroxyl of Tyr769 and $O_{\alpha\beta}$ of different NTP

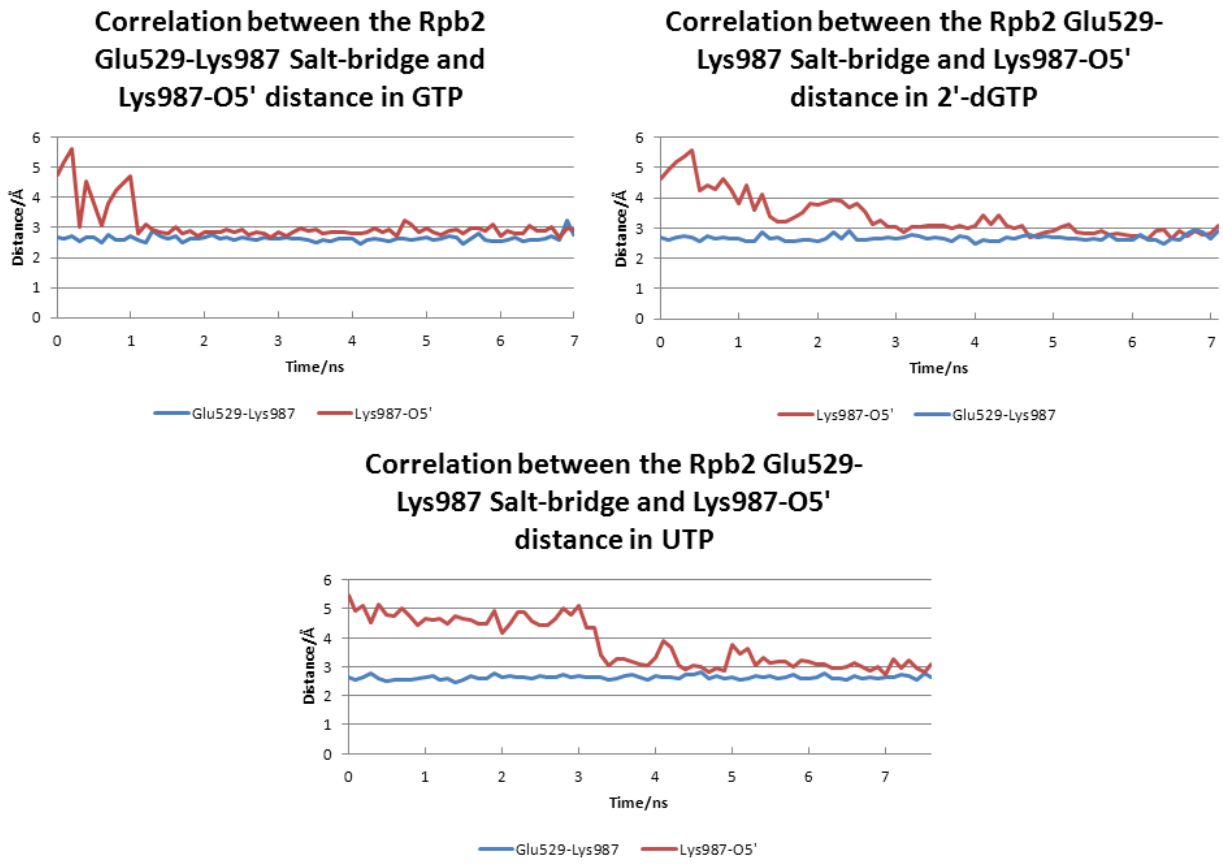


Figure S3-4: Correlation between the Rpb2 Glu529-Lys987 Salt-bridge and Lys987-O5' distance in different NTPs

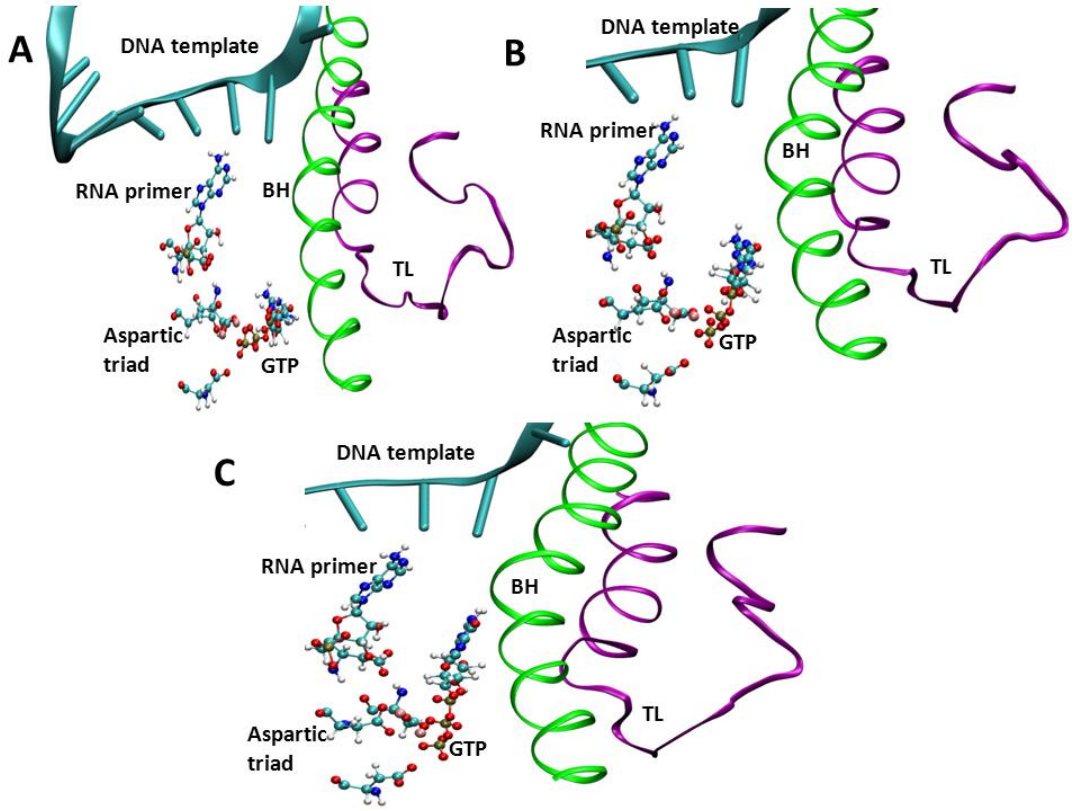


Figure S3-5: Initial interpolated structures between the addition site and entry site crystal structures

A) 70% entry site + 30% addition site B) 50% entry site + 50% addition site C) 30% entry site + 70% addition site where the trigger loop is colored in purple, the bridge helix in green, Mg ions in pink, carbon atoms in cyan, hydrogen atoms in white, oxygen atoms in red, nitrogen atoms in blue and phosphorus atoms in tan.

3.7 Bibliography

1. Ardehali, MB, Lis, JT. 2009. *Nat Struct Mol Biol* 16: 1123
2. Batada, NN, Westover, KD, Bushnell, DA, Levitt, M, Kornberg, RD. 2004. *P Natl Acad Sci USA* 101: 17361
3. Cramer, P, Bushnell, DA, Fu, JH, Gnatt, AL, Maier-Davis, B, Thompson, NE, Burgess, RR, Edwards, AM, David, PR, Kornberg, RD. 2000. *Science* 288: 640
4. Da, LT, Wang, D, Huang, XH. 2012. *J Am Chem Soc* 134: 2399
5. Feig, M, Burton, ZF. 2010. *Proteins* 78: 434
6. Gnatt, AL, Cramer, P, Fu, JH, Bushnell, DA, Kornberg, RD. 2001. *Science* 292: 1876
7. Larson, MH, Zhou, J, Kaplan, CD, Palangat, M, Kornberg, RD, Landick, R, Block, SM. 2012. *P Natl Acad Sci USA* 109: 6555

8. Tan, L, Wiesler, S, Trzaska, D, Carney, HC, Weinzierl, RO. 2008. *Journal of biology* 7: 40
9. Yuzenkova, Y, Bochkareva, A, Tadigotla, VR, Roghanian, M, Zorov, S, Severinov, K, Zenkin, N. 2010. *Bmc Biol* 8
10. Wang, D, Bushnell, DA, Westover, KD, Kaplan, CD, Kornberg, RD. 2006. *Cell* 127: 941
11. Westover, KD, Bushnell, DA, Kornberg, RD. 2004. *Cell* 119: 481
12. Feig, M, Burton, ZF. 2010. *Biophysical journal* 99: 2577
13. Wang, B, Predeus, AV, Burton, ZF, Feig, M. 2013. *Biophysical journal* 105: 767
14. Huang, XH, Wang, D, Weiss, DR, Bushnell, DA, Kornberg, RD, Levitt, M. 2010. *P Natl Acad Sci USA* 107: 15745
15. Uptain, SM, Kane, CM, Chamberlin, MJ. 1997. *Annu Rev Biochem* 66: 117
16. Foloppe, N, MacKerell, AD. 2000. *J Comput Chem* 21: 86
17. Lee, MS, Feig, M, Salsbury, FR, Brooks, CL. 2003. *J Comput Chem* 24: 1348
18. Brooks, BR, Bruccoleri, RE, Olafson, BD, States, DJ, Swaminathan, S, Karplus, M. 1983. *J Comput Chem* 4: 187
19. MacKerell, AD, Bashford, D, Bellott, M, Dunbrack, RL, Evanseck, JD, Field, MJ, Fischer, S, Gao, J, Guo, H, Ha, S, Joseph-McCarthy, D, Kuchnir, L, Kuczera, K, Lau, FTK, Mattos, C, Michnick, S, Ngo, T, Nguyen, DT, Prodhom, B, Reiher, WE, Roux, B, Schlenkrich, M, Smith, JC, Stote, R, Straub, J, Watanabe, M, Wiorkiewicz-Kuczera, J, Yin, D, Karplus, M. 1998. *J Phys Chem B* 102: 3586
20. Florian, J, Goodman, MF, Warshel, A. 2003. *J Am Chem Soc* 125: 8163
21. Oelschlaeger, P, Klahn, M, Beard, WA, Wilson, SH, Warshel, A. 2007. *Journal of molecular biology* 366: 687
22. Phillips, JC, Braun, R, Wang, W, Gumbart, J, Tajkhorshid, E, Villa, E, Chipot, C, Skeel, RD, Kale, L, Schulten, K. 2005. *J Comput Chem* 26: 1781
23. Humphrey, W, Dalke, A, Schulten, K. 1996. *J Mol Graph Model* 14: 33
24. Pohorille, CCaA. 2007. Springer Verlag
25. Chipot, C, Pearlman, DA. 2002. *Mol Simulat* 28: 1
26. Grossfield, A. *http://membrane.urmc.rochester.edu/content/wham/, version 2.0.6*
27. Cheung, ACM, Cramer, P. 2012. *Cell* 149: 1431
28. Domecq, C, Kireeva, M, Archambault, J, Kashlev, M, Coulombe, B, Burton, ZF. 2010. *Protein Expres Purif* 69: 83
29. Jovanovic, M, Burrows, PC, Bose, D, Camara, B, Wiesler, S, Zhang, XD, Wigneshweraraj, S, Weinzierl, ROJ, Buck, M. 2011. *J Biol Chem* 286: 14469
30. Hein, PP, Landick, R. 2010. *Bmc Biol* 8
31. Rui Zhang, AB, Dennis Salahub, Matin Field. 2013. *Unpublished*

CHAPTER FOUR: A GUIDE FOR QM/MM METHODOLOGY AND APPLICATIONS

4.1 Abstract

This review provides a general introduction to QM/MM methodology and techniques. It starts with the description of the effective QM/MM Hamiltonian and partitioning of the system, followed by geometry optimization and transition state search techniques. Subsequently, QM/MM sampling approaches are brought in for free energy calculations including thermodynamic integration, free energy perturbation, umbrella sampling and the path integral method to account for nuclear quantum effects. Lastly, QM/MM studies of DNA polymerases are reviewed as a state-of-the-art application to the study of enzymatic catalysis.

4.2 Introduction

Atomistic simulations and molecular modeling of complex biomolecular phenomena remain challenging, despite the enormous recent advances in computational capacity. Molecular mechanical (MM) methods have enabled us to perform molecular dynamics (MD) simulations of enormous chemical as well as biological systems, up to a few hundred thousand atoms. However, the molecular mechanical force fields are unable to describe the changes in the electronic structure of a system undergoing a chemical reaction. To meet this need, quantum mechanical (QM) methods need to be employed to account for bond-breaking and bond-forming, charge transfer, and electronic excitation. Unfortunately, QM methods are only applicable to relatively small systems, up to several hundred atoms, due to their prohibitive expense for large systems of, say, thousands of atoms.

A natural solution to this dilemma is to combine QM and MM together as a powerful hybrid entity. The combination of QM and MM, often denoted as QM/MM, allows us to investigate large and complicated systems at a reasonable cost while still yielding necessary

accuracy. Seminal contributions to the QM/MM methods, pioneered by Warshel and Levitt [1], and Singh and Kollman [2], have blazed the trail for the efforts that followed. Their early work was improved on by Field, Bash and Karplus [3] through the introduction of electrostatic embedding into the QM region. Since then, development of QM/MM methodology has become a red-hot subject [4-26] and its rapid growth and successful applications have greatly facilitated insightful understanding of the chemical properties of solutes in solution [27-37], chemical catalysts [38-45] and biological molecules [46-52] [10-26, 53] and its rapid growth and successful applications have greatly facilitated insightful understanding of the chemical properties of solutes in solution [27-37], chemical catalysts [38-45] and biological molecules [46-52, 54-59] [60-72] [46, 51, 59, 62, 73-84] [51, 64, 71, 80, 85-104].

Besides its wide use in the study of inorganic and organic chemical reactions, QM/MM has proven to be extremely successful in the study of biochemical, especially enzymatic, reactions and therefore has been widely applied in this field. Numerous reviews have been devoted to overviews of QM/MM studies of biochemical reactions over the last twenty years [44, 105-117]. In view of QM/MM's remarkable importance in theoretical studies of biological systems, the present review paper is primarily dedicated to QM/MM methodology and applications pertaining to bio-relevant processes.

In this article, essential concepts and related techniques are first introduced as the basis of QM/MM methods. Following the introduction of methodology, a brief review of QM/MM applications is provided, highlighting a state-of-the-art application in enzymatic catalysis. Methodologically this chapter is loosely split into three parts. The first deals with standard definitions of different terms in the effective Hamiltonian, with approaches to the decomposition of the system into classical, coupled and quantum terms. The second part deals with innovations

in techniques for optimizing geometries, for defining reaction paths and for locating transition states. The content of this part fits well into the rich tradition of quantum chemistry and its application to studies of chemical kinetics. The third part focuses on approaches that have to deal with a large number of degrees of freedom and yet have to treat at least part of the system quantum-mechanically. The methodology arises from the classical approaches of statistical mechanics (ensembles, collective variables, the projector-operator formalism) or quantum statistical mechanics (path integral methods, quantum Monte-Carlo approaches). To include effects of the environment on different chemical processes ranging from chemical reaction paths to ion solvation, one has to evaluate the free energy of the process. Collective variables and their averages are evaluated for a particular statistical ensemble and the QM treatment for part of the system provides the desired level of detail. The need for a good statistical description of the most relevant ensemble properties gives rise to different enhanced sampling techniques. One of the goals for this chapter is to provide the reader with answers to "Frequently Asked Questions" regarding the utilization of QM/MM techniques in studies of chemical reactions taking place in condensed phases based on numerous examples found in enzymology, ion channels and transporters, theoretical and physical chemistry.

4.3 Basic concepts of QM/MM methodology

Before going any further, it is useful to introduce an effective Hamiltonian that can be used for the description of the system's energetics and dynamics.

In a typical QM/MM scheme, a system is usually divided into two subsystems: the QM subsystem treated by "high-level" QM methods and the MM subsystem treated by "low-level" force field-based methods. The boundary between these two subsystems distinguishes the QM region from the MM region. Ideally, partition of the system should not cut any covalent bonds to

ensure the completeness of the QM subsystem. However, crossing covalent bonds is often unavoidable for large molecules, such as polymers and proteins. Here we will use the Link Atom method to illustrate the general features, returning to the issues arising from the boundary after the Hamiltonian is introduced.

In order for a QM/MM calculation to mimic the real system the QM chemical structure has to be complete, i.e. no dangling bonds are permitted. An intuitive way to remedy a dangling bond is to cap it with an artificial atom, which gives rise to the so-called Link Atom approach [2, 3, 13, 17, 53, 118]. In this approach, an additional atom, as a link between the QM and MM regions, is added to saturate the QM frontier atom at one end of the cut covalent bond. The link atom scheme is illustrated in Figure 4-1. This link atom in most cases is a hydrogen atom because of its simplicity and practicality.

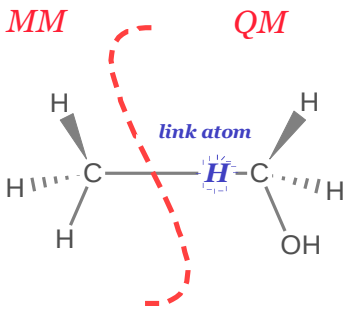


Figure 4-1: Illustration of the link atom scheme

This scheme uses ethanol as an example, in which the methyl group is treated as the MM subsystem and the rest is QM.

4.3.1 Energy expression

Two general schemes have been proposed in order to eliminate the artificial interaction between the link atom and the other QM atoms and that among link atoms – an additive scheme and a subtractive scheme.

4.3.1.1 Subtractive scheme

In the subtractive scheme, the QM subsystem with the link atom (QM+L) is calculated on both QM and MM levels and the entire system (S) without the link atom is treated on the MM level. The energy formulation is then:

$$E(S) = E_{MM}(S) + E_{QM}(QMS+L) — E_{MM}(QMS+L) \quad (1)$$

Since $E_{MM}(S)$ is the summation of the QM subsystem without the link atom (QMS), the MM subsystem (MMS) and the interaction between the two (QMS-MMS),

$$E_{MM}(S) = E_{MM}(QMS) + E_{MM}(MMS) + E_{MM}(QMS\text{-}MMS) \quad (2)$$

In the same fashion, we also have

$$E_{MM}(QMS+L) = E_{MM}(QMS) + E_{MM}(L) + E_{MM}(QMS\text{-}L) \quad (3)$$

Substituting (2) and (3) into (1), $E_{MM}(QMS)$ is cancelled and it gives

$$E(S) = E_{MM}(MMS) + E_{MM}(QMS\text{-}MMS) + E_{QM}(QMS+L) — [E_{MM}(L) + E_{MM}(QMS\text{-}L)] \quad (4)$$

Now it is evident that the correction comes from the last subtractive term in (4), which is expected to cancel the link atom contribution in $E_{QM}(QMS+L)$. The foregoing derivation has been clearly shown in the work of Bakowies and Thiel [9]. This subtractive scheme was implemented by Morokuma and co-workers in their IMOMM [12] and ONIOM [17, 119] protocols.

The upside of this scheme lies in its simple and implicit cancellation of the unwanted artifacts of the link atom. Nevertheless, its downside also resides in this cancellation, as it is controversial whether the link atom contribution at the QM level can be balanced by that at the MM level. Another drawback is the implied QM-MM interaction calculated at the MM level as indicated by the second term in (4).

4.3.1.2 Additive scheme

To explicitly calculate the QMS-MMS interaction on the QM level, an additive scheme has been reported. It is formulated as

$$E(S) = E_{MM}(MMS) + E_{QM}(QMS+L) + E_{QM}[(QMS+L)-MMS] \quad (5)$$

The last term in (5) accounts for the coupling between QM and MM subsystems. It consists of electrostatic, van der Waals and bonded interactions between QM and MM atoms, all of which will be detailed in following sections. This additive scheme is employed in the majority of QM/MM implementations [1-3, 13, 20, 53] [4, 6, 9, 120] .

Comparing (4) with (5), one can see two obvious differences. First, the subtractive term in (4) is omitted in (5), because $E_{MM}(L)$ is generally considered very small and thus makes little contribution to the total energy. And $E_{MM}(QMS-L)$ is usually regarded as a constant as its distance from the QM boundary atom is fixed in most boundary schemes [115] . Therefore when calculating the energy difference instead of the absolute energy, the link atom's contribution should be mostly cancelled out. Second, the QMS-MMS coupling may be calculated at the QM level, as indicated by the last term of (5). This term can either include or exclude the interaction between the link atom and MM atoms, but its inclusion proves to give better results. Although lacking physical cancellation for the link atom, the additive scheme turns out to perform well in practice.

However, when using the additive scheme for the link atom, one needs to take special care of the bonded QMS-MMS interaction at the boundary. The bending term QMA-QBA-MBA and torsional term QMA-QMA'-QBA-MBA (QMA or QMA' means any QM atom bonded to the QBA) should be eliminated since they are already accounted for by the link atom. The stretching term QBA-MMA should be retained to maintain the connection between QM and MM

subsystems. On the other hand, one need not worry about these terms in the subtractive scheme as they are corrected by calculating the QM subsystem at the MM level.

4.3.2 Electrostatic interactions

Both the energy expressions above contain an explicit term describing the QMS-MMS interaction. This interaction consists of the bonded interaction at the boundary (mentioned above) and also the electrostatic and van der Waals interactions. Two approaches describing electrostatic interactions are introduced here according to how the MM charges are embedded in the QM calculation.

4.3.2.1 Mechanical embedding

As implied in the subtractive formulation (4), E_{MM}(QMS-MMS) accounts for the QMS-MMS interaction at the MM level. Therefore the QM atoms are represented as point charges, bond dipoles or higher multipoles. In most cases, the point-charge model is adopted. However, this treatment is viewed as problematic because the charge density of the QM region is not actually polarized by the MM part.

Hagiwara and co-workers compared mechanical embedding in the subtractive scheme with electrical embedding (explained below) in the additive scheme by a QM/MM study of a protein-DNA complex. They found the HOMO energies differed by 23.7 kcal/mol as calculated by the two schemes [121]. This supports the viewpoint that the QM region has to be polarized.

4.3.2.2 Electrical embedding

To ensure the QM subsystem is polarized by MM charges, this charge-charge interaction has to enter into the QM Hamiltonian:

$$\hat{H}_{QM-MM} = -\sum_{i,m}\frac{q_m}{r_m} + \sum_{A,m}\frac{Z_A q_m}{R_{Am}} \qquad (6)$$

where q_m are the charges of MM atoms, Z_A the atomic number of QM atoms, i runs over all QM electrons, A over all QM atoms including link atoms and m over all MM atoms. The first term is a one-electron operator and the second accounts for the nuclei-MM charge interaction. When acting on the QM wave function, (6) results in the electrostatic interaction between QM and MM subsystems as a portion of E_{QM}[(QMS+L)-MMS] [13] .

As a matter of fact, electrical embedding can be implemented not only with the additive scheme but also with the subtractive scheme. Lin, Zhang and Truhlar demonstrated the formulation of a subtractive scheme with electrical embedding in their QMMM manual [122]. They showed

$$E(S) = [E_{bon}(S) — E_{bon}(QMS+L)] + [E_{vdW}(S) — E_{vdW}(QMS+L)] + E_{el}(MMS) + E_{QM}(QMS+L) \quad (7)$$

where bon stands for bonded interactions, vdW for van der Waals interactions, and el for electrostatic interactions. The last term of (7) includes (6) in the entire QM Hamiltonian. For a detailed derivation, readers may refer to Section 4G of [122]. By constructing electrical embedding into the subtractive scheme, the link atom artifacts are corrected at the MM level and thus should lead to a relatively better result than the additive scheme.

4.3.2.3 Electrical Embedding with explicit treatment of MM region polarization

Since the QM region is polarized by MM atoms, it would be unbalanced if the MM region were not affected reciprocally. A straightforward treatment based on the MM point-charge model is to include induced dipoles as a polarization effect. An early effort was conducted by Bakowies and Thiel [9], which is formulated as

$$E_{ind}(MM) = \frac{1}{2}\sum_m \mu_\alpha \langle F_\alpha \rangle \quad (8)$$

where the energy of induced MM dipoles is the summation of the m dipole moments μ_α of the MM subsystem multiplied by the electrical field F_α from the QM subsystem. μ_α depends on the polarizability tensor and the QM electrical field while the interactions between the dipoles are determined by the dipole moments and the dipole-dipole interaction tensor. Since the dipoles interact with each other, an iterative procedure must be applied to generate a self-consistent polarization.

Another polarizable force field, namely SIBFA (Sum of Interactions Between Fragments Ab initio computed), incorporates multipoles up to quadrupoles. It divides a macromolecule into elementary fragments comprising multipoles and having different polarizabilities, among which interactions are summed to obtain the total energy. A general energy expression for multipole-multipole interactions is

$$E_{multipole} = E_{mono-mono} + E_{mono-dip} + E_{mono-quad} + E_{dip-dip} + E_{dip-quad} + E_{quad-quad} \quad (9)$$

It is worth mentioning here that the monopole-monopole interaction is calculated by splitting the atom into core and valence electrons with a special parametrization, which, as a result, is different from the classical expression $E = \frac{q_i \cdot q_j}{ri_j}$. The same treatment is made for the monopole-dipole interaction. Details are in a recent review [123]. Linking SIBFA with QM engines is currently in progress [124].

A simpler approach to modeling electronic polarization is based on the Drude-oscillator model, in which a fictitious Drude particle with opposite sign is attached to a point charge by a harmonic spring thus introducing dipole induction in classical simulations [125]. The resulting electrostatic potential for a system containing Drude particles can be expressed as:

$$E_{elec} = \sum_{i}\sum_{i>j}\frac{q_i q_j}{r_{ij}} + (\sum_{i}\sum_{j'}\frac{q_i q_{j'}}{r_{i,j'}} + \sum_{i'}\sum_{i'>j'}\frac{q_{i'} q_{j'}}{r_{ij}}) + \frac{1}{2}\sum_{\alpha'} k_{\alpha'} d_{\alpha'}^2 \quad (10)$$

where the prime denotes Drude particles. The last term represents the oscillator self-energy expressed in familiar form with force constant k_α related to the site's polarizability (α) as $\alpha = \frac{q'^2}{k_\alpha}$

.

Higher-order multipoles are avoided by using the simple Drude model and the only difference with a non-polarizable function is the introduction of a new atom type –Drude. Parametrization of the Drude model in CHARMM [126, 127] for various systems has been undertaken by Lamoureux and co-workers [128-131]. It has been implemented recently in the QM/MM interface between CHARMM and deMon2k. Lev et al. [53] performed an analysis of the importance of the explicit treatment of the parameterization for the interaction energy in the water dimer. The interaction energies show little improvement with the inclusion of explicit polarization as both the polarizable and the classic model are parameterized to reproduce the energetics and geometry. However, MD simulations with explicit inclusion of dipole induction indicate that there is a significant component missing in the description of solvent dynamics around highly polarizable solutes. For the polarizable water around a sodium or potassium ion the magnitude of the induced dipole can be higher than 0.1 D, which can be very important in cases where electronic effects play a significant role.

4.3.2.4 First-principles electrostatic potential

Though polarized embedding includes the effect of electron polarization in the force field, its reliance on the point-charge model is nonetheless problematic as atomic charges in reality should be more distributed than simple monopoles. To mediate the point-charge model with over-concentrated charge, Darden and co-workers have proposed the Gaussian electrostatic

model (GEM) to represent the charge density [132-134]. In this model, the wave function of the system is first calculated with *ab initio* methods and the obtained electron density is then fitted with a set of Gaussian basis functions according to the variational principle [133]. The coefficients obtained from this procedure together with the basis set (currently s-type Gaussian functions) forms the frozen core to reproduce the Coulomb and exchange interactions. Polarization effects are described by parameterized dipoles at specific sites. Recently they have extended this model to higher-order multipoles and sped it up using reciprocal space methods to calculate long-range electrostatic interactions [134]. The GEM method has been tested with water dimers [132], the benzene dimer and water-metal complexes [134] and proved to perform consistently better than conventional point-charge models. It remains to be seen how useful it will be for large or very large systems, perhaps in conjunction with fragment-based approaches.

As a brief summary, electrostatic interactions have been demonstrated to be crucial to enzymatic reactions [135]. Therefore an accurate calculation of these interactions is required. More elaborate models are expected to reconcile accuracy and efficiency.

4.3.3 van der Waals interactions

Besides the bonded and electrostatic interactions, the QMS-MMS coupling also includes van der Waals (vdW) interactions. The vdW interaction is usually described by a Lennard-Jones 12-6 potential:

$$E_{ij} = \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \qquad (11)$$

where i runs over QM atoms and j over MM atoms, and A and B are constants pertaining to atom types. This contributes as a component of the last term in (5).

In a typical force field, the bonded, electrostatic and vdW parameters are optimized together using high-level calculations or experimental results. Using parameters of a certain term separately from others may sometimes cause trouble. In the case of QM/MM calculations, when electrical embedding is used, the vdW interaction of QMS-MMS could be incorrect as the corresponding electrostatic interaction is not the parameterized point charge-point charge interaction any more. To alleviate this issue, Friesner and co-workers have re-optimized the vdW parameters for amino acids in their QM/MM implementation [16]. It should be noted that they also included a hydrogen bond correction term in their scheme and thus the vdW parameters were parameterized accordingly. In their work, only the vdW radii were re-parameterized but not the well depth.

Recently, Mulholland and co-workers re-optimized vdW parameters from CHARMM 27 for nucleic acids with respect to the B3LYP DFT method [136]. Their results indicated that, for QM/MM investigations of nucleic acids, the standard force field vdW parameters might not be appropriate for atoms treated by QM. QM/MM interaction energies calculated with standard CHARMM27 parameters are found to be too large, by around 3 kcal/mol. They reasoned that this was because of an overestimation of electrostatic interactions and therefore reparameterized the vdW parameters to compensate for that.

However, Cui and co-workers have tested three sets of vdW parameters and concluded that the QM/MM energetics were not sensitive to the vdW parameters and efforts to improve QM/MM accuracy should focus elsewhere [137]. With SCC-DFTB as the QM method, they calculated a proton transfer process for a solvated enediolate and a solvated fused-ring molecule (FAD - flavin adenine dinucleotide) respectively. They found similar thermodynamic quantities

(the reduction potential deviated by 0.2 kcal/mol) for different vdW parameters although there were noticeable differences regarding solvent distribution functions around the solute.

An important difference between the modeled systems in [136] and [137] is solvation. While the former included only one water molecule for each base pair, and is thus essentially a gas-phase model, the latter employed a fully solvated system with explicit water. It was found in [137] that hydrogen bond length and energy deviated more for different parameters in the gas phase than in the condensed phase. On the other hand, the optimized vdW parameters from [136] were not tested in the condensed phase. Nonetheless, the choice of vdW parameters should be carefully considered as short-range vdW interactions could greatly affect the configuration of the QM region.

4.3.4 Boundary treatment

4.3.4.1 Link atom (LA)

When a covalent bond between the QM and MM subsystems is crossed by the boundary, and a link atom is introduced, it generates more problems for the QM part than for the MM part. A cascade of artificial effects is brought in by the link atom. First, it introduces unwanted interactions with the QM atoms and other link atoms, as discussed above. Second, a link atom bears three extra degrees of freedom that should not be present in the real system. Third, it is spatially too close to the MM frontier atom as it sits on the bond between the QM and MM frontier atoms. When the MM frontier atom is charged, this unrealistic closeness will cause an overestimated interaction between the MM frontier atom and the link atom, and hence an overpolarization of the QM subsystem.

To circumvent the second issue, the position of the link atom should be fixed in order to avoid the excess degrees of freedom. A straightforward way to achieve this goal is to relate its coordinates to its adjacent neighbors, i.e. the QM and MM boundary atoms. First proposed by Dapprich et al [17], the position of the link atom (LA) is defined as a function of the positions of the QM boundary atom (QBA) and the MM boundary atom (MBA) in Cartesian coordinates:

$$\vec{R}_{LA} = \vec{R}_{QBA} + \alpha (\vec{R}_{MBA} - \vec{R}_{QBA}) \qquad (12)$$

as is evident from

$$\alpha = \frac{\vec{R}_{QBA} - \vec{R}_{LA}}{\vec{R}_{MBA} - \vec{R}_{QBA}} \qquad (13)$$

where α can either be held constant as the ratio of the equilibrium bond lengths of QBA-LA and MBA-LA [17] or be varied as the ratio of the equilibrium bond lengths of QBA-LA and the distance of QBA-MBA [138, 139].

Moreover, in some work α is more elaborately defined, which involves the deviation of QBA-LA and that of QBA-MBA from the equilibrium bond lengths and the bond stretching constants of the QBA-LA bond (k_{QBA-LA}) and the QBA-MBA bond ($k_{QBA-MBA}$). In the work of Eichinge et al [27],

$$\alpha = \frac{(\vec{R}_{QBA} - \vec{R}_{LA}) - (\vec{R}_{QBA\,0} - \vec{R}_{LA\,0})}{(\vec{R}_{MBA} - \vec{R}_{QBA}) - (\vec{R}_{MBA\,0} - \vec{R}_{QBA\,0})} = \frac{k_{QBA-LA}}{k_{MBA-QBA}} \qquad (14)$$

By fixing the location of the link atom, elimination of the gradient on the link atom follows. Since the position of the LA is constructed by the positions of the QBA and the MBA, the gradient of LA should be accordingly projected onto the QBA and the MBA. Since α is the ratio of QBA-LA to QBA-MBA and the gradient is formulated as dE/dR, the LA's gradient portion to be projected on the QBA should be (1- α). When α is held constant, this gives [17]

$$\vec{G}_{QBA} = \vec{G}_{QBA0} + (1-\alpha)\cdot\vec{G}_{LA} \qquad (15)$$

$$\vec{G}_{MBA} = \vec{G}_{MBA0} + \alpha\cdot\vec{G}_{LA} \qquad (16)$$

where $\vec{G}_{QBA0}$ and $\vec{G}_{MBA0}$ are the gradients without the contribution from the link atom. When α varies, one will have to multiply the LA's gradient by a transformation matrix, as detailed in [115, 138, 139]. Thus the excess degrees of freedom from the link atom are avoided in geometry optimization and molecular dynamics.

The third issue regarding the artificial link atom is the overpolarization by the MM boundary atom when it is charged. An easy way to alleviate this drawback is to remove the charge of the MBA from the QM and MM subsystems interaction, which is termed the single link atom scheme [140]. However, this will result in an unrealistic extra charge in the MM subsystem. To eliminate artificially created charge, the deleted charge from the MBA can be redistributed over the rest of the residue group pertaining to the MBA [141]. In other schemes, charges of the entire MBA residue group are removed from the QM and MM subsystems interaction [2] or the MBA charge is redistributed among the rest of its residue group [142]. As opposed to the conventional point charge model, MM partial charges may be represented as Gaussian charge distributions centered on the respective atoms [15, 143, 144] . Double link atom [143] and charge shift [21] methods are also proposed to treat the boundary MM charges. More recently, Lin and Truhlar have developed two schemes to remedy this overpolarization problem: the redistributed-charge (RC) scheme and the redistributed-charge-and-dipole (RCD) scheme [23].

All the approaches mentioned above except the RC and RCD have been compared by Konig et al [141], with regard to different systems. They first calculated deprotonation energies

of alcohols and carboxylic acids, in which the alkyl group was treated with MM and the alkoxyl group with QM. As a result, the single link atom method always underestimated the energy by an average of 20 kcal/mol compared to high level calculations, whereas all the others overestimated the energy and the excluded group method yielded the closest result. When calculating the deprotonation energies of amino acids, the single link atom and excluded charge approaches produced the largest deviation whereas the Gaussian distributed charge and charge shift methods were the closest to the high level calculation result. It should be noted here that the Gaussian distributed charge model greatly depends on the blurred width and noticeable differences were found when different widths were chosen [141]. When evaluating the deprotonation energies of DNA bases, the single link atom method again deviated the most from high-level calculations while the others gave comparable results. And the two calculations using the Gaussian distributed charge model with different blurred widths yielded different energies. Activation energies and endothermicities were also calculated by Konig et al [141] with different treatments of the boundary MM charges. The comparison showed appreciable deviations among them in one enzyme (TIM – Triosephospate Isomerase) while the calculation in the MGS (Methyl Glyoxal Synthase) enzyme divided them into two groups: the single link atom and excluded group methods with lower energies and the others with higher energies. In summary, the single link atom method tends to underestimate the energies while other amended methods can improve the accuracy, which are case-dependent nonetheless.

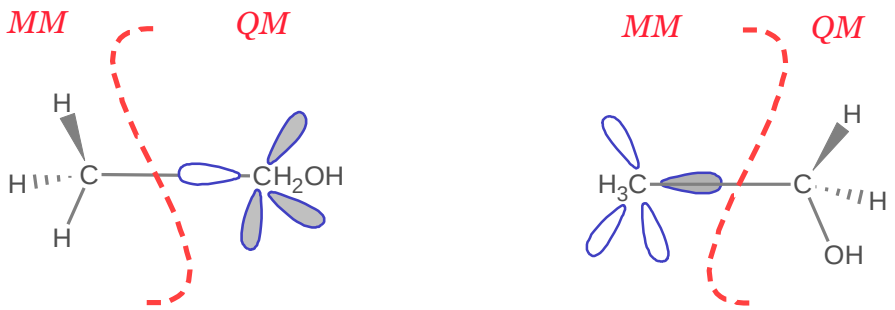
4.3.4.2 Frozen localized orbitals (FLO)

Even though extensive efforts have been made to eliminate the artificial effects introduced by the link atom, this approach still lacks a solid physical foundation and it is hence arbitrary and controversial. In order to completely root out the side effects of the link atom, the

dangling bond at the boundary may be capped by a frozen localized orbital instead of a link atom. This method dates back to Warshel and Levitt [1]. Another early attempt adopting this philosophy is the local self-consistent field (LSCF) approach as illustrated in Figure 4-2A. An atom with only s and p valence orbitals is chosen as the QM boundary atom and thus four hybrid orbitals are formed using one s and three p orbitals. Three of the hybrid orbitals participate in the normal QM calculation while the one left is strictly localized between the QM and MM boundary atoms, termed strictly localized bond orbital (SLBO) [145]. The respective coefficients for the four hybrid orbitals need to be parameterized and a -1e charge should be taken from the MM boundary atom as it donates 1 electron to the SLBO. Therefore the total QM energy comprises the energy of regular QM orbitals, the SLBO and the coupling between them. The gradient is calculated accordingly. As an update, the SLBO is then parameterized as a classical potential for different types of bonds [146].

Based on the LSCF procedure, Friesner and co-workers [15] [147] introduced more extensive parameterization in terms of electrostatics, van der Waals interactions and hydrogen bonds at the boundary instead of taking the parameters directly from the MM force field. Their parameterization is sensitive to different QM methods and basis sets. Contrary to the LSCF scheme, the frozen orbital can also be centered on the MM boundary atom instead of the QM boundary atom, which is termed the generalized hybrid orbital (GHO) method by Gao and co-workers [5-7, 148]. In this approach, one of the hybrid sp^3 orbitals participates in the QM calculation while the other three are kept frozen as shown in Figure 4-2B. The force field charge of the MM boundary atom is equally distributed over the four hybrid orbitals as opposed to the LSCF which has to calculate the charge density of the frozen orbital for different systems. This advantage ensures the transferability of the GHO. However, the classical potential parameters

involving the MM boundary atom, especially its bonding parameters, have to be re-optimized to accommodate the effects introduced by the three frozen orbitals. Or the interaction between the frozen orbitals and other normal QM orbitals need to be adjusted by scaling the integrals.



A **B**

Figure 4-2: Illustration of LSCF and GHO schemes

Orbitals are shown in blue spindles and those participating in the QM calculations are shadowed. A) LSCF scheme B) GHO scheme

4.3.4.3 Performance of LA and FLO: summary

The link atom scheme is advantageous for its simplicity while the frozen local orbitals scheme is convincing for its solid physical foundation. Their application and performance have been the concern of QM/MM investigators, who have made comparisons between these two methods. Karplus and co-workers [142] compared the LA with the LSCF with regard to the proton affinity and deprotonation enthalpy of propanal. In their test, there were two types of LA schemes, one that includes the electrostatic interaction between the link atom and the MM charges and the other one without. And the one with LA-MM charge interaction performed better, especially when a Na^+ ion was placed in the vicinity of the hydroxyl group. Upon calculation of the Mulliken charge of the link atom, they found that without including LA-MM charge interactions, the charge on the link atom is significantly greater than usual, hence biasing

the charge density of the rest of the QM orbitals. This finding indicates that polarization of the link atom by MM charges is important. The LA scheme with the LA-MM charge interaction was then compared with the LSCF scheme and they gave similar results. However, when a Na^+ ion is present in the neighborhood of the hydroxyl group, the LA was consistently better than the LSCF, suggesting that the assumption of a strictly localized bond orbital is not proper when the QM subsystem is strongly polarized by charges. As for geometry optimization of a tripeptide, the LSCF performed better than the LA, as is conceivable because of the geometry constraint of the link atom mentioned above.

Mulholland and co-workers compared the LA with the GHO with regard to the reaction mechanism of a virus protease [149]. In their LA scheme, the LA-MM charge interaction was included. With the same boundary, they found that the free energy barriers and reaction free energy differed by 8 kcal/mol and the locations of the reactant, product and transition state were rather different as well. The difference was attributed to the rotation of C_α-C_β of an aspartic residue that can form a hydrogen bond with the adjacent histidine residue and thus stabilize the reaction intermediate. When the LA was employed, the hydrogen bond was found to be broken, thus significantly destabilizing the intermediate. On the other hand, this did not occur for the GHO. This discrepancy is thought to be due to the one extra proton and more delocalized orbitals introduced by the link hydrogen atom. Be that as it may, it could also be caused by the improper partitioning of the system. For instance, if the link atom had been placed farther away from the rotational bond rather than on that bond, the result could have been improved. The reason why the link atom was positioned the same as the GHO was just for the convenience of the comparison and the GHO is only available for sp^3 hybridized carbon. Therefore the discrimination might have arisen from the way the system was partitioned as well.

As a matter of fact, the choice of the boundary placement does influence the result remarkably as verified in [141, 142, 149] . Moreover, when the size of the QM subsystem is enlarged, the result is consistently improved [141]. So the rule of thumb of partitioning the system is to place the boundary as far from the reaction center as possible and avoid boundary positions that can directly affect the active space as happened in [149]. The capped bond is preferred to be a C-C bond in order to circumvent the overpolarization by other, more intensely charged, MM boundary atoms.

4.3.4.4 Other boundary schemes

Besides the LA and FLO approaches, a boundary-atom scheme has also been implemented, in which the MM boundary atom is treated as a special boundary atom to cap the dangling bond at the QM frontier, thus relying on elaborate parameterization for different frontiers. Among them, we highlight Zhang and Yang's pseudobond approach [150, 151], Poteau and co-worker's effective group potential approach [152, 153] and DiLabio and co-worker's quantum capping potential treatment [154]. Notably, the quantum capping potential treatment has been applied by Salahub and co-workers [18] to study electron paramagnetic resonance and obtained consistently better results than the single link atom scheme. For a comprehensive review of this class of schemes, one can refer to Senn and Thiel's review [115]. Moreover, instead of fixing the boundary, it can also be adaptive during the calculation [155-157]. This adaptive scheme allows atoms to change between QM and MM subsystems and in principle should also allow charge transfer across the boundary once fully developed.

4.4 QM/MM optimization techniques for potential energy surfaces (PES)

A potential energy surface is typically a rugged landscape marked by various valleys and peaks. Among others, we are most interested in the stationary points: minima and saddle points. In this section, we will introduce the techniques for finding these points and the paths between them. We will start with geometry optimization methods to find the minima on QM/MM potential energy surfaces.

4.4.1 Geometry optimization

In general, QM calculations adopt quasi-Newton methods to optimize geometries. These estimate the Hessian matrix by gradient differences and then update it in various ways. In the case of MM optimization, the widely used methods are not only second-order methods but also first-order methods such as conjugate gradient and steepest descent methods. In the current context of QM/MM optimization, our focus will be on techniques specific to QM/MM rather than optimization algorithms in general.

In principle, the whole QM/MM system can be simultaneously optimized with a uniform optimizer using the QM/MM potential and gradient. Convergence, however, will be difficult to reach when the starting geometry is far from the minimum. Moreover, there are also technical issues if the subsystems are optimized at the same time. First, as just mentioned, QM and MM calculations prefer different optimizers. If one optimizer is chosen for both subsystems, the efficiency of optimizing either subsystem will be compromised and the same optimization parameters such as the trust radius may lead the configuration to an undesired space. Second, QM and MM optimizations are usually conducted in different coordinate systems. QM minimization often employs redundant internal coordinates or internal coordinates because of difficult convergence with strongly coupled Cartesian coordinates. Conversely, MM

minimization prefers Cartesian coordinates to avoid the laborious transformation between Cartesian and redundant internal coordinates.

4.4.1.1 Microiteration

Considering the different natures of QM and MM methods, optimization is easier to run separately for each subsystem on their respective levels. To this end, a macro/microiterative scheme has been proposed. There are two variants of this scheme, one termed the adiabatic scheme [12, 16, 158-162] and the other the alternating scheme [2, 4] . In the adiabatic scheme, the optimization is driven by the QM optimizer. The MM subsystem is optimized to convergence with the QM part frozen, and this is termed a microiteration. Thereafter, the QM region is optimized till convergence with the MM part frozen, and this is termed a macroiteration. These two iterations alternate until the whole system is fully optimized. In the alternating scheme, after the MM region is optimized, only one optimization step is taken in the QM region and it switches back to MM optimization again, thus iterating until both are fully optimized. Since the QM region is of principal interest and thus its optimization serves as the main driver, the adiabatic scheme is chosen more often. A diagram of both schemes is shown in Figure 4-3. Following this “divide and conquer” philosophy, the advantage of QM/MM partitioning is thoroughly exploited. First, the number of expensive QM energy and gradient calculations can be dramatically reduced. Second, the costly coordinate transformation is avoided in the MM optimization.

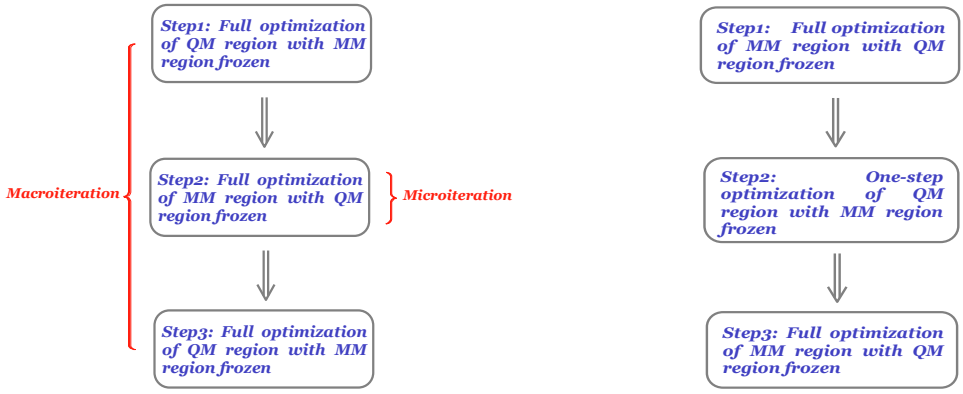


Figure 4-3: Diagrams of the adiabatic scheme and the alternating scheme
A) the adiabatic scheme B) the alternating scheme

However, it is oftentimes harder to unite than to divide. One encounters a problem when calculating the QM-MM coupling during the optimization. In the microiteration of optimizing the MM subsystem, even though the QM part's geometry is frozen, the electrostatic QM-MM interaction should be calculated on the QM level for electric embedding. So, ideally, a full QM SCF calculation should be run at every MM minimization step (termed S1 herein) to obtain the forces on the MM charges exerted by the QM atoms. This procedure is prohibitively expensive considering the fact that an MM minimization usually takes thousands of steps. Approximation comes into play at this point.

Yang and co-workers proposed to use the electrostatic potential (ESP) charges to represent the QM charge density and approximate the QM-MM electrostatic force on the classical level during the microiteration [31]. The gradient is calculated as

$$G_{el}(\text{QMS-MMS}) = G_{ESP} \qquad (17)$$

This treatment leads to a discrepancy between the gradient and the energy as the energy for the entire system still retains the electrostatic interaction on the QM level. To alleviate this double-

standard deficiency, Friesner and co-workers [16], Thiel and co-workers [162] introduced an ESP-based QM-MM interaction as a correction of the QM-based electrostatic interaction. Before the MM minimization steps, the *ab initio* calculated gradient on MM charges is calculated as G^0_{QM}(QMS-MMS), the ESP-based gradient G^0_{ESP} and the ESP charges are retained throughout the MM minimization. At each MM optimization step, the ESP-based gradient G_{ESP} is re-calculated (termed S2 herein). Therefore, the electrostatic QM-MM gradient is formulated as

$$G_{el}(\text{QMS-MMS}) = G^0_{QM}(\text{QMS-MMS}) + (G_{ESP} - G^0_{ESP}) \qquad (18)$$

The corresponding energy for MM minimization [161] is

$$E_{el}(\text{QMS-MMS}) = E^0_{QM}(\text{QMS-MMS}) + (E_{ESP} - E^0_{ESP}) + [G^0_{QM}(\text{QMS-MMS}) - G^0_{ESP}] (R_{MM} - R^0_{MM}) \qquad (19)$$

where E^0_{QM}(QMS-MMS) is the total energy corresponding to G^0_{QM}(QMS-MMS), E_{ESP} the total energy calculated with the ESP model, and R the coordinates of the MM atoms.

As (17) is improved by the perturbation in (18), the gradient now becomes consistent with the energy (19). The ESP charges were then further improved by a 1SCF procedure by Lluch and coworkers [159]. They assumed the QM wave function to be frozen during the MM minimization and thus only the electron-MM charge and nuclei-MM charge interactions need to be calculated because the other terms in the QM Hamiltonian stay the same. Therefore, only one SCF calculation is performed at each MM minimization step to evaluate the QM-MM electrostatic coupling (termed S3 herein).

Nonetheless, freezing the QM wave function during the MM minimization is not quite convincing as it should be polarized in reality. Morokuma and co-workers [158] proposed to re-calculate the wave function when the MM subsystem is fully optimized (termed S4 herein). Thereafter, the MM minimization is started again with the newly polarized QM densities.

Moreover, they also suggested using a fast multipole method instead of the ESP charge model for the QM-MM electrostatic interaction.

To evaluate the different optimization schemes above, Thiel and co-workers [162] first tested whether the QM wave function should be re-calculated after the MM subsystem is optimized. They tested S2 and (S2 + S4) with a water cluster and ESP charges and found the optimized energy by S2 alone was lower and reached with fewer iterations, though at a slightly different geometry. They then compared S1, (S2 + S4) and (S3 + S4), and found (S3 + S4) yielded the lowest optimized energy and second-fewest QM calculations after S1. It seems that S2 gives the fastest convergence whereas (S3 + S4) produces the best accuracy.

In addition to the schemes above, Moliner and co-workers have implemented a dual level scheme for QM/MM optimization [160]. They use semi-empirical methods to calculate the QM-MM electrostatic coupling in the MM minimization [160]. A similar approach has been proposed by Warshel and co-workers to introduce a reference potential using the empirical valence bond (EVB) method [8].

4.4.1.2 Macroiteration

So far, we have discussed the QM-MM coupling concerning the MM minimization. A subsequent question would be: Is this coupling a problem in the QM minimization? In fact, it is not a problem in the gradient calculation, but it is a disturbing one in the Hessian update. In the widely used adiabatic scheme, a QM optimization step is taken after the MM region is optimized and hence the QM Hessian should be updated from the last one. Although the gradients only relate to the current geometry and wave function, the Hessians are decided by the gradient difference between the previous step and the current one. Contributions to the gradient change include the QM wave function and the MM coordinates as well because the latter are altered

during the MM minimization. Therefore, the gradient change caused by the MM coordinate changes should be eliminated. To this end, Morokuma and co-workers have incorporated a quadratic QM-MM coupling [161] in the macro-iteration that is realized by coordinate transformation and Hessian manipulation. They managed to demonstrate its moderately improved performance for small molecules and better convergence behavior than schemes without the quadratic coupling.

4.4.1.3 Convergence criteria

The adiabatic scheme in principle requires a completely optimized MM region to obtain a good QM convergence behavior. To ensure this, the ratio of the QM optimization convergence criterion to that of MM optimization should be appreciable, as it was set to 10 in [162]. Moreover, there is no presumption in the macro-iteration that the forces on the MM atoms are exactly zero, so less tight convergence criteria can be used in the micro-iteration [161, 162] .

4.4.1.4 Size of QM region and starting geometry

As we have stressed throughout this article, the size of the subsystems greatly influences the results. As demonstrated in [159] , when using the S3 scheme, the total CPU time for optimization does not increase monotonically with the size of the QM region, which indicates that there is a medium QM size leading to a minimum CPU time for optimization. When the MM environment configuration is complicated and hard to converge, a larger QM core can be chosen as this leads to less QM-MM coupling and hence fewer micro-iterations. However, with the ESP charges, the larger the QM core, the greater the error caused by the charge screening effects as the dielectric constant is unknown in the QM region.

Different starting geometries could well result in different minima, especially for complicated proteins [163]. This indicates the importance of sufficient sampling and averaging

as a single optimized geometry is not quite meaningful. This is especially important in reaction path calculations.

4.4.2 Transition state search on the potential energy surface

Having located the minima on the energy surface, we come to a natural question: How are these minima connected with each other? Chemically speaking, this question equals: How to find the reaction path between the reactant and product? According to the transition state (TS) theory, there is always a transition state(s) on the path from the reactant to the product. This transition state is a first-order saddle point on the potential energy surface, which has a negative eigenvalue in only one direction. If we can identify the TS, the reaction rate can be calculated according to the transition state theory. Knowing the TS, an intrinsic reaction coordinate (IRC) method [164] is usually adopted to draw out the reaction path to understand the reaction mechanism. However, the TS is often unknown for complicated reactions, e.g. enzymatic reactions. Hence searching for the TS becomes a principal task for the study of chemical reactions.

Ideally, the TS can be found via an eigenvector following approach based on the first-order saddle point nature of the TS. However, this procedure requires a good initial guess when starting from the vicinity of the TS, which is often impractical, especially for high dimensional systems. Therefore, the search for the TS is often combined with finding the reaction path that connects the reactant and product minima. As soon as the reaction path is found, the TS is evidently identified as the highest point along the path.

Similar to the philosophy of reference [22] , del Campo and Köster have proposed a hierarchical TS search algorithm [165]. They first use the saddle method similar to reaction coordinate driving, which sequentially optimizes to zero the system's force perpendicular to the

path, so as to bracket the TS between two highest points. The subsequent TS finder adopts the uphill trust region method, which constrains the step to ascend the potential energy surface in the direction of the normal mode associated with the TS vector and to descend in the remaining normal modes. This hierarchical approach has been effectively implemented in deMon2k [166] and has proven to work well for TS searches with QM methods. Its further incorporation of QM/MM methods is of great interest. To obtain an initial guess in a simple way, instead of the elaborate saddle method, the multicoordinate driving (MCD) scheme is proposed by Berente and Naray-Szabo [167] . Their method differs from the regular Reaction Coordinate Driven (RCD) method by including multiple reaction coordinates. And it has been tested with a QM method for hydrolysis by dUTPase [167].

In the following section we discuss the Minimum Energy Path technique. More complex methods, involving more or less extensive sampling in many coordinates are described in Section 4.5.

4.4.2.1 Minimum energy path (MEP)

Identification of a reaction path relies on the definition of reaction coordinate(s), which is often based on one's chemical intuition. Once the reaction coordinate (RC) is chosen, the rest is just to determine the system configuration along the RC. To this end, a simple way is to optimize the geometry at different RC, thus forming a minimum energy path (MEP). Thereupon, all the minimization techniques can be employed in this approach. And the foresaid macro/micro iterative optimization scheme can also be fully utilized for the MEP search on the QM/MM potential energy surface.

One simple method to find the MEP is the Reaction Coordinate Driven (RCD) approach [168]. In this approach, the RC is changed step wise and the geometry is optimized in every step with the RC kept frozen. This method has been tested by Frischer and co-workers with a QM/MM potential for a proton transfer reaction in an enzyme [169]. The RCD led to a TS with unrealistically high energy and produced discontinuities along the path. This happened because the frozen RC during the optimization over-parameterizes the reaction path and drags the system to higher points than the true TS. Moreover, this method is also inefficient due to its sequential walk along the RC.

4.5 QM/MM approaches to the simulation of kinetics and thermodynamics in condensed phases

The main goal of this part of our review is to provide a comprehensive description of methods used in molecular simulations for studies of chemical processes in condensed phases. Similar to classical simulations, one can use a QM/MM energy function to perform Monte Carlo (MC) and molecular dynamics (MD) simulations. First, let's focus on the classical propagation of nuclear dynamics. In this case, one can use QM/MM-derived gradients (forces) as a part of a standard integration scheme. QM/MM correction to the full Hamiltonian of this system will provide important information on electronic degrees of freedom relevant to the problem at hand. Since we can impose a canonical distribution of states with a known energy function (classical or QM/MM), it is also possible to perform QM/MM Monte-Carlo with the Metropolis algorithm. Unlike the examples in the sections above, a single structure (for example minimized) does not bear substantial significance of its own, and one will have to get an ensemble of structures to get proper averages and then to evaluate thermodynamic functions. It was shown for many systems that interaction energies evaluated from high-level *ab-initio* computations do not always provide

accurate descriptions for chemical processes that happen in a solvent and at finite temperature [170].

A good illustration of the need for extensive sampling over the entire configurational space may be found in studies of ion binding to membrane proteins. Yu and Roux [171] have examined the distribution of states from MD simulations in ion binding to membrane proteins with classical and polarizable force-fields and compared it to high-level *ab-initio* computations. It was found that, although high-level ab-initio structures may represent "true" local minima, an accurate estimate for ion binding may only be obtained from the analysis of a distribution of states. They analyzed simplified toy-model systems consisting of a monovalent ion (Na^+ or K^+) coordinated by 8 waters or N-methylacetamide molecules. They were able to show that because the PES for these systems is very complex and multiple local minima exist, a simple minimization will not provide a reliable estimate for the thermodynamics of ion binding to proteins. The problem gets much worse if we consider metal binding to proteins and all degrees of freedom available to the system. It was also found that the harmonic approximation is insufficient in this case and ensemble averaging is required to understand the complex thermodynamics of ion binding to proteins. This finding underlines the role of thermal fluctuations and overall protein flexibility in the modulation of ion binding to proteins and complex compounds. Their findings are in excellent accord with the conclusions of QM/MM MD simulations reported from the Guidoni and Klein groups [172, 173].

Bucher at al. [174] were able to show the importance of local charge transfer and electronic polarization effects, thus providing a welcome correction to the results of classical simulations. In their study of K^+ and Na^+ binding to the KcsA channel, the selectivity filter of the protein was represented as the QM region and the rest of the system was described with MM [89,

174]. The electronic structure was investigated using the maximally localized Wannier function centers of charge and Bader's atoms-in-molecules charge analysis. The results obtained were able to outline polarization effects on the channel backbone carbonyls and significant charge transfer from the backbone to the ions.

Detailed technical information on how to run these simulations is outside the scope of this chapter but it is worthwhile to provide a short summary for the interested and engaged reader. Major technical challenges are well understood e.g. the need for long-range electrostatic treatment and the introduction of periodic boundary conditions for studies of processes in condensed phases, the accurate introduction of thermostats into the system.

The cutoff schemes adopted in standard force-field simulations may or may not cause problems in QM/MM calculations [175-177]. Hu and Wang [109] showed that long-range corrections to the electrostatic interactions for systems with cut-offs larger than 14 Å play no significant role in the dynamics of the QM part. Nevertheless, a correct account of long-range corrections to electrostatic interactions may be important, if not critical, for studies that involve polyelectrolytes such as DNA or RNA molecules. To address this problem, an extension of the particle-mesh Ewald (PME) method has been developed by York and co-workers for QM/MM calculations under periodic boundary conditions [175]. They use conventional point-charge interactions and a reciprocal space component for the MM part and a real-space multipole expansion for the QM region. The method enables the partition of the total Ewald potential into a short-ranged real-space interaction and a long-range periodic correction. To compute the periodic correction term one requires only a Mulliken charge representation of the charge density (or any other method to map the charge density) and hence it can be used with any efficient linear

scaling Ewald method for point charge (or multipolar) systems, such as the particle mesh Ewald method.

It is evident that the performance of this method relies substantially on the number of quantum atoms being sufficiently small. The electrostatic energy has to be efficiently evaluated at each SCF iteration by a Fock matrix multiplication with the charge vector. Although very robust approaches to the problem of PBC simulations with infinite cutoffs exist, this method is likely to be very expensive for large QM regions. A convenient alternative to PBC simulations may be found in schemes that further reduce the dimensionality of the system. A good example of such a scheme is the Generalized Solvent Boundary Potential (GSBP), where the system is divided into inner (explicitly represented) and outer shells (described implicitly by solving the Poisson-Boltzmann (PB) equation for an external field imposed on the inner shell) [177, 178]. In this approach, all atoms within the inner region (usually a 20 Å sphere) are treated explicitly while the outer environment is represented as a solvent potential field. This procedure has been successfully implemented with the self-consistent charge tight binding DFT (SCC-DFTB) method and applied to pK_a calculations [176] by Cui and co-workers. More recently, Benighaus and Thiel extended this method with a semi-empirical approach and further proved its validity [179]. Moreover, they stressed the fact that, despite the success of the GSBP QM/MM approach, special care needs to be taken as to the physical parameters of GSBP such as the size of the inner region and the number of basis functions for the reaction field evaluations (Legendre polynomials), which might be highly system dependent. Nevertheless, all these corrections are important if one is to study chemical processes in condensed phases.

Therefore, MD and MC simulations may provide very useful insights into dynamics and thermodynamics of the system. However, to make an important step towards linking theory and experiment one has to compute observable properties. One of the most important thermodynamic properties to deal with is the free energy or relative free energy underlying chemical reactions or ion partitioning or any other chemical process. It is possible to use molecular simulations with QM/MM energy functions to compute directly the free energy for the process. The methods for evaluating Gibbs or Helmholtz free energies described above dealt mainly with the finding of transition states and thus reaction paths with the assumption that only a relatively small number of degrees of freedom are important. Although mighty and very successful for small systems where sampling of all important degrees of freedom are readily available, direct computations of free energies encounter significant difficulties if one wants to account explicitly for collective degrees of freedom due to the environment and its effect on the reaction path.

4.5.1 Free energy simulations and the QM/MM formalism

The benefits of direct computations of free energies are evident; they can be directly compared to measurements of reaction quotients and equilibrium constants. Before discussion of the QM/MM FE simulations, it might be useful to provide a general overview of theoretical foundations for the evaluation of the free energy from molecular simulations and its extension to QM/MM Hamiltonians.

4.5.1.1 PMF evaluation with thermodynamic integration (TI)

A common ingredient in all free energy simulation techniques is the use of an effective potential W that corresponds to reversible thermodynamic work done by the average force acting on the system. This thermodynamic function is also commonly referred to as a potential of mean force. The PMF can be evaluated along a chosen reaction coordinate providing a unique metric

for the free energy along the chosen transformation path. The introduction of the PMF as a measure of the free energy change is significant, since we can start working with forces along the pathway without the requirement for accurate energy computations for each pathway point. For example re-organization of water molecules far away from the reaction center may have a significant impact on the potential energy of the entire system and yet a minimal impact on the reaction path.

Theoretical foundations for PMF evaluation were initially developed by Kirkwood for the distribution functions in liquids. In his formalism to compute PMF, now known as thermodynamic integration (TI), one can introduce a dimensionless coupling parameter λ varying between 0 and 1. Let's illustrate its application for an ion solvation problem in the classical approximation. The state with λ=1 is the normal system with all interactions between ion and solvent turned on and the state with λ=0 is a reference state in which all interactions between ion and solvent have been turned off. The PMF is a function of collective coordinates ($W(\mathbf{r_1},\mathbf{r_2},\ldots)$) and characterizes a difference in free energy for these states which can be written as the ratio of two respective partition functions:

$$e^{-W(\mathbf{r_1},\mathbf{r_2},\ldots)/k_BT} = \frac{\int d\mathbf{X} e^{-E(\mathbf{r_1},\mathbf{r_2},\ldots,\lambda=1)/k_BT}}{\int d\mathbf{X} e^{-E(\mathbf{r_1},\mathbf{r_2},\ldots,\lambda=0)/k_BT}} \qquad (20),$$

or in the more convenient form of the thermodynamic integral of Kirkwood:

$$W(\mathbf{r_1},\mathbf{r_2},\ldots) = \int_0^1 \left\langle \frac{\partial E}{\partial \lambda} \right\rangle d\lambda \qquad (21).$$

The PMF in the equation above does not contain any mass terms and thus is an equilibrium thermodynamic function independent of time-scale provided convergence and quality of the

force-field/QM basis-set/functional etc. It is clear that these criteria apply to all *E(λ)* averages over the integration path.

Free energy simulations have been extensively developed for classical simulations, but they have also started to appear in QM and QM/MM studies. The starting point for calculation of the PMF is the definition of the partition function for the ensemble. The canonical partition function for a system described with the QM/MM Hamiltonian can be expressed as [180]:

$$Z_O = \int \mathrm{e}^{-E(\mathbf{r}_{QM},\mathbf{r}_{MM})/k_B T}\,\mathrm{d}r_{QM}\,\mathrm{d}r_{MM} \qquad (22)$$

where E is the total energy and a function of collective coordinates for atoms in QM and MM parts ($\mathbf{r}_{QM}$ and $\mathbf{r}_{MM}$), respectively. Some degrees of freedom similar to all states with different λ could be effectively averaged out and the reaction path can be defined for a smaller system sub-set (often for only the QM part of the system).

4.5.1.2 Free energy perturbation (FEP) techniques

Although powerful, the use of TI implies a continuous integration with respect to λ, which might be problematic in many situations. In 1954, Zwanzig introduced a free energy perturbation theory that relates the free energy difference between state A and state B to the potential energy difference between these two states [181]. If we assume that the perturbation required to transform system A to system B is small (<2 kT), it can be shown that:

$$\Delta G_{(A->B)} = -k_B T \ln\left\langle \exp\left(-\frac{E_B - E_A}{k_B T}\right)\right\rangle_A \qquad (23),$$

where the potential energy difference between two states is weighted by the energy of the initial state. In case of larger perturbations, one can always use additional windows to connect starting

and ending points of this perturbation. The free energy is a path-independent property of the system and can be evaluated regardless of how "alchemical" the perturbation path may be. Similar to TI, FEP can easily be extended to accommodate a QM/MM description of the system since the partition function of the system can be specified. Reddy et al. [182] provided an intuitively appealing formulation of FEP for QM/MM simulations of enzymatic reactions and solvation. The Hamiltonian describing the system which is changing from state A to state B during FEP calculations can be re-written as

$$H(\lambda) = \lambda E_A + (1 - \lambda) E_B \qquad (24),$$

where E_A and E_B represent distributions of states for the two end-points of the perturbation. The dynamics of two replicas of the system each corresponding to endpoints with λ=0 and λ=1 is treated explicitly and simultaneously and appropriate weighting (see eq 21) is applied to reconstitute the Hamiltonian for intermediate windows. This method is known as a dual-topology FEP method reflecting the simultaneous presence of two end-point systems. The dual-topology FEP method is especially useful for calculations such as solvation [182], studying the effects of mutation and pKa values in a protein [176, 183], ligand binding [184] or enzymatic reactions etc [185]. An interesting area of potential application of QM/MM FEP methods is the study of ion solvation and potentially mechanisms of selectivity in ion channels or ion-coupled transporters. Ion selectivity has been extensively studied at the MM level with both classical and polarizable force-fields [170, 186-191]. However, classical simulations may be compromised by their inability to account for charge transfer and electronic polarization, thought to be critical for ion binding to proteins. We have extended QM/MM FEP to studies of Na^+/K^+ solvation and selectivity by water clusters with variable numbers of ligands [53], of which results are illustrated in Figure 4-4. In that article QM/MM FEP calculation have been performed, for ion-

water clusters with different numbers of water molecules, where the ion was treated as the QM region using B88-lyp with the DZVP basis set implemented in deMon, and water were represented by the polarizable Drude force-field implemented in CHARMM. The same approach can be used for selectivity calculations in more complex biological systems. Also, the close connection between QM/MM calculations and those with polarizable force-fields developed to account for electronic effects is visible. Therefore, the polarizable force fields can be sufficient for some cases.

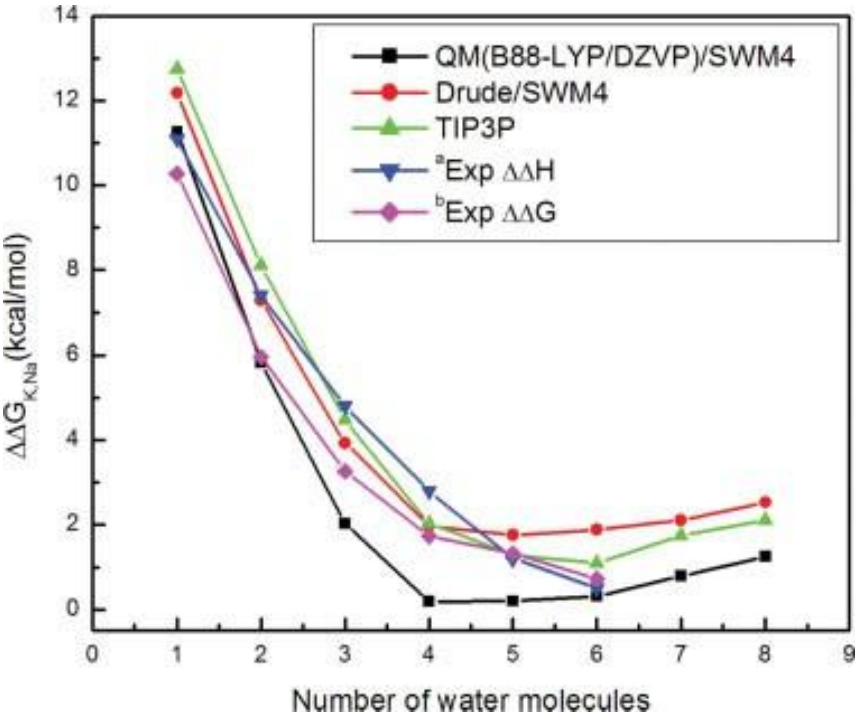


Figure 4-4: Relative (to the bulk) free energy of selectivity for Na+/K+ in water clusters as a function of cluster size

Experimental information is taken from a. [192] and b. [193]

Although QM/MM FEP is significantly more expensive than classical FEP simulations, the ability to compute free energies (or relative free energies) provides a unique opportunity for validation of classical potential functions as well as for relating QM computations at finite temperature to experimental measurements. Several attempts were made to reduce the number of degrees of freedom in FE QM/MM simulations. Yang and co-workers introduced the QM region into sampling via ESP charges included in the energy expression used to obtain the forces and to propagate the system along the reaction coordinate. At the beginning of the sampling, a full QM/MM calculation is run to obtain the reference potential and the energy difference caused by the MM configuration change is estimated as coulombic charge-charge interactions, which is then added to the reference potential to approximate the system's potential [4]. This reference potential approach, as mentioned before, has been employed as well in QM/MM geometry optimizations [8, 160]. Since this reference potential is sensitive to the MM configuration, it is more reasonable to use an average electrostatic field over an MM ensemble as done by Yang and co-workers [180]. The same authors also found that a short MD simulation was usually enough to obtain a consistent reference potential so that a self-consistent procedure could be avoided. The PMF profiles have been obtained for all methods and have been compared. The results have proven to be very close to each other. A full QM sampling is rather expensive for regular QM/MM calculations. Yang and co-workers [180, 194] have simplified the reaction path search on the FEP to that on the PMF surface e.g. using partial derivatives of the PMF along the reaction path with respect to position. To remove an apparent need for extensive and expensive sampling of the QM part of their system, they froze the QM part and used the obtained ESP charges to evolve their system in classical (MM) space. To render out the minimum free energy path, the QM free energy gradient is employed as a criterion in the QM geometry optimization.

Lastly, the above steps must be iterated to reach self-consistency.

A similar approach to the study of solvation effects has also been used by Warshel and co-workers [195]. To reduce the conformational space in FEP simulations, they kept the QM region frozen. Detailed comparisons between different schemes for the performance of free energy techniques applied to studies of enzymatic reactions have been made by Senn and Thiel [19, 116]. The major conclusion was that, if appropriate sampling is achieved, estimated activation barriers and reaction thermodynamics may be described very accurately by either TI, FEP or Umbrella Sampling techniques (see below).

TI, FEP and other similar techniques imply the existence of a well-defined reaction coordinate (reaction path, permeation pathway or conformational pathway). What if we don't have a pre-conceived idea about the reaction path? Providing a comprehensive sampling of free energy surface, one can find the pathway post-factum. Extensive sampling along the reaction coordinate may allow the complete removal of the dependence of the results on starting configurations. Several simulation techniques have emerged recently to address the problem of efficient sampling and are often referred to as "enhanced sampling" techniques.

4.5.2 Enhanced sampling techniques

4.5.2.1 Multiple time-step (MTS) approaches

The conceptually simplest approach to enhance sampling in molecular dynamics simulations would be to introduce a larger time-step. The time-step in classical simulations is defined by an integrator and the apparent need to integrate the fast dynamics of covalent bonds. It is usually set to 1 or 2 femto-seconds. The multiple time-step (MTS) method was first proposed for molecular mechanical dynamics to separately treat motions of high and low frequencies in the same system [196]. The MTS method uses a smaller time-step Δt for the fast

motion and $n\Delta t$ for the slow motion. The fast degrees of freedom are first advanced for n steps at step size Δt with the slow degrees of freedom fixed and the latter is then updated with a step size $n\Delta t$, for which n was found to be 5 ~ 10 to sample effectively [196]. As an analogue to the fast motion, the MM region possesses more configurational variability than the QM part, thus demanding an enhanced sampling. Wei and Salahub [11, 197] have adopted this MTS approach to study solvation effects on a QM water molecule, in which a step size of 1 fs was used for the MM region and 15 fs for the QM water. While this enables oversampling of the MM region, the conventional MTS is liable to yield biased sampling due to the frozen slow degrees of freedom when propagating the fast ones. To remedy this drawback, Tuckerman et al proposed reversible reference system propagator algorithms (RESPA) [198] and implemented them within the CPMD program package [199]. This methodology has been adopted by Woo et al in their QM/MM MD scheme [200]. In their work, the MM part is first propagated from t_0 to $(t_0 + n\Delta t/2)$ at the step size of Δt. The QM part is then evolved from t_0 to $(t_0 + n\Delta t)$ at the step size of $n\Delta t$ with the MM force averaged from the forces at $(t_0 — n\Delta t/2)$ and $(t_0 + n\Delta t/2)$. Following this, the MM part then moves from $(t_0 + n\Delta t/2)$ to $(t_0 + n\Delta t)$ at the step size of Δt with the QM force averaged from the forces at t_0 and $(t_0 + n\Delta t)$. Hence, the problem caused by fixed degrees of freedom is partially solved by force average and smaller propagation steps. Moreover, the MM part is further extensively sampled by assigning smaller masses to the MM atoms, resulting in faster motion.

4.5.2.2 Umbrella sampling (US)

The US method offers a simple method of configurational sampling for processes with activation barriers. Instead of using an order parameter to force the transition between two states in TI and FEP, umbrella sampling resorts to a biasing (restraining) harmonic potential (V_b) to

overcome the transition barrier and thus enhance the sampling [201]. In the US method the ensemble average <A> of a function A can be expressed as [202]:

$$< A >= \frac{< A / e^{-\beta V_b} >_b}{< 1 / e^{-\beta V_b} >_b} \qquad (25)$$

where β is the Boltzmann factor, the brackets emphasize the ensemble average over the biased non-Boltzmannian distribution defined by exp(-β (H+V_b)), that includes the biasing potential V_b and the system's Hamiltonian H. It is apparent that one can use either a classical or a QM/MM expression for the Hamiltonian of the system.

With the biasing potential, the high-energy and low-probability regions of the phase space are better sampled. Several different implementations of this algorithm have been developed. For computing the PMF, a US simulation would ideally lead to a uniform distribution along the reaction coordinate. The difficulty in practice is to determine the strength and spacing for placement of umbrella potentials along the reaction coordinate and the common practice is to use multiple trial-and-error assessments. This represents a clear challenge to multi-dimensional PMFs involving costly computations such as QM/MM simulations. To overcome this problem, the classical US scheme was extended to develop an adaptive US algorithm by early work of M. Mezei [203] and then was extended by a number of research groups [202, 204]. In the adaptive US algorithm the biasing potential is adapted to PMF information extracted from the preceding US window and the analysis of the reaction pathway is usually carried over with the weighted histogram method (WHAM) [205]. Rajamani et al. have reported an application of adaptive-US for sampling of multi-dimensional PMFs for the proton transfer reaction in [NH_3-H-NH_3]$^+$ in water [202]. Aside from its apparent methodological importance, this study reports on the quantitative assessment of solvent effects on the rate of the proton transfer reaction. The

presence of solvent molecules leads to an increase in the free energy barrier by ~5 kcal/mol. This large adjustment in the barrier height was related to charge delocalization for the transition state as compared to products and reactants. Furthermore, the explicit account of solvent effects leads to changes in the shape of the two-dimensional PMF profiles for this reaction as compared to the gas-phase.

4.5.2.3 Replica Exchange

Conventional QM MD simulations are not capable of sampling rare events because of their prohibitive expenditure for long-time dynamics. This barrier in turn hinders QM/MM dynamics and hence, an efficient sampling method is desirable. Parallel tempering, also known as replica exchange, is originally an MD approach to simulate replicas of the system simultaneously and exchange their configurations at different temperatures. One may think about ensembles of replicas as an example of a Markov chain of states. That is, two conformational states in an ensemble can be understood as the state of the ensemble before and after a pair of replicas (i,j) have exchanged their respective configurations. It is possible to then use the Metropolis algorithm to determine the exchange probability and set acceptance criteria[206]. This exchange enables replicas at low temperatures to access regions that are hard to sample in phase space. As recently reviewed by Earl and Deem [207], parallel tempering has been widely used in MM simulations of polymers and proteins and QM simulations of clusters. The temperature-based version of the RE algorithm may become a double-edged sword, since there is always a possibility that the free-energy landscape is drastically different in the high-temperature region. There are numerous extensions of RE simulations that are being run at the same temperature, but contain a gradual perturbation of the potential energy in replicas similar to that of the FEP method [208, 209]. This is a promising alternative for parallel tempering simulation

of a quantum system, as the QM self-consistent field calculation is often difficult to converge at high temperatures. Li and Yang [210] have added a QM replica to the MM replicas in their Hamiltonian parallel tempering simulation with the module in CHARMM [211]. They formulated the exchange acceptance ratio to satisfy detailed balance and obtained a more complete sampling at a much shorter time length than the QM dynamics. As noted by the same authors, the QM potential can well be replaced by a QM/MM potential to increase the resolution. Several notable applications of QM/MM REMD are worthwhile mentioning and we direct interested readers to the particulars of systems in references [212-217].

4.5.2.4 Reaction Coordinate Driven (RCD) methods

An extension of enhanced sampling algorithms and particularly US-based routines may be found in the recent implementation of a chain-of-replica approach [118] for QM/MM from the Brooks group to allow for some flexibility along the reaction coordinate. In this approach, a tentative RC is first defined to connect the reactant and the product. And then a chain of geometries (replicas) differing only in their RC values is interpolated between the reactant and the product. A spring force is exerted between the replicas as a function of their RMS distance. Hence, the object function to be minimized is the actual QM/MM potential plus the spring potential of each replica. An intrinsic advantage of this replica path scheme is the trivial effort of parallelization as the replicas can be optimized separately from each other. Another benefit from this approach is its capability of discriminating the important atoms vs. the unimportant ones. Since the RMS distance is used for adjacent replicas, different weights can be assigned to critical atoms, which play more important roles than other atoms far from the reaction centre. Furthermore, the RMS distance can be used to calculate the potential of mean force evaluating forces along a specific path [118]. However, it is noted that the spring force constant should be

chosen carefully. Since there is also a restraint on the angle between replicas, this force constant demands a careful choice as well. Yang and co-workers have further expanded this promising approach with the introduction of the macro/micro iteration scheme into their chain-of-replica method [218]. The QM core and MM environment forces and energies along the path points could be computed and optimized separately by alternating iterations. As pointed out earlier, the coordinate system plays a role in the QM/MM optimization. Hence redundant and Cartesian coordinates are employed for QM and MM subsystems respectively, and the calculation of the distance between the points along the path excludes the MM degrees of freedom. To avoid the translation and rigid body rotation of the QM core, Cartesian coordinates of at least three core atoms are included as redundant internal coordinates. In contrast to this treatment, a rotation matrix to best fit the adjacent path images is used to ensure the same coordinate frame in the work of [118, 219] as only Cartesian coordinates are used. Another major difference in [218] is the application of the TS search scheme by Ayala and Schlegel [22] . An explicit TS finder is added to the regular replica path method, in which the highest point of the path is moved toward the true TS and the replicas are redistributed afterwards in each path relaxation cycle. Since this Hessian update requires gradients of the neighboring replicas, discontinuity of the environment along the path can seriously bias the updated Hessian. This problem is remedied to some extent by checking the Hessian after each update and reinitializing it with an empirical estimate when necessary, which possibly results in slower convergence.

As an alternative to the replica path method, a nudged elastic band (NEB) method has also been implemented [219]. As a member of the chain-of-replica methods, the NEB is similar to the replica path scheme in terms of replicated points along the reaction path to be optimized separately. The difference is that in NEB, the force perpendicular to the path is optimized below

threshold instead of the total force. Brooks and co-workers have tested both NEB and replica path with a QM/MM potential [220]. They found both methods converge at nearly the same rate when proper optimizers were chosen. While the replica path method can effectively use the adopted basis Newton-Raphson (ABNR) optimizer, the NEB requires a combination of steepest descent (SD) and ABNR, as the ABNR alone with the NEB is unstable owing to the projection of the forces.

A variant of the NEB method for QM/MM calculations has been proposed by Yang and co-workers [221] . Their approach differs from that of reference [219] in two respects: the optimization method and the definition of the distance between points along the path. In [221] , the path is optimized with the projected velocity Verlet algorithm. However, the convergence efficiency of MD minimization based on quenched Newtonian MD has been shown to be inferior to ABNR [219]. Pure minimizers should work better for optimization on the potential energy surface and they significantly reduce the number of expensive QM calculations. To avoid the floppy degrees of freedom in the MM environment, in reference [221] a transformation was used of a set of interatomic distances concerning all chemical or hydrogen bonds formed/broken during the reaction. However, this treatment could cause discontinuity of the environment when the environmental conformation is adjusted to different reaction coordinates. To alleviate this drawback, Yang and co-workers started from a reference system with a relatively rigid environment and gradually decreased the spring force constants for the environment. On the other hand, the assignment of different weights in [118, 219, 220] seems more straightforward although the weights should be carefully chosen nonetheless.

4.5.2.5 Transition path sampling (TPS)

Despite the fact that the above search methods include temperature effects, they still find only one path and one TS, which in reality should be an ensemble. To adequately sample the various possible paths, a transition path sampling (TPS) method has been proposed [222]. Schwartz and co-workers have applied the TPS in CHARMM [211] to study lactate dehydrogrenase using a QM/MM potential [62, 223, 224] . To distinguish the reactant, product and transition state regions, an order parameter, e.g. atomic distances in [62, 223, 224], is first defined. Restraining the order parameter, a biased QM/MM MD simulation of time length t (t=500fs in [224]) is run to obtain the initial trajectory from the reactant to the product. A time slice of the trajectory is then randomly chosen and its momenta are changed by a small amount while the total momenta and energy are conserved. Following this, the dynamics is run both forward and backward to complete a trajectory of t. If this dynamics arrives at both the reactant and the product regions, it is considered reactive and the subsequent dynamics starts from a time slice from it. Otherwise, when it is not reactive, another time slice is chosen from the old trajectory to continue the dynamics until a reactive one is obtained. Practically, the acceptance ratio of reactive trajectories depends on the momenta changes, which is adjusted to get a population of 26.5% in [62]. From this transition path ensemble, one can draw out the transition state ensemble by the definition of the order parameter [62, 223, 224]. Analyzing the trajectories and the transition state ensemble, one can identify important movements of the atoms surrounding the reaction centre as detailed in [62, 223, 224]. However, while inspiring qualitative conclusions have been reached [62, 223] [224], no quantitative result such as reaction rates have been calculated with QM/MM methods to compare with experimental observations, although the procedure to calculate the free energy barrier was already illustrated by Chandler and co-workers for classical case in [222] .

4.6 Beyond conventional QM/MM dynamics: explicit account of nuclear quantum effects

The free energy calculations described in the section above form the cornerstone of the study of chemical kinetics. Together transition state theory (TST) and reaction coordinate sampling techniques (Umbrella Sampling, FEP etc) enable computations of approximate rate constants in the Born-Oppenheimer approximation. A QM treatment is commonly used to treat electronic effects and the MM part provides a necessary account of environmental effects. Examples of the QM/MM molecular simulations are numerous and several excellent reviews were recommended above. Thus far, our chapter has focused as well on the latest state of the art methods developed to accurately compute free energies, with emphasis on the QM nature of the PES used to drive classical nuclear dynamics. However, it is accepted [225] that nuclear quantum effects, such as tunneling, play an important role, in particular, in enzymatic catalysis. Tunneling is important in reactions that involve the abstraction of hydrogen atoms, protons or hydrides. Its importance has been largely shown by the experimental observation of temperature-independent kinetic isotope effects (KIE) in different enzymes. These observations cannot be understood solely on the basis of differences in the Zero-Point Energy (ZPE) between the isotopes involved [225]. Therefore, to have a complete description of enzymatic catalysis it is necessary to bring into the picture the quantum nature of light nuclei, beyond the calculation of simple frequency-based ZPE.

Once again, it is desirable to have in place methods with the computational convenience of MM and the theoretical robustness of QM. Fortunately, a description at the TST level allows one to bypass the problem of solving the time-dependent Schrödinger equation for the nuclear system. Similar to full electronic structure calculations for bio-molecules, to solve the nuclear quantum dynamics is a formidable endeavor, possibly intractable with the current available

computational power for systems with thousands of atoms. However, nuclear quantum effects on equilibrium properties, barriers height and free energies could and should be accounted for. In this arena, several theoretical methodologies have been developed. Herein, however, we shall focus on the approaches rooted in the path integral (PI) formulation of quantum statistical mechanics [226]. This twist in our discussion is motivated by the natural extension of the QM/MM and free energy techniques to PI methods thanks to the quantum–classical isomorphism provided by Feynman's alternative formulation of QM.

The central goal of the methodologies described below is to calculate rate constants. The quantum mechanical rate constant can be written in terms of the partition functions (PF):

$$k = \kappa \frac{k_B T Q^t}{h Q^r} = \kappa \frac{k_B T}{h} \exp\left(-\beta \Delta G^t\right) \qquad (26)$$

where κ is the transmission coefficient, k_B is the Boltzmann constant, T the temperature and h is Planck's constant. Q^t and Q^r are the PF of the TS and the reactants respectively. The largest quantum effects are associated with the PF. Using PI, one can show that the canonical partition function of a quantum particle (the generalization to many particles is straightforward) under the influence of a potential V(x) is given by

$$Q = \lim_{P\to\infty}\left(\frac{mP}{2\pi\beta\hbar^2}\right)^{P/2} \int dx_1 ... dx_P \exp\left[-\sum_{i=1}^{P}\left(\frac{mP}{2\beta\hbar^2}(x_{i+1}-x_i)^2 + \tau V(x_i)\right)\right]_{x_{P+1}=x_1} \qquad (27)$$

In the limit of infinite P, this expression is exact. In practice however, finite values of P produce converged results [226]. Note the mathematical correspondence of the partition function of a single quantum particle with the partition function of a classical "ring polymer". Herein, P classical particles are connected by harmonic "springs" and each of them is subject to a fraction V(x)/P of the potential (τ=β/P). This correspondence renders the use of MM sampling

techniques based in MC or canonical MD suitable to evaluate expectation values of ensemble averages of thermodynamical properties using the PI partition function. We should clarify that the dynamics so obtained are devoid of physical meaning, and serve only as a configurational sampling of the quantum partition function (QPF).

It is worth pointing out that the correct PIMD implementation is not free of subtleties, mostly related to temperature control and integration schemes. These details, however, are outside the scope of this review and for more specific information we recommend the excellent literature on the topic found elsewhere[227, 228]

An important notion within the PI framework is the centroid variable[229]. It is defined as the centre of mass of the ring polymer and its average value along the path is given by:

$$x_c = \frac{1}{\beta\hbar}\int_0^{\beta\hbar} d\tau x(\tau) \qquad (28)$$

Hwang and Warshel [230] have developed a method that exploits the fact that the QPF can be recast in terms of the centroid variable. Using this formulation, the quantum correction to the classical free energy along the Reaction Coordinate (RC) can be written as a double average of the form:

$$G^t_{qm} - G^t_{TST} = -\frac{1}{\beta}\ln\frac{Q_P}{Q_P^{cl}} = -\frac{1}{\beta}\ln\left\langle\left\langle \exp\left(\frac{\beta}{P}\sum_i^P \Delta V_i\right)\right\rangle_{FP,x_c}\right\rangle_V \qquad (29)$$

where $\Delta V_i = V(x_i) - V(x_c)$. The outer average over V is obtained over the distribution generated by an MD simulation driven by $V(x_c)$. The inner average over FP (free particle),x_c is over the so-called free-particle distribution.

The quantized classical path (QCP) method developed by Hwang and Warshel, utilizes the trajectory obtained from classical mechanics simulations to obtain the QM correction by

performing free-particle path integral averaging with the centroid constrained to the classical position. This methodology has been successfully applied in the study of unusual KIE in several enzymatic systems [231-234].

More recently Gao and coworkers [235] introduced a bisection-sampling algorithm extensively used in quantum MC simulations of condensed matter [236, 237] into the QCP approach to sample the free-particle paths. The bisection QCP (BQCP) method has been used to study KIE in condensed phase reactions and enzymes [238] . Further development of the BQCP has been done, aimed at determining analytical expressions for the effective centroid potential [238]. These methodologies expedite the calculations and are promising for the study of large systems.

The two methodologies described above have been successfully used in conjunction with MM force fields, QM/MM and empirical valence bond (EVB) methods The main importance of these techniques is their suitability for free energy calculations using US and FEP methods described in the previous section.

For completeness, we briefly summarize other methodologies developed to study quantum effects in proton transfer reactions. These methodologies incorporate at least one of the following: QM/MM, MM forcefields or PI.

Wang and Hammes-Schiffer [239] have used a mixed quantum/classical PIMC (QC-PIMC) approach to study proton transfer in the enzyme dihydrofolate reductase. In this approach, the classical PMF along a reaction coordinate is calculated using MD trajectories propagated according to an EVB-based mapping potential and US. The nuclear quantum correction is determined separately by standard (without the centroid constraint) PIMC calculations based on an effective mapping potential. Hammes-Schiffer and coworkers have

developed and refined a hybrid quantum /classical grid method to study proton transfer reactions [240-242]. In this approach, the classical PMF for the reaction is obtained using US simulations along a mapping potential based on EVB or QM/MM techniques. Nuclear quantum effects are incorporated perturbatively into the PMF by representing the proton as a multi-dimensional vibrational wavefunction.

Semiclassical theory can be used together with MM (and in principle with QM/MM) approaches to obtain quantum corrections to rate constants. Truhlar and coworkers [243] have worked on variational transition state theories (VTST) and semiclassical quantum tunneling (QT) corrections. In VTST the position of the TS along the RC is determined variationally, as the point that minimizes the reactive flux or maximizes the free energy of activation, i.e. minimizes the rate constant [243]. Different semiclassical flavors of tunneling correction can be used in conjunction with VTST. In particular the ensemble average (EA) TST/QT has been successfully used to study KIE in a series of enzymatic reactions [236, 244, 245].

Finally, to obtain exact QM rate constants, i.e. beyond TST, it is unavoidable to determine the quantum dynamics of the system. As stated at the beginning of this section, this is a task that is currently prohibited, computationally. Be that as it may, greater efforts have been devoted to develop approximate quantum dynamical methods that rely on the computational machinery already built for MD simulations. Some of these methods, such as centroid molecular dynamics [229, 246], ring polymer molecular dynamics [247] and techniques based on analytical continuation of imaginary time correlation functions [248] are based on a PI description. Others, such as the semiclassical initial value representation are based on semiclassical ideas [249]. These methods are currently still in formal development and their applicability and limitations are being scrutinized, however they soon will reach maturity and certainly will become another

element of the toolbox for studying enzyme catalysis. A thorough discussion of these techniques is beyond the scope of this review. Needless to say, the future is bright for these methodologies and we expect that in the not so distant future applications of these methods in systems of biological relevance will begin to appear in the literature.

4.7 Summary of alternatives to QM/MM methodology

QM/MM methods are known for their high accuracy and efficiency treating large systems due to the integration of high-level and low-level methods. In fact, this philosophy of integration can also adopt pure high-level methods for the entire system or even lower-level methods for the outer region.

For the integration of all high-level methods, an example is the "divide and conquer" approach [250-253]. This approach divides the system into small fragments, calculates the electron density of each fragment with high-level QM methods and then sums up the interactions between fragments, thus conquering the whole system. While this procedure offers a promising alternative to QM/MM, it is still quite expensive to calculate each polarized fragments in a self-consistent way and the boundary treatment could cause larger errors than in QM/MM because more boundaries are involved. Moreover, sufficient sampling of the system configuration space is compromised because of high cost. So far, this approach has been mainly applied to the study of materials with highly repetitive structures. Nonetheless, this approach is of great significance in that it extraordinarily improves computational accuracy. Recently, it has been extended with MM methods to save computational cost and yielded satisfactory results [254] .

For the integration of lower-level methods than MM, an example is the quantum chemical cluster approach whose application in biological systems was recently reviewed in [255]. In this scheme, a subsystem is chosen to be treated with high-level methods and its

peripheral atoms are fixed according to the X-ray structure. Subsequently, this cluster is embedded in a polarizable continuum electrostatic field to mimic the environmental effects. In fact, a similar method, termed quantum mechanical charge field (QMCF) in [256] , has been used to study the hydration of alkali ions. In [256] , Rode and co-workers demonstrated that the QMCF treatment performed better than a QM/MM approach with mechanical embedding as long as the QM system was big enough to include the first hydration shell. Under this condition, implicit solvent can represent explicit solvent well, indicating that the configuration of the outer region described is not very important to the QM region.

But how about biochemical reactions? Sevastik and Himo [257] showed that when the QM system was enlarged (from 77 atoms to 177 atoms) for a proton transfer reaction in a protein, the calculated pK_a values conformed better to experimental results. They also made a comparison with previous studies on the same enzyme using QM/MM methods [258, 259] and found disagreements with their results. However, this finding is conceivable as there were only 30 atoms in the QM regions of [258] and [259]. Therefore, it was a question of the QM system size rather than the different methodologies. It is also found in [257] that the energies are quite different at different dielectric constants when the QM cluster is not big enough, suggesting the important role of the dielectric constant in this approach. One should also note that there are two intrinsic problems with the quantum chemical cluster method: fixed QM boundary atoms and no van der Waals interaction between the QM region and the charge field.

By and large, QM/MM methods are still more prevalent compared with other approaches, because of the accuracy of the QM portion and the efficient configurational sampling of the MM portion.

4.8 Applications to biochemical simulation

There have been numerous QM/MM applications in computational biochemical studies thanks to the rapid development of this methodology. When combined with experimental studies, the QM/MM methods have been widely used as a tool to assist the interpretation of biological spectroscopy. In electron paramagnetic resonance (EPR) spectroscopy, the hyperfine structures of the paramagnetic active sites in blue copper proteins were examined by Salahub and co-workers [18]. In nuclear magnetic resonance (NMR) spectroscopy, the QM/MM calculated chemical shifts have been compared with the experimental data to determine the binding mode of substrate to protein by Karplus and co-workers [260] and QM/MM calculations have also been used for sequence-specific NMR assignments by Lula et al [261] . In X-ray spectroscopy, QM/MM methods have been employed to refine the crystal structure of proteins [262] and to include environmental effects in the determination of the substrate's electron density in proteins [263]. On the theoretical methodology side, QM/MM has also been adopted to parameterize MM force fields to include environmental effects [264] .

Particularly, QM/MM methods find their most popular application in enzymatic reaction studies which start from experimental structures and arrive at predictions through computations. To illustrate the power of the QM/MM approach in biochemical studies, we select a single example to describe here: QM/MM simulation of DNA polymerases. This choice is based on our own interest in the related RNA polymerases [265, 266]. We believe this example nicely represents the state of the art and should give the reader a faithful representation of the excitement in the field. Other examples may be found in the references.

4.8.1 DNA polymerases

DNA polymerases (DNA pol) are crucial constituents of the complex cellular machinery for replicating and repairing DNA. They discriminate the matched deoxynucleoside triphosphate (dNTP) against the mismatched dNTPs and NTPs in an intricate mechanism. Understanding the origin of DNA pols' high fidelity on the atomic level is important to the full revelation of their exquisite cellular functions. Mammalian DNA polymerase β (pol β), a small (39 kDa) member of the X-family, has been extensively studied with computational tools by many groups. As revealed by X-ray crystallography, to ensure the high fidelity, this polymerase exhibits closed (inactive) and open (active) forms, the transition between which is triggered by the correct dNTP and hampered by the wrong dNTPs. Radhakrishnan and Schlick [267] applied the transition path sampling method at the MM level to obtain the transition states of the activation process, employed a QM/MM method for the following nucleotidyl transfer reaction and integrated these results in a kinetic Monte Carlo model to explain the kinetic difference between the correct and mismatched dNTPs. To further identify which step is the rate-limiting one, the pre-chemistry or chemistry step, Schlick and co-workers investigated the chemistry step elaborately with hybrid QM/MM and with the QM method alone.

The nucleotidyl transfer reaction in DNA pols is illustrated below in Figure 4-5. As commonly acknowledged, it involves a nucleophilic attack of the DNA primer hydroxyl oxygen (O3') on the α–phosphorus (P_α) of the incoming dNTP and P_α-O bond breaking which results in pyrophosphate group leaving. There are two possibilities for the nucleophilic attack by O3': direct attack without deprotonation of this oxygen and attack following a prior deprotonation. The first possibility has been tested by Pedersen and co-workers [268] and an extremely high energy (47 kcal/mol) was found at a short O3'-P_α distance while the H3'-O3' bond was still

present. The second possibility has been scrutinized by Bojin and Schlick [269] with a pure QM method in respect to different pathway models: direct proton transfer from O3' to O (P_α), proton transfer to adjacent Asp residues and to a water molecule. They simplified the active site by changing all aspartates to formates and the ribose ring to methyl, resulting in a model of 49 atoms and an overall charge of -3. To find the transition states, they constrained the reaction coordinates and performed constrained geometry optimization accordingly, as similar to the reaction coordinate driven method mentioned above. As a result, the direct proton hopping from O3' to O (P_α) was found to be most energetically favorable.

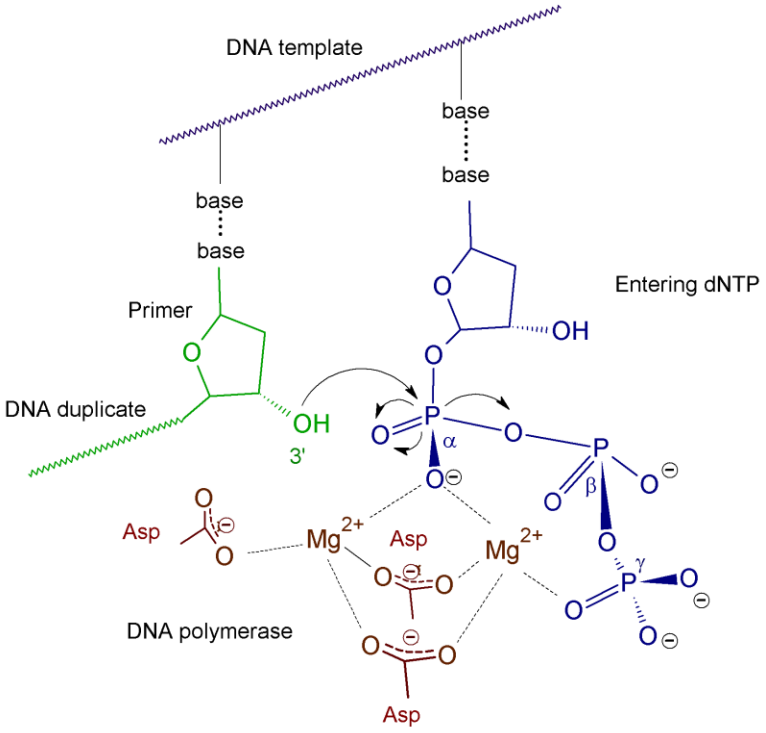


Figure 4-5: Illustration of the nucleotidyl transfer reaction

In view of the inability to include environmental effects by QM methods, Schlick and co-workers [51, 270] also recruited QM/MM methods to study this reaction and a very different

proton transfer pathway was found. They employed 10 ps QM/MM dynamics in conjunction with umbrella sampling to estimate free energies of intermediates along the reaction coordinate and discovered the proton transfer to water molecules instead of O (P_α) [270]. To validate this finding, they then compared the different proton transfer pathways with QM/MM methods [51]. A system of more than 40000 atoms including 65 QM atoms was built where all atoms within 15 Å of any QM atom were free, atoms within 15-25 Å semi-constrained and atoms further than 25 Å were fixed. To find the minimum energy path, harmonic constraints were added along the reaction coordinate instead of totally constraining it. Results of this study agreed with their previous QM/MM calculations: The O3' is deprotonated by a water molecule. It is worth noting here that hierarchical levels of QM methods were employed in this work, i.e. Hartree-Fock/STO-3G to identify preliminary geometries, B3LYP/3-31G* for geometry optimizations and MP2/6-311+G(d,p) for final energy calculations.

While supported by abundant evidence, the water-mediated proton transfer mechanism has been challenged recently by Pedersen and co-workers [59, 268] . Based on a more recent crystal structure of DNA pol β, they proposed a two-stage mechanism through their QM/MM calculations, i.e. B3LYP/6-311G**/AMBER ff99. The primer terminal O3' first replaces one of the water molecules bound to the catalytic Mg ion resulting in a prechemistry state and thereupon the reaction starts with O3' deprotonated by an aspartic residue. The prechemistry state was found to be stable after 5 ns of unconstrained MD simulation and the proton transfer barrier 6 kcal/mol. A similar study has also been conducted on pol λ, another member of the X-family polymerases by Cisneros et al [59]. The transition state (TS) search was performed using the chain-of-replica method augmented with an explicit TS finder by Yang and co-workers as

introduced above. With this approach, the proton transfer to an aspartic acid was found to be much more energetically favorable than that to a water molecule.

The question of whether there is similarity across the DNA polymerase families has stimulated great interest. DNA pol IV from the Y-family and T7 pol of phage have been investigated by Zhang and co-workers [271, 272]. Both works utilized the pseudobond approach to cap the QM subsystem, the micro/macro-iteration scheme to optimize the geometry, the reaction coordinate driven method for the reaction path search and the free energy perturbation method for free energy calculations along the path. It was found that the nucleophilic attack was the rate-limiting step and the initial proton transfer was assisted by water.

4.9 Conclusions and perspectives

To conclude this review, we look at the QM/MM methods from the viewpoint of multi-scale methodology. QM/MM is actually a combination of methods on two scales which divides a system into two levels and treats them accordingly. However, we will have to march on along the scales as the size of the systems and time length of the events grow. When investigating systems such as multiple-unit proteins, coarse-grained methods would have to be adopted to sufficiently sample their configuration space. When studying events such as protein folding in milliseconds, kinetic simulations instead of regular dynamics simulation will probably have to be used. Nevertheless, QM/MM methods are indispensible, since they can be used as benchmark calculations for coarse-grained models and to obtain rate constants for kinetic models, especially when chemical reactions are concerned.

Our groups have been applying the QM/MM-based multi-scale methodology to the study of the RNA elongation process catalyzed by RNA polymerase II. Since this event happens on a second time scale, we have adopted a kinetic Monte Carlo method to simulate this process [273].

The whole process was divided into four steps: Diffusion of the substrates, substrate moving from the entry site to the addition site, nucleotidyl transfer reaction in the addition site and the release of the pyrophosphate group. Although our tentative kinetic model based on empirical rate constants in general agreed with the experimental results, the rate constants had to be obtained through many trials, which lacks a solid physical basis. Therefore, rate constants from first-principles are still desired. The rate constant of the nucleotidyl transfer reaction is being pursued using a QM/MM method which combines CHARMM and deMon2k [53].

Our review has focused on two classes of problems. In the first, exemplified by DNA and RNA Polymerases, the QM region is the initial point of focus – getting the quantum mechanics right is essential to a correct description of the reaction. Environmental and solvent effects are then brought in to gain quantitative, and sometimes qualitative, insight. In the second class of problems, the solvent or protein-solvent dynamics are essential to the very existence of the phenomena investigated, ion solvation and transport providing the prototypical example. We hope that the "creative tension" between these two perspectives will lead to even more powerful simulation methodologies, that will be able to provide better treatments of environmental sampling for the former and the incorporation of larger and more complex quantum regions for the latter.

In summary, while QM/MM methods are important in their own right, they will also contribute to a multi-scale systems approach as a powerful component.

4.10 Bibliography

1. Warshel, A, Levitt, M. 1976. *J Mol Biol* 103: 227
2. Singh, UC, Kollman, PA. 1986. *J Comput Chem* 7: 718
3. Field, MJ, Bash, PA, Karplus, M. 1990. *J Comput Chem* 11: 700
4. Zhang, YK, Liu, HY, Yang, WT. 2000. *J Chem Phys* 112: 3483
5. Pu, JZ, Gao, JL, Truhlar, DG. 2004. *J Phys Chem A* 108: 632

6. Gao, JL, Amara, P, Alhambra, C, Field, MJ. 1998. *J Phys Chem A* 102: 4714
7. Garcia-Viloca, M, Gao, JL. 2004. *Theor Chem Acc* 111: 280
8. Bentzien, J, Muller, RP, Florian, J, Warshel, A. 1998. *J Phys Chem B* 102: 2293
9. Bakowies, D, Thiel, W. 1996. *J Phys Chem-Us* 100: 10580
10. Lyne, PD, Hodoscek, M, Karplus, M. 1999. *J Phys Chem A* 103: 3462
11. Wei, DQ, Salahub, DR. 1994. *J Chem Phys* 101: 7633
12. Maseras, F, Morokuma, K. 1995. *J Comput Chem* 16: 1170
13. Woodcock, HL, Hodoscek, M, Gilbert, ATB, Gill, PMW, Schaefer, HF, Brooks, BR. 2007. *J Comput Chem* 28: 1485
14. Freindorf, M, Shao, YH, Furlani, TR, Kong, J. 2005. *J Comput Chem* 26: 1270
15. Philipp, DM, Friesner, RA. 1999. *J Comput Chem* 20: 1468
16. Murphy, RB, Philipp, DM, Friesner, RA. 2000. *J Comput Chem* 21: 1442
17. Dapprich, S, Komaromi, I, Byun, KS, Morokuma, K, Frisch, MJ. 1999. *J Mol Struc-Theochem* 462: 1
18. Moon, S, Patchkovskii, S, Salahub, DR. 2003. *J Mol Struc-Theochem* 632: 287
19. Kastner, J, Senn, HM, Thiel, S, Otte, N, Thiel, W. 2006. *J Chem Theory Comput* 2: 452
20. Cui, Q, Elstner, M, Kaxiras, E, Frauenheim, T, Karplus, M. 2001. *J Phys Chem B* 105: 569
21. Sherwood, P, de Vries, AH, Guest, MF, Schreckenbach, G, Catlow, CRA, French, SA, Sokol, AA, Bromley, ST, Thiel, W, Turner, AJ, Billeter, S, Terstegen, F, Thiel, S, Kendrick, J, Rogers, SC, Casci, J, Watson, M, King, F, Karlsen, E, Sjovoll, M, Fahmi, A, Schafer, A, Lennartz, C. 2003. *J Mol Struc-Theochem* 632: 1
22. Cisneros, GA, Liu, HY, Lu, ZY, Yang, WT. 2005. *J Chem Phys* 122
23. Lin, H, Truhlar, DG. 2005. *J Phys Chem A* 109: 3991
24. Aqvist, J, Warshel, A. 1993. *Chem Rev* 93: 2523
25. Rosta, E, Klahn, M, Warshel, A. 2006. *J Phys Chem B* 110: 2934
26. Woo, TK, Margl, PM, Deng, L, Cavallo, L, Ziegler, T. 1999. *Catal Today* 50: 479
27. Eichinger, M, Tavan, P, Hutter, J, Parrinello, M. 1999. *J Chem Phys* 110: 10452
28. Gao, JL. 1996. *Accounts Chem Res* 29: 298
29. Gao, JL, Xia, XF. 1992. *Science* 258: 631
30. Schwenk, CF, Loeffler, HH, Rode, BM. 2001. *J Chem Phys* 115: 10808
31. Shoeib, T, Ruggiero, GD, Siu, KWM, Hopkinson, AC, Williams, IH. 2002. *J Chem Phys* 117: 2762
32. Stanton, RV, Hartsough, DS, Merz, KM. 1993. *J Phys Chem-Us* 97: 11868
33. Thompson, MA. 1996. *J Phys Chem-Us* 100: 14492
34. Thompson, MA, Glendening, ED, Feller, D. 1994. *J Phys Chem-Us* 98: 10465
35. Tongraar, A, Liedl, KR, Rode, BM. 1997. *J Phys Chem A* 101: 6299
36. Tongraar, A, Rode, BM. 2004. *Chem Phys Lett* 385: 378
37. Tunon, I, MartinsCosta, MTC, Millot, C, RuizLopez, MF. 1997. *J Chem Phys* 106: 3633
38. Bernstein, N, Kermode, JR, Csanyi, G. 2009. *Rep Prog Phys* 72
39. Bo, C, Maseras, F. 2008. *Dalton T*: 2911
40. Costabile, C, Cavallo, L. 2004. *J Am Chem Soc* 126: 9592
41. Deng, LQ, Woo, TK, Cavallo, L, Margl, PM, Ziegler, T. 1997. *J Am Chem Soc* 119: 6177
42. Froudakis, GE. 2001. *Nano Lett* 1: 179
43. Goldfuss, B, Steigelmann, M, Khan, SI, Houk, KN. 2000. *J Org Chem* 65: 77

44. Sauer, J, Sierka, M. 2000. *J Comput Chem* 21: 1470
45. Xu, ZT, Vanka, K, Ziegler, T. 2004. *Organometallics* 23: 104
46. Altarsha, M, Benighaus, T, Kumar, D, Thiel, W. 2009. *J Am Chem Soc* 131: 4755
47. Altun, A, Guallar, V, Friesner, RA, Shaik, S, Thiel, W. 2006. *J Am Chem Soc* 128: 3924
48. Altun, A, Shaik, S, Thiel, W. 2007. *J Am Chem Soc* 129: 8978
49. Banerjee, A, Yang, W, Karplus, M, Verdine, GL. 2005. *Nature* 434: 612
50. Bathelt, CM, Zurek, J, Mulholland, AJ, Harvey, JN. 2005. *J Am Chem Soc* 127: 12900
51. Alberts, IL, Wang, Y, Schlick, T. 2007. *J Am Chem Soc* 129: 11100
52. Bhattacharyya, S, Stankovich, MT, Truhlar, DG, Gao, JL. 2007. *J Phys Chem A* 111: 5729
53. Lev, B, Zhang, R, de la Lande, A, Salahub, DR and Noskov SY 2009. *Journal of Computational Chemistry* In press
54. Blumberger, J, Klein, ML. 2006. *J Am Chem Soc* 128: 13854
55. Bucher, D, Guidoni, L, Rothlisberger, U. 2007. *Biophys J* 93: 2315
56. Bukowski, MR, Koehntop, KD, Stubna, A, Bominaar, EL, Halfen, JA, Munck, E, Nam, W, Que, L. 2005. *Science* 310: 1000
57. Callis, PR, Liu, TQ. 2006. *Chem Phys* 326: 230
58. Cao, Z, Mo, Y, Thiel, W. 2007. *Angew Chem Int Edit* 46: 6811
59. Cisneros, GA, Perera, L, Garcia-Diaz, M, Bebenek, K, Kunkel, TA, Pedersen, LG. 2008. *DNA Repair* 7: 1824
60. Crespo, A, Marti, MA, Estrin, DA, Roitberg, AE. 2005. *J Am Chem Soc* 127: 6940
61. Crespo, A, Scherlis, DA, Marti, MA, Ordejon, P, Roitberg, AE, Estrin, DA. 2003. *J Phys Chem B* 107: 13728
62. Basner, JE, Schwartz, SD. 2005. *J Am Chem Soc* 127: 13822
63. Cui, Q, Elstner, M, Karplus, M. 2002. *J Phys Chem B* 106: 2721
64. Dal Peraro, M, Llarrull, LI, Rothlisberger, U, Vila, AJ, Carloni, P. 2004. *J Am Chem Soc* 126: 12661
65. Dinner, AR, Blackburn, GM, Karplus, M. 2001. *Nature* 413: 752
66. Faulder, PF, Tresadern, G, Chohan, KK, Scrutton, NS, Sutcliffe, MJ, Hillier, IH, Burton, NA. 2001. *J Am Chem Soc* 123: 8604
67. Ferrer, S, Ruiz-Pernia, JJ, Tunon, I, Moliner, V, Garcia-Viloca, M, Gonzalez-Lafont, A, Lluch, JM. 2005. *J Chem Theory Comput* 1: 750
68. Garcia-Viloca, M, Truhlar, DG, Gao, JL. 2003. *Biochemistry-Us* 42: 13558
69. Gherman, BF, Goldberg, SD, Cornish, VW, Friesner, RA. 2004. *J Am Chem Soc* 126: 7652
70. Greco, C, Bruschi, M, De Gioia, L, Ryde, U. 2007. *Inorg Chem* 46: 5911
71. Guallar, V, Wallrapp, F. 2008. *J R Soc Interface* 5: S233
72. Guimaraes, CRW, Repasky, MP, Chandrasekhar, J, Tirado-Rives, J, Jorgensen, WL. 2003. *J Am Chem Soc* 125: 6892
73. Guo, H, Cui, Q, Lipscomb, WN, Karplus, M. 2001. *P Natl Acad Sci USA* 98: 9032
74. Hermann, JC, Hensen, C, Ridder, L, Mulholland, AJ, Holtje, HD. 2005. *J Am Chem Soc* 127: 4454
75. Hermann, JC, Ridder, L, Hotje, HD, Mulholland, AJ. 2006. *Org Biomol Chem* 4: 206
76. Ishida, T. 2006. *Biochemistry-Us* 45: 5413
77. Ishida, T, Kato, S. 2003. *J Am Chem Soc* 125: 12035

78. Ishida, T, Kato, S. 2004. *J Am Chem Soc* 126: 7111
79. Jayapal, P, Sundararajan, M, Hillier, IH, Burton, NA. 2008. *Phys Chem Chem Phys* 10: 4249
80. Konig, PH, Ghosh, N, Hoffmann, M, Elstner, M, Tajkhorshid, E, Frauenheim, T, Cui, Q. 2006. *J Phys Chem A* 110: 548
81. Kubar, T, Elstner, M. 2008. *J Phys Chem B* 112: 8788
82. Lyne, PD, Mulholland, AJ, Richards, WG. 1995. *J Am Chem Soc* 117: 11345
83. Marti, S, Andres, J, Moliner, V, Silla, E, Tunon, I, Bertran, J, Field, MJ. 2001. *J Am Chem Soc* 123: 1709
84. Masson, F, Laino, T, Tavernelli, I, Rothlisberger, U, Hutter, J. 2008. *J Am Chem Soc* 130: 3443
85. Metz, S, Wang, DQ, Thiel, W. 2009. *J Am Chem Soc* 131: 4628
86. Olsson, MHM, Hong, GY, Warshel, A. 2003. *J Am Chem Soc* 125: 5025
87. Pang, JY, Hay, S, Scrutton, NS, Sutcliffe, MJ. 2008. *J Am Chem Soc* 130: 7092
88. Park, H, Brothers, EN, Merz, KM. 2005. *J Am Chem Soc* 127: 4232
89. Piana, S, Bucher, D, Carloni, P, Rothlisberger, U. 2004. *J Phys Chem B* 108: 11139
90. Rousseau, R, Kleinschmidt, V, Schmitt, UW, Marx, D. 2004. *Phys Chem Chem Phys* 6: 1848
91. Ryde, U. 2003. *Curr Opin Chem Biol* 7: 136
92. Schoneboom, JC, Lin, H, Reuter, N, Thiel, W, Cohen, S, Ogliaro, F, Shaik, S. 2002. *J Am Chem Soc* 124: 8142
93. Schoneboom, JC, Neese, F, Thiel, W. 2005. *J Am Chem Soc* 127: 5840
94. Soderhjelm, P, Ryde, U. 2006. *J Mol Struc-Theochem* 770: 199
95. Suarez, D, Merz, KM. 2001. *J Am Chem Soc* 123: 3759
96. Sundararajan, M, Hillier, IH, Burton, NA. 2006. *J Phys Chem A* 110: 785
97. Sundararajan, M, Hillier, IH, Burton, NA. 2007. *J Phys Chem B* 111: 5511
98. Szefczyk, B, Claeyssens, F, Mulholland, AJ, Sokalski, WA. 2007. *Int J Quantum Chem* 107: 2274
99. Szefczyk, B, Mulholland, AJ, Ranaghan, KE, Sokalski, WA. 2004. *J Am Chem Soc* 126: 16148
100. Topf, M, Richards, WG. 2004. *J Am Chem Soc* 126: 14631
101. Wang, Y, Hirao, H, Chen, H, Onaka, H, Nagano, S, Shaik, S. 2008. *J Am Chem Soc* 130: 7170
102. Wymore, T, Deerfield, DW, Hempel, J. 2007. *Biochemistry-Us* 46: 9495
103. Xu, D, Guo, H, Cui, G. 2007. *J Am Chem Soc* 129: 10814
104. Zhang, XD, Zhang, XH, Bruice, TC. 2005. *Biochemistry-Us* 44: 10443
105. Field, MJ. 2002. *J Comput Chem* 23: 48
106. Friesner, RA, Beachy, MD. 1998. *Curr Opin Struc Biol* 8: 257
107. Friesner, RA, Guallar, V. 2005. *Annu Rev Phys Chem* 56: 389
108. Gao, JL, Truhlar, DG. 2002. *Annu Rev Phys Chem* 53: 467
109. Hu, H, Yang, WT. 2008. *Annu Rev Phys Chem* 59: 573
110. Hu, H, Yang, WT. 2009. *J Mol Struc-Theochem* 898: 17
111. Kamerlin, SCL, Haranczyk, M, Warshel, A. 2009. *J Phys Chem B* 113: 1253
112. Lin, H, Truhlar, DG. 2007. *Theor Chem Acc* 117: 185
113. Mordasini, TZ, Thiel, W. 1998. *Chimia* 52: 288

114. Mulholland, AJ. 2005. *Drug Discov Today* 10: 1393
115. Senn, HM, Thiel, W. 2007. *Curr Opin Chem Biol* 11: 182
116. Senn, HM, Thiel, W. 2009. *Angew Chem Int Edit* 48: 1198
117. Warshel, A. 2003. *Annu Rev Bioph Biom* 32: 425
118. Woodcock, HL, Hodoscek, M, Sherwood, P, Lee, YS, Schaefer, HF, Brooks, BR. 2003. *Theor Chem Acc* 109: 140
119. Vreven, T, Byun, KS, Komaromi, I, Dapprich, S, Montgomery, JA, Morokuma, K, Frisch, MJ. 2006. *J Chem Theory Comput* 2: 815
120. Trajbl, M, Hong, GY, Warshel, A. 2002. *J Phys Chem B* 106: 13333
121. Hagiwara, Y, Ohta, T, Tateno, M. 2009. *J Phys-Condens Mat* 21
122. Lin, HZYaT, DG. *Online manual:* http://comp.chem.umn.edu/qmmm/
123. Gresh, N. 2006. *Curr Pharm Design* 12: 2121
124. Piquemal, J-Pac-w.
125. Lamoureux, G, Roux, B. 2003. *J Chem Phys* 119: 3025
126. Lu, ZY, Zhang, YK. 2008. *Journal of Chemical Theory and Computation* 4: 1237
127. Brooks, BR, Brooks, CL, Mackerell, AD, Nilsson, L, Petrella, RJ, Roux, B, Won, Y, Archontis, G, Bartels, C, Boresch, S, Caflisch, A, Caves, L, Cui, Q, Dinner, AR, Feig, M, Fischer, S, Gao, J, Hodoscek, M, Im, W, Kuczera, K, Lazaridis, T, Ma, J, Ovchinnikov, V, Paci, E, Pastor, RW, Post, CB, Pu, JZ, Schaefer, M, Tidor, B, Venable, RM, Woodcock, HL, Wu, X, Yang, W, York, DM, Karplus, M. 2009. *Journal of Computational Chemistry* 30: 1545
128. Anisimov, VM, Lamoureux, G, Vorobyov, IV, Huang, N, Roux, B, MacKerell, AD. 2005. *J Chem Theory Comput* 1: 153
129. Lamoureux, G, Harder, E, Vorobyov, IV, Roux, B, MacKerell, AD. 2006. *Chem Phys Lett* 418: 245
130. Lamoureux, G, Roux, B. 2006. *J Phys Chem B* 110: 3308
131. Lopes, PEM, Lamoureux, G, Mackerell, AD. 2009. *J Comput Chem* 30: 1821
132. Cisneros, GA, Piquemal, JP, Darden, TA. 2006. *J Phys Chem B* 110: 13682
133. Piquemal, JP, Cisneros, GA, Reinhardt, P, Gresh, N, Darden, TA. 2006. *J Chem Phys* 124
134. Cisneros, GA, Piquemal, JP, Darden, TA. 2006. *J Chem Phys* 125
135. Warshel, A, Sharma, PK, Kato, M, Xiang, Y, Liu, HB, Olsson, MHM. 2006. *Chem Rev* 106: 3210
136. Pentikainen, U, Shaw, KE, Senthilkumar, K, Woods, CJ, Mulholland, AJ. 2009. *J Chem Theory Comput* 5: 396
137. Riccardi, D, Li, GH, Cui, Q. 2004. *J Phys Chem B* 108: 6467
138. Field, MJ, Albe, M, Bret, C, Proust-De Martin, F, Thomas, A. 2000. *J Comput Chem* 21: 1088
139. Swart, M. 2003. *Int J Quantum Chem* 91: 177
140. Vasilyev, VV. 1994. *Theochem-J Mol Struc* 110: 129
141. Konig, PH, Hoffmann, M, Frauenheim, T, Cui, Q. 2005. *J Phys Chem B* 109: 9082
142. Reuter, N, Dejaegere, A, Maigret, B, Karplus, M. 2000. *J Phys Chem A* 104: 1720
143. Das, D, Eurenius, KP, Billings, EM, Sherwood, P, Chatfield, DC, Hodoscek, M, Brooks, BR. 2002. *J Chem Phys* 117: 10534
144. Amara, P, Field, MJ. 2003. *Theor Chem Acc* 109: 43

145. Thery, V, Rinaldi, D, Rivail, JL, Maigret, B, Ferenczy, GG. 1994. *J Comput Chem* 15: 269
146. Ferre, N, Assfeld, X, Rivail, JL. 2002. *J Comput Chem* 23: 610
147. Murphy, RB, Philipp, DM, Friesner, RA. 2000. *Chem Phys Lett* 321: 113
148. Amara, P, Field, MJ, Alhambra, C, Gao, JL. 2000. *Theor Chem Acc* 104: 336
149. Rodriguez, A, Oliva, C, Gonzalez, M, van der Kamp, M, Mulholland, AJ. 2007. *J Phys Chem B* 111: 12909
150. Zhang, YK, Lee, TS, Yang, WT. 1999. *J Chem Phys* 110: 46
151. Zhang, YK. 2005. *J Chem Phys* 122
152. Bessac, F, Alary, F, Carissan, Y, Heully, JL, Daudey, JP, Poteau, R. 2003. *J Mol Struc-Theochem* 632: 43
153. Carissan, Y, Bessac, F, Alary, F, Heully, JL, Poteau, R. 2006. *Int J Quantum Chem* 106: 727
154. DiLabio, GA, Hurley, MM, Christiansen, PA. 2002. *J Chem Phys* 116: 9578
155. Kerdcharoen, T, Liedl, KR, Rode, BM. 1996. *Chem Phys* 211: 313
156. Yague, JI, Mohammed, AM, Loeffler, H, Rode, BM. 2001. *J Phys Chem A* 105: 7646
157. Kerdcharoen, T, Morokuma, K. 2002. *Chem Phys Lett* 355: 257
158. Vreven, T, Morokuma, K, Farkas, O, Schlegel, HB, Frisch, MJ. 2003. *J Comput Chem* 24: 760
159. Prat-Resina, X, Gonzalez-Lafont, A, Lluch, JM. 2003. *J Mol Struc-Theochem* 632: 297
160. Marti, S, Moliner, V. 2005. *J Chem Theory Comput* 1: 1008
161. Vreven, T, Frisch, MJ, Kudin, KN, Schlegel, HB, Morokuma, K. 2006. *Mol Phys* 104: 701
162. Kastner, J, Thiel, S, Senn, HM, Sherwood, P, Thiel, W. 2007. *J Chem Theory Comput* 3: 1064
163. Klahn, M, Braun-Sand, S, Rosta, E, Warshel, A. 2005. *J Phys Chem B* 109: 15645
164. Fukui, K. 1981. *Accounts Chem Res* 14: 363
165. del Campo, JM, Koster, AM. 2008. *J Chem Phys* 129
166. A. M. Köster, PC, M. E. Casida, R. Flores-Moreno, G. Geudtner, A. Goursot, T. Heine, A. Ipatov, F. Janetzko, J. M. del Campo, S. Patchkovskii, J. U. Reveles, A. Vela, and D. R. Salahub. 2006. *The International deMon Developers Community, Cinvestav-IPN, México*
167. Berente, I, Naray-Szabo, G. 2006. *J Phys Chem A* 110: 772
168. Scharfenberg, P. 1982. *J Comput Chem* 3: 277
169. Bondar, N, Elstner, M, Fischer, S, Smith, JC, Suhai, S. 2004. *Phase Transit* 77: 47
170. Yu, HB, Noskov, SY, Roux, B. 2009. *Journal of Physical Chemistry B* 113: 8725
171. Yu, HB, Roux, B. 2009. *Biophysical Journal* 97: L15
172. Guidoni, L, Carloni, P. 2002. *Biochimica Et Biophysica Acta-Biomembranes* 1563: 1
173. Bucher, D, Guidoni, L, Maurer, P, Rothlisberger, U. 2009. *Journal of Chemical Theory and Computation* 5: 2173
174. Bucher, D, Raugei, S, Guidoni, L, Dal Peraro, M, Rothlisberger, U, Carloni, P, Klein, ML. 2006. *Biophys Chem* 124: 292
175. Nam, K, Gao, JL, York, DM. 2005. *J Chem Theory Comput* 1: 2
176. Riccardi, D, Schaefer, P, Cui, Q. 2005. *J Phys Chem B* 109: 17715
177. Schaefer, P, Riccardi, D, Cui, Q. 2005. *J Chem Phys* 123

178. Im, W, Berneche, S, Roux, B. 2001. *Journal of Chemical Physics* 114: 2924
179. Benighaus, T, Thiel, W. 2008. *J Chem Theory Comput* 4: 1600
180. Hu, H, Lu, ZY, Yang, WT. 2007. *J Chem Theory Comput* 3: 390
181. Zwanzig, RW. 1954. *J Chem Phys* 22: 1420
182. Reddy, MR, Erion, M.D. 2007. *J Am Chem Soc* 129: 9296
183. Li, GH, Zhang, XD, Cui, Q. 2003. *Journal of Physical Chemistry B* 107: 8643
184. Reddy, MR, Singh, UC, Erion, MD. 2007. *Journal of Computational Chemistry* 28: 491
185. Hu, H, Yang, WT. 2005. *J Chem Phys* 123
186. Noskov, SY, Berneche, S, Roux, B. 2004. *Nature* 431: 830
187. Noskov, SY, Roux, B. 2008. *Journal of Molecular Biology* 377: 804
188. Luzhkov, VB, Aqvist, J. 2001. *Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology* 1548: 194
189. Fowler, PW, Tai, KH, Sansom, MSP. 2008. *Biophysical Journal* 95: 5062
190. Tieleman, DP, Borisenko, V, Sansom, MSP, Woolley, GA. 2003. *Biophysical Journal* 84: 1464
191. Shrivastava, IH, Tieleman, DP, Biggin, PC, Sansom, MSP. 2002. *Biophysical Journal* 83: 633
192. Dzidic, I, Kebarle, P. 1970. *Journal of Physical Chemistry* 74: 1466
193. Tissandier, MD, Cowen, KA, Feng, WY, Gundlach, E, Cohen, MH, Earhart, AD, Tuttle, TR, Coe, JV. 1998. *Journal of Physical Chemistry A* 102: 9308
194. Hu, H, Lu, ZY, Parks, JM, Burger, SK, Yang, WT. 2008. *J Chem Phys* 128
195. Rosta, E, Haranczyk, M, Chu, ZT, Warshel, A. 2008. *J Phys Chem B* 112: 5680
196. Teleman, O, Jonsson, B. 1986. *J Comput Chem* 7: 58
197. Wei, DQ, Salahub, DR. 1994. *Chem Phys Lett* 224: 291
198. Tuckerman, M, Berne, BJ, Martyna, GJ. 1992. *J Chem Phys* 97: 1990
199. Tuckerman, ME, Parrinello, M. 1994. *J Chem Phys* 101: 1316
200. Woo, TK, Margl, P, Blochl, PE, Ziegler, T. 2002. *J Phys Chem A* 106: 1173
201. Torrie, GM, Valleau, JP. 1977. *J Comput Phys* 23: 187
202. Rajamani, R, Naidoo, KJ, Gao, JL. 2003. *Journal of Computational Chemistry* 24: 1775
203. Mezei, M. 1987. *Journal of Computational Physics* 68: 237
204. Dellago, C, Bolhuis, PG, Chandler, D. 1999. *Journal of Chemical Physics* 110: 6617
205. Roux, B. 1995. *Computer Physics Communications* 91: 275
206. Faraldo-Gomez, JD, Roux, B. 2007. *Journal of Computational Chemistry* 28: 1634
207. Earl, DJ, Deem, MW. 2005. *Phys Chem Chem Phys* 7: 3910
208. Sugita, Y, Okamoto, Y. 2000. *Chemical Physics Letters* 329: 261
209. Sugita, Y, Kitao, A, Okamoto, Y. 2000. *Journal of Chemical Physics* 113: 6042
210. Li, HZ, Yang, W. 2007. *J Chem Phys* 126
211. Brooks, BR, Bruccoleri, RE, Olafson, BD, States, DJ, Swaminathan, S, Karplus, M. 1983. *J Comput Chem* 4: 187
212. Seabra, GD, Walker, RC, Roitberg, AE. 2009. *Journal of Physical Chemistry A* 113: 11938
213. Yang, W, Nymeyer, H, Zhou, HX, Berg, B, Bruschweiler, R. 2008. *Journal of Computational Chemistry* 29: 668
214. Woods, CJ, Manby, FR, Mulholland, AJ. 2008. *Journal of Chemical Physics* 128

215. Bockmann, M, Peter, C, Delle Site, L, Doltsinis, NL, Kremer, K, Marx, D. 2007. *Journal of Chemical Theory and Computation* 3: 1789
216. Cummins, PL, Rostov, IV, Gready, JE. 2007. *Journal of Chemical Theory and Computation* 3: 1203
217. Moskovsky, AA, Vanovschi, VV, Konyukhov, SS, Nemukhin, AV. 2006. *International Journal of Quantum Chemistry* 106: 2208
218. Liu, HY, Lu, ZY, Cisneros, GA, Yang, WT. 2004. *J Chem Phys* 121: 697
219. Chu, JW, Trout, BL, Brooks, BR. 2003. *J Chem Phys* 119: 12708
220. Woodcock, HL, Hodoscek, M, Brooks, BR. 2007. *J Phys Chem A* 111: 5720
221. Xie, L, Liu, HY, Yang, WT. 2004. *J Chem Phys* 120: 8039
222. Bolhuis, PG, Chandler, D, Dellago, C, Geissler, PL. 2002. *Annu Rev Phys Chem* 53: 291
223. Quaytman, SL, Schwartz, SD. 2007. *P Natl Acad Sci USA* 104: 12253
224. Quaytman, SL, Schwartz, SD. 2009. *J Phys Chem A* 113: 1892
225. Nagel, ZD, Klinman, JP. 2009. *Nature Chemical Biology* 5: 696
226. Feynman, RP, Hibbs, A.R. 1965. *Quantum Mechanics and Path Integrals*. New York: McGraw Hill
227. Paesani, F, Voth, GA. 2009. *Journal of Physical Chemistry B* 113: 5702
228. Tuckerman, M. 2002. *NIC Series and References therein*
229. Cao, JS, Voth, GA. 1994. *Journal of Chemical Physics* 100: 5093
230. Hwang, JK, Warshel, A. 1993. *Journal of Physical Chemistry* 97: 10053
231. Olsson, MHM, Parson, WW, Warshel, A. 2006. *Chemical Reviews* 106: 1737
232. Olsson, MHM, Siegbahn, PEM, Warshel, A. 2004. *Journal of Biological Inorganic Chemistry* 9: 96
233. Olsson, MHM, Siegbahn, PEM, Warshel, A. 2004. *Journal of the American Chemical Society* 126: 2820
234. Wang, Q, Hammes-Schiffer, S. 2006. *Journal of Chemical Physics* 125
235. Major, DT, Gao, JL. 2005. *Journal of Molecular Graphics & Modelling* 24: 121
236. Gao, JL, Ma, SH, Major, DT, Nam, K, Pu, JZ, Truhlar, DG. 2006. *Chemical Reviews* 106: 3188
237. Major, DT, Garcia-Viloca, M, Gao, JL. 2006. *Journal of Chemical Theory and Computation* 2: 236
238. Wong, KY, Gao, J. 2008. *Journal of Chemical Theory and Computation* 4: 1409
239. Wang, ML, Lu, ZY, Yang, WT. 2006. *Journal of Chemical Physics* 124
240. Chakravorty, DK, Soudackov, AV, Hammes-Schiffer, S. 2009. *Biochemistry* 48: 10608
241. Carra, C, Iordanova, N, Hammes-Schiffer, S. 2008. *Journal of the American Chemical Society* 130: 8108
242. Hatcher, E, Soudackov, AV, Hammes-Schiffer, S. 2004. *Journal of the American Chemical Society* 126: 5763
243. Truhlar, DG, Garrett, BC. 1984. *Annual Review of Physical Chemistry* 35: 159
244. Pang, JY, Pu, JZ, Gao, JL, Truhlar, DG, Allemann, RK. 2006. *Journal of the American Chemical Society* 128: 8015
245. Bhattacharyya, S, Stankovich, SMMT, Truhlar, DG, Gao, JL. 2005. *Biochemistry* 44: 16549
246. Cao, JS, Voth, GA. 1996. *Journal of Chemical Physics* 105: 6856
247. Craig, IR, Manolopoulos, DE. 2004. *Journal of Chemical Physics* 121: 3368

248. Rabani, E, Krilov, G, Berne, BJ. 2000. *Journal of Chemical Physics* 112: 2605
249. Miller, WH. 2001. *Journal of Physical Chemistry A* 105: 2942
250. Yang, WT. 1991. *Phys Rev Lett* 66: 1438
251. Li, W, Li, SH. 2004. *J Chem Phys* 121: 6649
252. Deev, V, Collins, MA. 2005. *J Chem Phys* 122
253. Ganesh, V, Dongare, RK, Balanarayan, P, Gadre, SR. 2006. *J Chem Phys* 125
254. Li, H, Li, W, Li, SH, Ma, J. 2008. *J Phys Chem B* 112: 7061
255. Siegbahn, PEM, Himo, F. 2009. *J Biol Inorg Chem* 14: 643
256. Azam, SS, Hofer, TS, Randolf, BR, Rode, BM. 2009. *J Phys Chem A* 113: 1827
257. Sevastik, R, Himo, F. 2007. *Bioorg Chem* 35: 444
258. Cisneros, GA, Liu, HY, Zhang, YK, Yang, WT. 2003. *J Am Chem Soc* 125: 10384
259. Tuttle, T, Thiel, W. 2007. *J Phys Chem B* 111: 7665
260. Spichty, M, Taly, A, Hagn, F, Kessler, H, Barluenga, S, Winssinger, N, Karplus, M. 2009. *Biophys Chem* 143: 111
261. Lula, I, Denadai, AL, Resende, JM, de Sousa, FB, de Lima, GF, Pilo-Veloso, D, Heine, T, Duarte, HA, Santos, RAS, Sinisterra, RD. 2007. *Peptides* 28: 2199
262. Ryde, U, Nilsson, K. 2003. *J Mol Struc-Theochem* 632: 259
263. Mladenovic, M, Arnone, M, Fink, RF, Engels, B. 2009. *J Phys Chem B* 113: 5072
264. Otte, N, Bocola, M, Thiel, W. 2009. *J Comput Chem* 30: 154
265. Kornberg, RD. 2007. *P Natl Acad Sci USA* 104: 12955
266. Cramer, P, Armache, KJ, Baumli, S, Benkert, S, Brueckner, E, Buchen, C, Damsma, GE, Dengl, S, Geiger, SR, Jaslak, AJ, Jawhari, A, Jennebach, S, Kamenski, T, Kettenberger, H, Kuhn, CD, Lehmann, E, Leike, K, Sydow, JE, Vannini, A. 2008. *Annu Rev Biophys* 37: 337
267. Radhakrishnan, R, Schlick, T. 2005. *J Am Chem Soc* 127: 13245
268. Lin, P, Batra, VK, Pedersen, LC, Beard, WA, Wilson, SH, Pedersen, LG. 2008. *P Natl Acad Sci USA* 105: 5670
269. Bojin, MD, Schlick, T. 2007. *J Phys Chem B* 111: 11244
270. Radhakrishnan, R, Schlick, T. 2006. *Biochem Bioph Res Co* 350: 521
271. Wang, LH, Yu, XY, Hu, P, Broyde, S, Zhang, YK. 2007. *J Am Chem Soc* 129: 4731
272. Wang, LH, Broyde, S, Zhang, YK. 2009. *J Mol Biol* 389: 787
273. Zhu R, dlLA, Zhang R, Salahub DR. 2009. *Interdisplinary Science: Computational Life Science* 1: 91

CHAPTER FIVE: THE QM-MM INTERFACE FOR CHARMM-DEMON

5.1 Abstract

We present a new QM/MM interface for fast and efficient simulations of organic and biological molecules. The CHARMM/deMon interface has been developed and tested to perform minimization and atomistic simulations for multi-particle systems. The current features of this QM/MM interface include readability for molecular dynamics, tested compatibility with Free Energy Perturbation simulations (FEP) using the dual topology/single coordinate method. The current coupling scheme uses link atoms, but further extensions of the code to incorporate other available schemes are planned. We report the performance of different levels of theory for the treatment of the QM region, while the MM region was represented by a classical force-field (CHARMM27) or a polarizable force-field based on a simple Drude model. The current QM/MM implementation can be coupled to the dual-thermostat method and the VV2 integrator to run molecular dynamics simulations.

5.2 Introduction

Treating large and complex systems at the atomic level remains a challenge in molecular modeling and simulation of condensed phase and biomolecular phenomena even though computational possibilities are greatly increased nowadays. Still, most modern electronic structure calculations are performed in the gas-phase. However, many processes of great interest for modern chemistry and biochemistry happen in the bulk solution or in the core of a protein. So the quest for rapid yet reasonably cheap methods to include electronic structure in biomolecular simulations remains one of the highest priorities of modern computational chemistry. Traditional molecular mechanical (MM) methods enable simulations of atomic systems composed of hundreds of thousands of atoms. However, the force- fields are inadequate in situations where the

electronic structure of a system plays a role. The correct treatment of electronic structure changes is an absolute requirement for any process involving electron transfer, protonation, charge transfer or the formation/breaking of covalent bonds. A natural solution to this problem is a combination of the QM and MM treatments in one scheme. The combination of QM and MM, often termed QM/MM, allows the investigation of large systems in complex environments at a reasonable cost, while remaining accurate. The origins of the QM/MM approach can be traced back to the mid 70s, when Warshel and Levitt [1] and later Singh and Kollman [2] developed methods to combine classical and quantum simulations together. Later, Field, Bash and Karplus [3] developed an interface between a semiempirical program and the CHARMM simulation package, opening a broad avenue for similar efforts directed at a "two programs under the same roof" ideology. The advantage of schemes linking two separate program packages is evident, the large body of users for popular biomolecular simulation packages such as CHARMM [4], Amber [5, 6], Gromacs [7] etc. will be able to set up QM/MM simulations with the same syntax and scripting logic. At the same time, users of QM packages can access all features developed for the treatment of large systems with already implemented thermostats, barostats, minimization algorithms etc.

A large number of interfaces have been implemented already, to enable different QM/MM protocols, such as the energy expression, QM/MM coupling protocols and QM/MM boundary treatments.[6, 8, 9] For example, within the CHARMM project, users can access Q-Chem,[10] Gamess [11] and some other quantum-chemical packages. So, the logical question is why we need yet another QM/MM interface for biomolecular simulation. The answer lies in the large heterogeneity of methods development within different quantum chemistry software projects. For example, some of the DFT functionals, to study many cofactors (e.g., metals), are

available only in ADF or deMon2k packages. For instance, the linkage between CHARMM and deMon provides the advantage of the abundant choices of functionals in deMon, such as GGA functionals—PBE [12] and meta-GGA functionals—TPSS-TPS [13], and TPSS-3 [14], of the ability to mimic systems involving considerable hydrogen bonding and van der Waals interactions. The high efficiency of deMon facilitated by its auxiliary basis sets [15] and powerful parallel scheme16 also improves considerably the speed of QM/MM calculations. In our current work, we report the extension of CHARMM's QM/MM capabilities to utilize deMon2k's [17] efficient density functional theory (DFT) method. Earlier QM/MM works utilizing demon capabilities can be seen in refs [18, 19]. In the interface between CHARMM and deMon2k, an additive scheme, electronic embedding and the link atom method are implemented. In the following sections, we briefly review the theoretical basis of QM/MM method- ology and free energy perturbation theory within the QM/MM framework. A number of test cases are also provided to validate our implementation.

5.3 Computational Methodology

5.3.1 QM/MM Decomposition

In a typical QM/MM approach, the entire system is primarily divided into two subsystems, one that requires a quantum mechanical treatment, e.g., a chemical reaction, and the other which may be treated on the MM level. Intuitively, the total energy of the whole system is the sum of the two subsystems, which is referred to as the additive scheme. However, a subtractive scheme, where the entire system is treated by MM, the QM region treated both by QM and MM, then the part calculated at the MM level is subtracted, is also sometimes applicable, see details in refs [20–24]. The additive scheme has been implemented by [3, 10, 25–27] and the subtractive one by [20–24].

The coupling interactions between the QM and MM subsystems include van der Waals interactions and electrostatic interactions. The latter are generally described in two ways, mechanical embedding and electrical embedding. Mechanical embedding accounts for the interaction between the two subsystems on the classical mechanical level. Electrical embedding involves inserting a one-electron operator for the electrostatic interactions into the Hamiltonian of the QM system. Mechanical embedding has been employed in refs. [10, 25–27] and the more popular electrical embedding is present in refs. [4, 20–24,] and [28–30]. The result of a QM/MM simulation is highly sensitive to the treatment of the boundary between the QM and MM subsystems when chemical bonds have to be crossed to partition the system. To circumvent the side effects of the dangling bonds, two approaches have been proposed. One is the general link atom scheme using a hydrogen atom as the frontier atom, [3, 10, 20–22, 26] a pseudobond [31] or a quantum capping potential [32] to cap the unsaturated bond. The other is the so-called local self-consistent field (LSCF) algorithm [33–35] employing strictly localized bond orbitals (SLBOs). We have chosen to implement the link atom scheme in our interface at this time, where QM/MM electrostatics for link host groups can be removed completely.

In addition to standard QM/MM functionality, the interface between CHARMM and deMon2k is also capable of tight binding (TB) self-consistent field calculations and free energy perturbation (FEP) calculations (see below). The particular QM/MM scheme employed in the interface between CHARMM and deMon2k is based on the work of Field et al.[3] and Woodcock et al. [10] adopting an additive scheme. Following the same notations proposed by Field et al., the effective Hamiltonian of the entire system is formalized as

$$\hat{H}_{\text{eff}} = \hat{H}_{\text{QM}} + \hat{H}_{\text{MM}} + \hat{H}_{\text{QM/MM}} \qquad (1)$$

where $\hat{H}_{QM}$ is the pure Hamiltonian of the QM subsystem including the link atom(s), $\hat{H}_{MM}$ is the pure classical Hamiltonian described by the force field, and $\hat{H}_{QM/MM}$ is the Hamiltonian accounting for the coupling between the two subsystems.

According to the electrical embedding formula, $\hat{H}_{QM/MM}$ is given as

$$\hat{H}_{QM/MM} = -\sum_{i,M} \frac{q_M}{r_{iM}} + \sum_{A,M} \frac{Z_A q_M}{R_{AM}} + \sum_{A,M} \left(\frac{A_{AM}}{R_{AM}^{12}} - \frac{B_{AM}}{R_{AM}^{6}} \right) \quad (2)$$

where the first term is a single-electron operator generated by the external MM point charges, the second term describes the Coulomb interaction between the QM nuclei and external MM charges and the last accounts for the Pauli repulsion and van der Waals attraction between QM and MM atoms in the Lennard-Jones formalism.

Polarization effects are included in the given expressions for the hybrid QM/MM scheme. They can be accounted for by employing the Drude model [36], which has been implemented in CHARMM. Those artificial atoms can be passed to deMon as additional embedded charges. Therefore $\hat{H}_{QM/MM}$ would be expanded for those interactions. Simple tests for that approach have been performed within the Q-Chem/CHARMM interface [10]; we have used similar test cases for the QM/MM decomposition implemented.

5.3.2 Molecular Dynamics Simulations: Polarizable and Nonpolarizable Force-Fields

The simulations with PARAM274 and Drude polarizable [36] force fields were performed with CHARMM c35b1 modified to include the interface between CHARMM and deMon2k. For the evaluation of dipole moments around an ion a box of 216 water molecules was simulated at 298 K and 1 atm. The free energy perturbation for water clusters was performed

with a confining potential. A water droplet was maintained by a steep half-harmonic potential of 100 kcal/mol/Å acting only if water oxygens were displaced by more than 3.5 Å. Free Energy Perturbation simulations has been per-formed incorporating the dual topology single coordinate method [37]. The hamiltonian describing the system which undergoes change from state A to state B during FEP calculations can be written as

$$H(\lambda, t) = \lambda HA(t) + (1 - \lambda)HB(t), \quad (3)$$

where HA and HB share coordinates for all objects included. The incorporation of the QM/MM potentials into dual-topology method has been described in ref. 38.

The free energy difference between Na^+ and K^+ were evaluated using the dual topology method. All runs were short and presented not to claim precise values, but rather to show that this type of computations works within the CHARMM/deMon interface.

5.3.3 Technical Details of the Implementation

The CHARMM/deMon interface has been implemented based on the approach used to interface Q-Chem and CHARMM [10], which means that the QM/MM interaction has been included at the source code level in CHARMM as well as in deMon and no further code modifications are needed in order to perform QM/MM calculations. Each program can be run separately and deal with the tasks they usually deal with. However, CHARMM is the host program for QM/MM calculations. It sets up the QM job which has to be performed and after it is done it patches all energies and gradients needed to perform further calculations at the MM level. A simple scheme is shown in Figure 5-1.

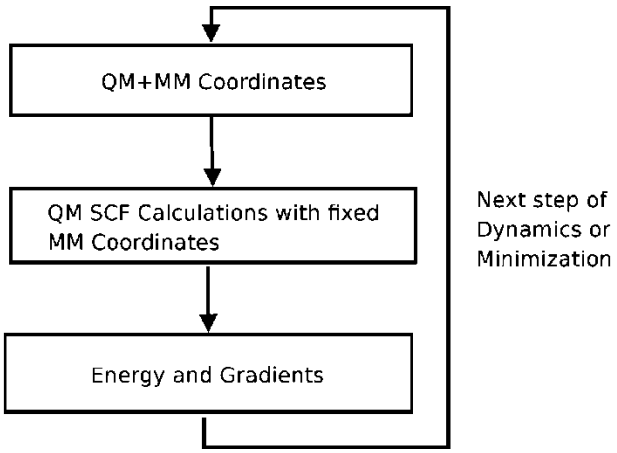


Figure 5-1: Simplified QM/MM Scheme of CHARMM/deMon

5.4 Results and Discussion

In the process of testing the CHARMM/deMon interface, a set of simple tests has been performed, such as the water dimer minimization, free energy perturbation calculations, and tests on the link atoms. All of them incorporated different levels of QM theory and some of them used polarizable models for the classical part. Three major issues have been tested: the accuracy of computation, the ability to perform complex jobs and the ability to deal with the boundary between QM and MM regions i.e., the link atoms.

5.4.1 Link Atoms

As a test for the link atoms, we have taken the same alanine test case as for the CHARMM/Q-Chem interface. The system has been split in such a way that the C-terminus is in the QM region and the rest of the molecule is in MM; the link atom is between Cα and C. Minimization has been performed using the ABNR (adapted basis Newton-Raphson) method [39] until the RMS gradients were less then 0.001 kcal/mol/Å. Then the results obtained for CHARMM/deMon were compared to those from CHARMM/Q-Chem calculations performed within the same conditions i.e., the QM region has been treated by using the B88-LYP functional

and the STO-3G basis set. The RMS deviation between those two obtained structures is 0.0085 Å, which can be explained by differences of math engines implemented in deMon and Q-Chem.

5.4.2 Water Dimer

The water dimer has been chosen as one of the tests for the CHARMM/deMon interface to confirm accuracy. One of the water molecules has been treated as the QM part of the system, the other one has been treated as the MM part, employing the TIP3P or Drude polarizable model in CHARMM. Therefore, two major configurations are possible, when the QM water is the donor and the MM water is the acceptor of the hydrogen bond and vice versa. Different levels have been used to treat the quantum part, B88-LYP and PBE-PBE functionals along with STO-3G, DZVP, PVTZ-FIP1, and AUG-CC-PVQZ basis sets. The tolerance for QM SCF calculations was 10−8 and the maximum number of steps set to be 100. Hybrid QM/MM minimization has been performed by CHARMM using the internal ABNR method. All minimizations were stopped after the RMS gradients become less than or equal to 0.01 kcal/mol/Å.

All QM/MM water dimer calculations, irrespective of the MM model used, showed that results agree better with the experimental data, see Tables 5-1 and 5-2 if the QM water is the acceptor. However, within a polarizable model the results seem to estimate the energies a little bit better. For the QM part treated with PBE-PBE/STO-3G and MM treated as Drude, the QM dipole moment, for QM being acceptor and donor, has been calculated to be 1.638 debye and 1.632 debye, respectively. Calculations of the QM water dipole shown no significant difference between acceptor and donor cases, therefore this is excluded as a possible cause for the difference in energies as well as improbable cause for the overestimation of the binding.

Table 5-1: Ab Initio QM/MM Results for the Water Dimer

For All Calculations the Classical Portion Employed the TIP3P Model.[a]

Method	Basis	Type	ΔE	$d(H\cdots O)$	$\angle(O\cdots OH)$	$d(O\cdots O)$
B88-LYP	AUG-CC-PVQZ	Donor	−8.34	1.69	179.0	2.68
PBE-PBE	AUG-CC-PVQZ	Donor	−8.33	1.69	179.0	2.68
B88-LYP	DZVP	Donor	−8.14	1.70	179.3	2.70
PBE-PBE	DZVP	Donor	−8.24	1.70	179.0	2.69
B88-LYP	PVTZ-FIP1	Donor	−8.34	1.69	179.0	2.69
PBE-PBE	PVTZ-FIP1	Donor	−8.34	1.69	179.0	2.69
B88-LYP	STO-3G	Donor	−4.81	1.77	172.3	2.81
PBE-PBE	STO-3G	Donor	−4.82	1.77	172.4	2.81
B88-LYP	AUG-CC-PVQZ	Acceptor	−6.06	1.84	178.3	2.81
PBE-PBE	AUG-CC-PVQZ	Acceptor	−6.06	1.83	178.7	2.80
B88-LYP	DZVP	Acceptor	−6.1	1.91	174.3	2.87
PBE-PBE	DZVP	Acceptor	−6.78	1.83	177.5	2.79
B88-LYP	PVTZ-FIP1	Acceptor	−6.15	1.85	177.3	2.82
PBE-PBE	PVTZ-FIP1	Acceptor	−6.05	1.85	177.7	2.82
B88-LYP	STO-3G	Acceptor	−3.9	1.91	174.3	2.87
PBE-PBE	STO-3G	Acceptor	−4.18	1.86	178.4	2.83
TIP3P/TIP3P			−6.14	1.83	178.7	2.81
Drude/Drude			−5.93	1.87	171.3	2.82
Extrapolated[40]			5.02 ± 0.05			2.91
Focal point[41]			5.02 ± 0.07			2.91
Experimental[42, 43]			5.44 ± 0.7		174 ± 20	2.98

a Binding energies ΔE, are reported in kcal/mol. Geometric parameters are reported in Ångstroms (bond distances) and degrees (bond angles).

Table 5-2: Ab Initio QM/MM Results for the Water Dimer

For all Calculations the Classical Portion Employed the Polarizable Drude Water Model.[a]

Method	Basis	Type	ΔE	$d(H\cdots O)$	$\angle(O\cdots OH)$	$d(O\cdots O)$
B88-LYP	AUG-CC-PVQZ	Donor	−8.39	1.71	173.0	2.69
PBE-PBE	AUG-CC-PVQZ	Donor	−8.47	1.71	173.1	2.69
B88-LYP	DZVP	Donor	−7.90	1.73	174.7	2.72
PBE-PBE	DZVP	Donor	−8.03	1.73	174.7	2.72
B88-LYP	PVTZ-FIP1	Donor	−8.30	1.71	173.0	2.70
PBE-PBE	PVTZ-FIP1	Donor	−8.28	1.71	173.1	2.70
B88-LYP	STO-3G	Donor	−3.90	1.85	176.5	2.89
PBE-PBE	STO-3G	Donor	−4.01	1.85	176.2	2.89

B88-LYP	AUG-CC-PVQZ	Acceptor	−5.81	1.88	173.1	2.83
PBE-PBE	AUG-CC-PVQZ	Acceptor	−5.91	1.88	172.9	2.83
B88-LYP	DZVP	Acceptor	−6.51	1.86	173.8	2.81
PBE-PBE	DZVP	Acceptor	−6.51	1.86	173.6	2.81
B88-LYP	PVTZ-FIP1	Acceptor	−5.80	1.90	173.4	2.85
PBE-PBE	PVTZ-FIP1	Acceptor	−5.80	1.89	173.3	2.84
B88-LYP	STO-3G	Acceptor	−3.87	1.91	174.2	2.87
PBE-PBE	STO-3G	Acceptor	−3.88	1.91	174.3	2.86
TIP3P/TIP3P			−6.14	1.83	178.7	2.81
Drude/Drude			−5.93	1.87	171.3	2.82
Extrapolated[40]			5.02 ± 0.05			2.91
Focal point[41]			5.02 ± 0.07			2.91
Experimental[42, 43]			5.44 ± 0.7		174 ± 20	2.98

a Binding energies ΔE, are reported in kcal/mol. Geometric parameters are reported in Ångstroms (bond distances) and degrees (bond angles).

5.4.3 Solvation of Na+ and K+ in Water

Understanding of ion solvation at the molecular level has been one of the cornerstones of physical chemistry of solutions for over a hundred years. The ability to decompose ion–solvent and solvent–solvent interactions with high precision is vital to the understanding of processes as diverse as cell signaling via ion channels, functions of nanopores and enzymatic catalysis. Recent developments in the analytical theory of ion solvation come hand in hand with progress in molecular simulations. One of the main issues for the accurate description of ion solvation has to do with accurate accounting for polarization effects missing in most of the conventional force-fields [44]. The fixed dipole moment assigned to the solvent helps to obtain an accurate description of thermodynamics, but polarization effects are expected to play a dominant role in the situation where the environment cannot be described as a structureless continuum and induced dipoles become substantial. In this case, QM/MM simulations provide an indispensable tool to evaluate polarization effects [45]. To illustrate the potential importance of the polarizable

model we have obtained the average induced dipole moment dependence of a water molecule on oxygen-ion distance as shown in Figure 5-2.

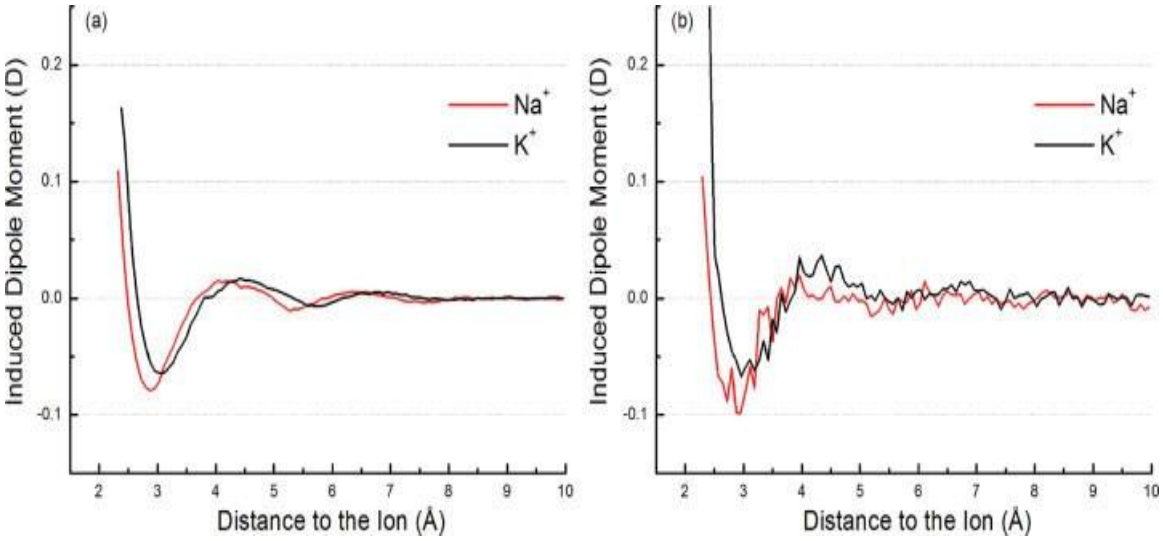


Figure 5-2: Induced dipole moment of water dependence on distance from oxygen to the ion

(a) Drude ion/Drude water, (b) QM ion/ Drude water (short statistics).

The induced dipole moment of water molecules as a function of distance is shown for (a) Drude ion—polarizable (SWM4) water and (b) quantum ion—Drude (SWM4) water systems. Importantly, both simulations for two different ions provided qualitatively the same estimate for the induced dipole. To provide validation for the developed interface between CHARMM and deMon we focused on two very well studied cationic systems - Na+ and K+ in water. Both ions play an important role in biology and chemistry and thus qualitative models for solvation will provide a necessary technical tool to further our understanding of sodium and potassium chemistry [46]. This finding indicates that SWM4 is able to capture polarization effects in water accurately and thus provide a comparable description though less detailed than the full QM system. As expected, the dipole moment slowly converges, with distance, to that for bulk

water simulations (~2.5 D), but the average dipole moment of water molecules in the direct vicinity of the ion displays a very nonlinear behavior, Figure 5-2b; the overall behavior of the dipole moment curve has been reproduced, despite limited sampling. It was found that the maximal induced dipole of water molecules in direct contact with K^+ is about 0.16 D (Drude model) and 0.21 D (QM system), whereas the induced dipole for water molecules in contact with Na^+ is about 0.11 D for both polarizable MM and QM systems. The highest value in the induced dipole moment of water's oxygen bound to an ion corresponds to the ion–oxygen direct pairing at around 2.3–2.5 Å for Na^+ and 2.6–2.7 Å for K^+. The induced dipole moment is rapidly damped as the ion–oxygen distance increases. These numbers are consistent with those estimated from Car-Parrinello dynamics [49] and recent DFT studies on large water cluster dynamics [50]. The total and induced dipole moments of water molecules follow closely the shape of the potential of mean force (PMF), depicted on Figure 5-3. The dipole moment of water molecules in transit between the first and second solvation shells or in the vicinity of the potential barrier in the ion-oxygen PMF is smaller than that of the bulk solution (2.5 D). The results obtained with both polarizable and full QM simulations are in excellent accordance with previous studies on the polarization effects in salt solutions [51].

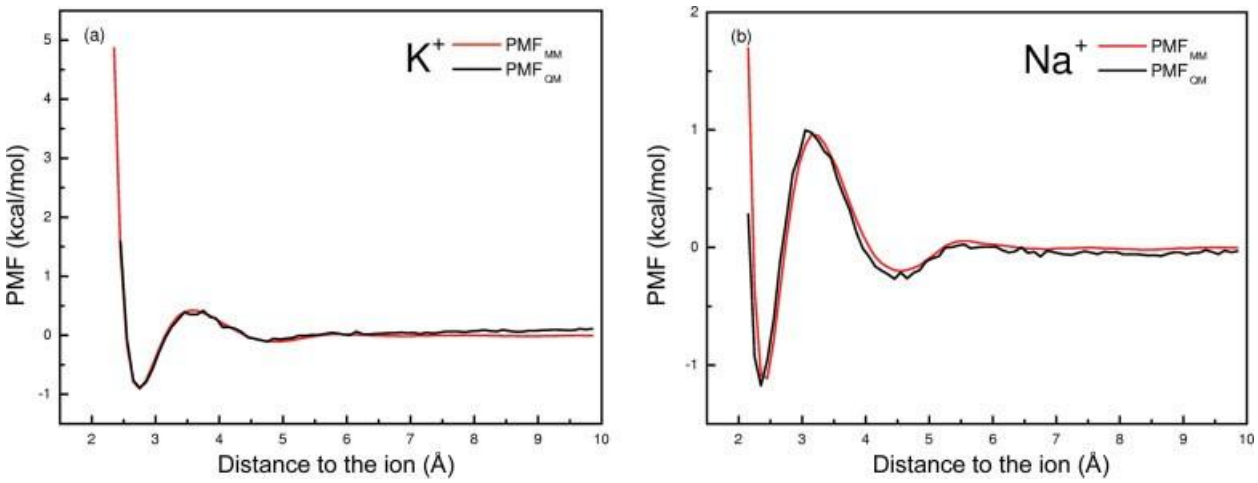


Figure 5-3: Potential of mean force for water around the alkali ions for the SWM4 model The MM functions are red, the QM functions are black lines. Distances are in Å, graph (a) for K^+, (b) for Na^+.

The simplest approach to describe the structure of the solvation shell is to evaluate the ion–oxygen and ion–hydrogen radial 2 distribution functions and respective running integration numbers. It is still very challenging to run full QM simulations and to evaluate solvent structure from first principles. To validate the current release of the interface we opted for a reduced case. We chose the solvent to be treated with the developed polarizable model (SWM4) and the ion to be either quantum or classical, but with polarization included via the Drude method. The resulting radial distribution functions of the alkali ions and oxygen or hydrogen sites (treated as QM and Drude) for the SWM4 model are shown in Figure 5-4. The positions of the maxima and minima as well as the peak values do agree for both QM/MM and pure MM (Drude) calculations. They also display an excellent agreement, see Table 5-3 for analysis, with a large number of papers published to date on the structure of K^+ and Na^+ aqueous solutions [52-55].

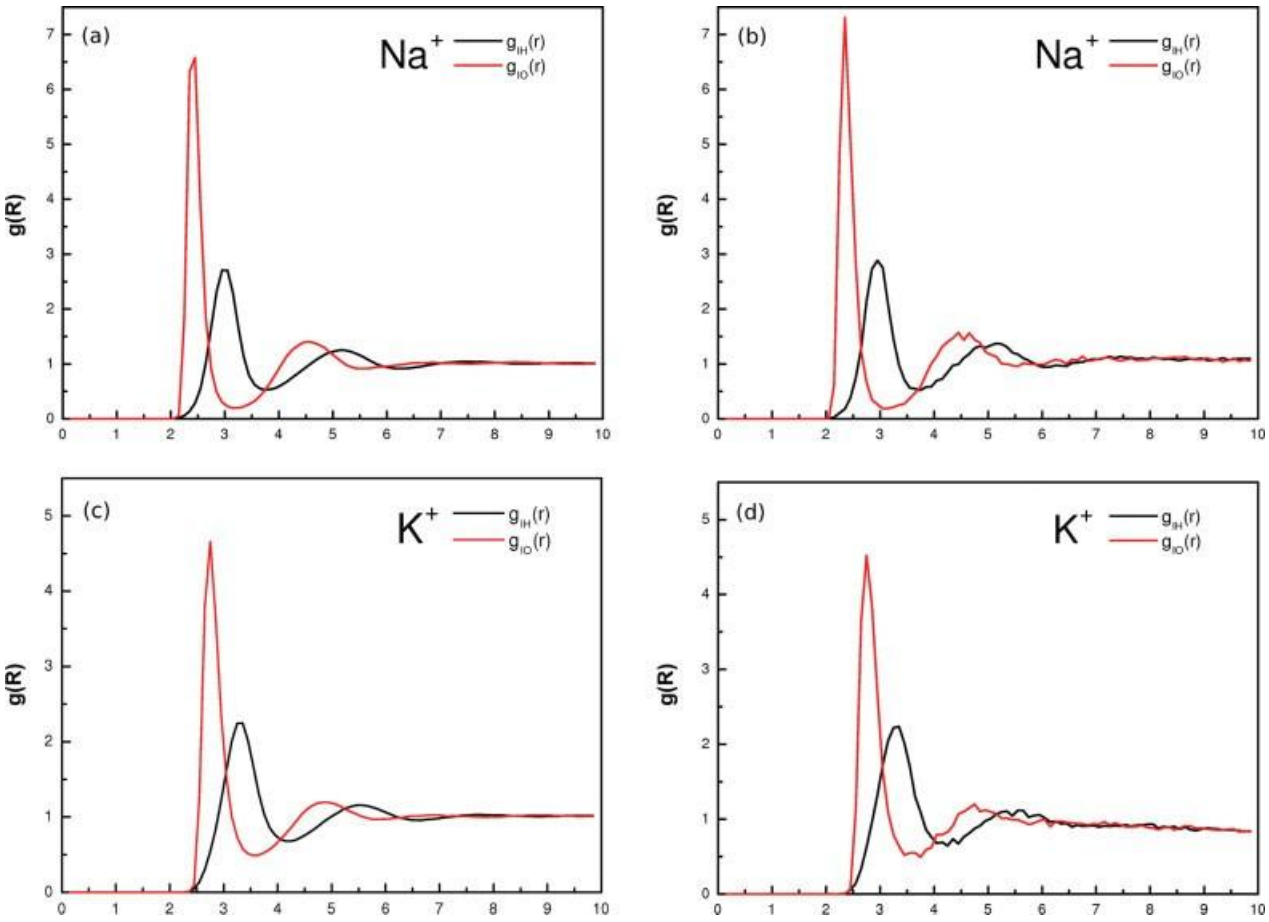


Figure 5-4: Radial hydration structure of the alkali ions for the SWM4 model

The $g_{IO}(r)$ functions are red, the $g_{IH}(r)$ functions are black lines. Distances are in Å. Cases (a) and (c): ion treated as a Drude oscillator, (b) and (d) as a QM ion.

Table 5-3: Radial distribution data analysis

Type	R_{max}	h_{max}	R_{min}	n
Na^+ Drude	2.45	6.58	3.15	5.77
Na^+ QM	2.35	7.30	3.05	5.69
Na^+ Exp[47]	2.43			5.68
K^+ Drude	2.75	4.65	3.55	7.13
K^+ QM	2.75	4.52	3.45	7.05
K^+ Exp[48]	2.65		3.45	6.7–6.9

Positions of maxima and minima (R_{max}, R_{min}) in Å along with value at maxima h_{max} and coordination number n are presented for both Na^+ and K^+

The initial minimization employed the ABNR (adapted basis Newton-Raphson) method. The VV2 algorithm, the velocity-Verlet algorithm created to simulate efficiently the motion of Drude oscillators, has been used to treat the system dynamics. Two separate thermostats have been used; the one for the Drude oscillators has been set to 0.1 K to keep them from breaking away from their parental particles and the other one to 315K to control the overall temperature of the system. The run length for the Drude ion was 5 ns, but only 0.1 ns for the quantum ion. Therefore, that data show important differences between polarizable and non-polarizable models, as well as it can be considered as another successful test for the CHARMM/deMon interface.

5.4.4 Free Energy Perturbation: Thermodynamics of Ion–Water Clusters

Several papers published recently raise serious concerns about the quality of nonpolarizable force-fields if applied to study the thermodynamics of ion solvation [56]. It was proposed that lack of explicit polarization may invalidate studies of ion selectivity. To clarify this issue and to test the developed free energy perturbation protocol further, we have performed free energy perturbation for Na^+ and K^+ in water droplets. The advantage of this testing system is evident. Abundant experimental data on the cluster thermodynamics may be combined with results obtained with different force-fields. Hybrid QM/MM calculations have been performed to compute free energy differences ΔG for the perturbation from a potassium ion to a sodium ion with respect to the number of water molecules surrounding those ions. In this particular case, one eliminates the force-field for the ion all together. The ion is simply being treated quantum-mechanically, while the solvent surrounding it can be represented by any potential model. To further improve the quality of the model system, we have chosen to use the polarizable Drude model. All QM calculations have been performed using the B88-LYP functional and the DZVP

basis set. The initial minimization has been done using the ABNR method included in CHARMM.

The VV2 algorithm has been employed to deal with system dynamics. Two separate thermostats have been used, the one for the Drude oscillators has been set to 0.1 K and the other one to 315 K and the temperature coupling parameter was set to 10,000 to establish a simulation regime which corresponds to Langevin Dynamics. The reaction path from Na^+ to K^+ and backward has been approximated by 10 windows 20 ps per window with spacing between windows of $\lambda = 0.1$. The performance of the FEP for full quantum calculations remains computationally prohibitive.

The dependence of $\Delta\Delta G_{K\rightarrow Na}$ on the number of water molecules present is shown on Figure 5-5 and it agrees with predicted [55] and experimental [58, 59] data. The upturn at high n values is an artifact of the confined droplet model adopted in this article. The constraint acting on the oxygen atoms of water to prevent droplet evaporation results in entropic penalties for "crowded clusters" where $n > 6$ [57]. The "$\Delta\Delta$" precursor reflects the fact that the difference in bulk hydration free energies was subtracted and plotted dependencies represent the free energy of the transfer from bulk to the confined cluster. The power of such an approach is that ion hydration can be analyzed in the simplified context of the confined cluster with a small number of ligands. It should be noted that results have been proven comparable with experimentally predicted ΔG^{bulk} for the ion in bulk solution, ≈18 kcal/mol once the coordination number has reached $n = 6$. In all models used, the relative free energy starts to approach a bulk-like value after $n = 4$ in good agreement with experimental data from the group of Kebarle [58, 59] and others including theoretical estimates from first principles [60, 61].

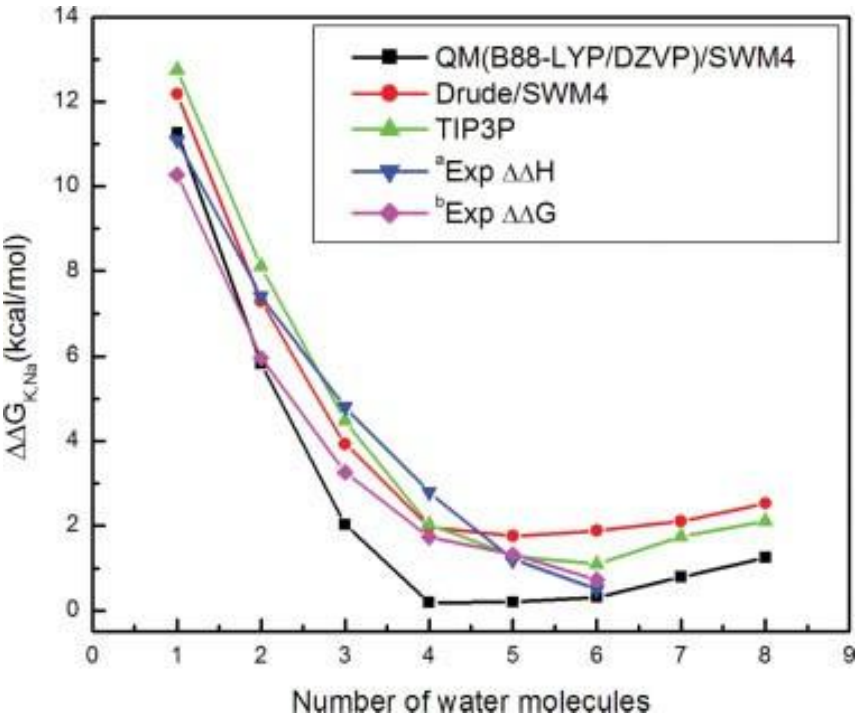


Figure 5-5: Free energy difference versus number of surrounding water molecules [a]Exp,[59] [b]Exp.[58]

Although this is a powerful approach, the outcome shows a nontrivial dependence on the protocol adopted for model building and ligand confinement protocols [56, 57]. Nevertheless, these are good reasons to expect that relative free energies and not enthalpies are more accurate than indicated by this comparison, and to some extent, meaningful computational studies are made possible by relying on the cancellation of errors in relative free energy simulations of reduced models [57, 61].

To illustrate the potential danger in the evaluation of absolute interaction energies alone, we have performed enthalpy computations for Na^+-water and K^+-water systems summarized in Table 5-4. The interaction energies were computed in 8 discrete MD simulations for different values of *n* (1–8) with the temperature being imposed by the Langevin thermostat. It can be noted that absolute values are strongly overestimated. This can be attributed to several factors.

First, for the sake of the method developed we have used the polarizable model of water, instead of treating the whole system quantum-mechanically. Second, many functionals describing ion dynamics are known to overestimate interaction energies being developed to reproduce structural features of the system and proper calibration of the basis set may be necessary [50]. Only the ion was treated quantum-mechanically and thus this is a lesser concern. However, the challenge in computing absolute interaction energies between water and ion is further exaggerated by the lack of ligand exchange with the bulk reservoir, re-orientation of the first solvation shell, etc. Nevertheless, the agreement for the relative free energies is remarkable between all models used in this study (see Fig. 5-4 for details), including nonpolarizable (TIP3P), polarizable (Drude) and QM systems. Importantly, inclusion of polarization or even treating the ion quantum mechanically did not lead to any better results as compared to experiment. Therefore, nonpolarizable models, despite all evident setbacks, can be a very powerful tool to study thermodynamics of ion solvation, given that force-field parameters are developed in a robust and self-consistent way [62]. The correct account of polarization, however, is expected to have a major impact on the description of finer effects such as hydrogen bonding, formation of stable hydrophobic interactions or solvation of multivalent ionic species and/or heavy metals.

Table 5-4: Enthalpy Computations for Na^+-Water and K^+-Water Systems Versus Number of Water Molecules

N	H_{Na^+}	$H_{Na^+}^{exp}$ [58]	H_{K^+}	$H_{K^+}^{exp}$ [58]	$\Delta H_{K^+ \rightarrow Na^+}$	$\Delta H_{K^+ \rightarrow Na^+}^{exp}$ [58]
1	−26.96	−25.0	−19.41	−18.1	−7.54	−6.9
2	−50.96	−44.8	−37.76	−34.2	−13.19	−10.6
3	−70.75	−60.2	−52.33	−47.4	−18.41	−12.8
4	−86.70	−73.4	−66.65	−59.2	−20.04	−14.2
5	−97.78	−84.9	−77.42	−70.6	−20.35	−14.3
6	−108.57	−95.6	−88.43	−79.9	−20.14	−15.7
7	−120.53		−101.69		−18.83	
8	−133.00		−116.57		−16.42	

Enthalpies H, are reported in kcal/mol.

5.5 Conclusions

A hybrid QM/MM interface between CHARMM and deMon has been developed. Like the Q-Chem/CHARMM interface [10] it does not enforce joint compilation and employs external data sharing between those two programs. Therefore, it is able to use all methods implemented in deMon as well as those implemented in CHARMM. To evaluate the interface several test cases were performed. The results were compared to those from the Q-Chem/CHARMM interface, as well as to those experimentally known or predicted. The agreement between those results allows us to conclude that the interface is working and it is capable of dealing with complex problems. The water dimer test showed some overestimation for binding energies, and the influence of the polarizable Drude model implementation, however those results were expected, as similar data has been obtained before [10]. The FEP test proved to be successful as it qualitatively as well as quantitatively reproduced expected results. And, last

but not least, the link atom test produced structures which are indistinguishable from those produced by a similar, working interface.

One of the future directions for this work would be to look closer at the problem of overpolarization and test the "blurred" option for MM charges or other approaches for damping the coulomb interaction.

5.6 Bibliography

1. Warshel, A.; Levitt, M. J Mol Biol 1976, 103, 227249.
2. Singh, U. C.; Kollman, P. A. J Comp Chem 1986, 7, 71830.
3. Field, M. J.; Bash, P. A.; Karplus, M. A. J Comp Chem 1990, 11, 70033.
4. MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E. III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. J Phys Chem B 1998, 102, 3586.
5. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
6. Torras, J.; Seabra Gde, M.; Deumens, E.; Trickey, S. B.; Roitberg, A. E. J Comput Chem 2008, 29, 1564.
7. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. J Chem Theory Comput 2008, 4, 435.
8. Lin, H.; Truhlar, D. G. Theor Chem Acc 2007, 117, 185.
9. Senn, H. M.; Thiel, W. Angew Chem Int Ed 2009, 48, 1198.
10. Woodcock, H. L.; Hodosceck, M.; Gilbert, A. T. B.; Gill, P. M. W.; Schaefer, H. F.; Brooks, B. R. J Comp Chem 2007, 28, 1485.
11. Guest, M. F.; Bush, I. J.; van Dam, H. J. J.; Sherwood, P.; Thomas, J. M. H.; van Lenthe, J. H.; Havenith, R. W. A.; Kendrick, J. Mol Phys 2005, 103,719.
12. Perdew, J. P.; Burke, K.; Ernzerhof, M. Phys Rev Lett 1996, 77, 3865.
13. Perdew, J. P.; Tao, J. M.; Staroverov, V. N.; Scuseria, G. E. J Chem Phys 2004, 120, 6898.
14. Zhang, Y.; Vela, A.; Salahub, D. R. Theor Chem Account 2007, 118, 693.
15. Köster, A. M.; Reveles, J. U.; del Campo, J. M. J Chem Phys 2004, 121, 3417.
16. Geudtner, G.; Janetzko, F.; Köster, A. M.; Vela, A.; Calaminici, P. J Comput Chem 2006, 27, 483.
17. deMon2k, Köster, A. M.; Calaminici, P.; Casida M. E.; Flores-Moreno, R.; Geudtner, G.; Goursot, A.; Heine, T.; Ipatov, A.; Janetzko, F.; del Campo, J. M.; Patchkovskii, S.; Reveles, J. U.; Salahub, D. R.; Vela, A. deMon developers, 2006.
18. Wei, D. Q.; Salahub, D. R.; J Chem Phys 1994, 101, 7633.
19. Wei, D. Q.; Salahub D. R.; Chem Phys Letter 1994, 224, 291.

20. Lin, H.; Truhlar, D. G. J Phys Chem A 2005, 109, 3991.
21. Maseras, F.; Morokuma, K. J Comp Chem 1995, 16, 1170.
22. Froese, R. D. J.; Morokuma, K. Hybrid methods. In von Ragu Schleyer, P (ed); Encyclopedia of Computational Chemistry, vol 2, Wiley, Chichester, 1998; 1244.
23. Humbel, S.; Sieber, S.; Morokuma, K. J Chem Phys 1996, 105, 1959.
24. Vreven, T.; Byun, K. S.; Komromi, I.; Dapprich, S.; Montgomery, J. A., Jr.; Morokuma, K.; Frisch, M. J. J Chem Theory Comput 2006, 2, 815.
25. Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjvoll, M.; Fahmi, A.; Schfer, A.; Lennartz, Ch. J Mol Struct (THEO CHEM) 2003, 632, 1.
26. Elstner, M.; Frauenheim, T.; Suhai, S. J Mol Struct (THEO CHEM) 2003, 632, 29.
27. Walker, R. C.; Crowley, M. F.; Case, D. A. J Comput Chem 2008, 29, 1019.
28. Rizzo, R. C.; Jorgensen, W. L. J Am Chem Soc 1999, 121, 4827.
29. Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. J Phys Chem B 2001, 105, 6474.
30. Price, M. L. P.; Ostrovsky, D.; Jorgensen, W. L. J Comput Chem 2001, 22, 1340.
31. Zhang, Y.; Lee, T.-S.; Yang, W. J Chem Phys 1999, 110, 46.
32. Di Labio, G. A.; Hurley, M. M.; Christiansen, P. A. J Chem Phys 2002, 116, 9578.
33. Ferenczy, G. G.; Rivail, J.-L.; Surjan, P. R.; Naray-Szabo, G. J Comput Chem 1992, 13, 830.
34. Assfeld, X.; Rivail, J.-L. Chem Phys Lett 1996, 263, 100.
35. Ferr Nicolas; Assfeld, X.; Rivail, J.-L. J Comput Chem 2002, 23, 610.
36. Lamoureux, G.; Mackerell, A. D., Jr.; Roux, B. J Chem Phys 2003, 119, 5185.
37. Hao, H,; Weitao, Y. J Chem Phys 2005, 123, 041102.
38. Li, G.; Zhang, X.; Cui, Q. J Phys Chem B 2003, 107, 86438653.
39. Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. J. Comp Chem 1983, 4, 187.
40. Klopper, W.; van Duijneveldt-van de Rijdt, J. G. C. M.; van. Duijneveldt, F. B. Phys Chem Chem Phys 2000, 2, 2227.
41. Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer, H. F.; Quack, M. J Chem Phys 2002, 116, 690.
42. Curtiss, L. A.; Frurip, D. J.; Blander, M. J Chem Phys 1995, 98, 2198.
43. Jorgensen, W.; Chandraserkhar, J.; Mabdura, J.; Impey, R.; Klein, M. J Chem Phys 1983, 79, 926.
44. Grossfield, A.; Ren, P.; Ponder, J. W.; J Am Chem Soc 2003, 125, 15671.
45. Bucher, D.; Raugei, S.; Guidoni, L.; Dal Peraro, M.; Rothlisberger, U.; Carloni, P.; Klein, M. L. Biophys Chem 2006, 124, 292.
46. Roux, B.; Allen, T.; Bernèche, S.; Im, W. Q Re Biophys 2004, 37, 15.
47. Megyes, T.; Bálint, S.; Peter, E.; Grósz, T.; Bakó, I.; Krienke, H.; Bellissent-Funel, M.-C. J Phys Chem B 2009, 113, 4054.
48. Soper, A. K.; Wechström, K. Biophys Chem 2006, 124, 180.
49. Ramania, L. M.; Bernasconi, M.; Parinello, M. J Chem Phys 1999, 111, 1587.
50. Krekeler, W.; Hess, B.; Site, D. L. J Chem Phys 2006, 125, 054305.

CHAPTER SIX: REACTION MECHANISM IN RNAP II – PROTON RELAY VIA COMPETING ROUTES

6.1 Abstract

RNA polymerase II catalyzes the nucleotidyl transfer reaction for messenger RNA synthesis in eukaryotes. Two crystal structures of this system have been resolved, each with its own defects in the coordination sphere of Mg^{2+}(A) resulting from chemical modifications. To remedy these defects, three models were built and equilibrated by molecular dynamics simulations. For each model, a reaction pathway search was performed with quantum mechanical/molecular mechanical potential. The results revealed a proton-transfer-facilitated mechanism. While the acceptor of the initial proton transfer may vary depending on the particular conformation of the active site, all possible routes converge to the same destination. Moreover, comparison between different models indicates that the role of Mg^{2+}(A) is more structural than catalytic.

6.2 Introduction

DNA and RNA are two of the major macromolecules essential for all known forms of life. While DNAs are replicated by various DNA polymerases, RNA transcription is catalyzed by RNA polymerases. RNA polymerase II (RNAP II), responsible for synthesizing messenger RNA, has become the most studied type of RNA polymerase [1-9] due to its critical role in the life cycle of eukaryotes. *In vivo* experiments have shown that RNAP II is capable of selecting nucleotide triphosphates (NTP) complementary to the DNA template with an error rate of 1 per 10^5 nucleotides (nt) [1, 7]. Underlying such high accuracy is the widely acknowledged two-metal ion catalytic mechanism [10], where one metal ion activates the attacking sugar hydroxyl and the other coordinates and stabilizes the departing phosphate group. In the case of RNAP II, these two

metal ions are divalent Mg^{2+}, both of which coordinate with the incoming nucleotide and the enzyme surroundings, as depicted in Fig. 6-1. This figure also sketches a sequence of events during the nucleotidyl transfer reaction: When the 3'O is activated, it approaches P_α to conduct a nucleophilic attack and when a bond is formed between these atoms, the P-O bond between P_α and $O_{\alpha\beta}$ is broken resulting in the departure of the pyrophosphate group (PPi). Although this general scheme provides a promising lead, more details of the reaction are still to be filled in. Numerous studies have investigated in detail the steps of the nucleotidyl transfer reaction in DNA polymerases (DNAP), which are valuable references for RNAP II because of the commonalities between these two polymerases. In this light, questions raised and addressed by studies of DNA polymerases should be worth reviewing.

1. The first key issue of this reaction is how the phosphodiester bond (3'O - P_α) formation coordinates with the P_α - $O_{\alpha\beta}$ bond breaking. Does 3'O - P_α formation precede (associative) or follow (dissociative) P_α - $O_{\alpha\beta}$ breaking? Or rather do they take place simultaneously (concerted)?

Quantum mechanical calculations of the nucleotidyl transfer in DNAP β by Abashkin et al. [11] support an associative mechanism over a dissociative one. Warshel and co-workers also proposed both associative and concerted mechanisms which produce comparable energy barriers on the quantum mechanical/molecular mechanical (QM/MM) free energy surface[12]. The dissociative path in DNAPs has not been reported in the literature, to our knowledge.
Either the associative or concerted mechanism entails activation of the 3'-OH group. Therefore

2. how does 3'-OH of the ribose become active for the nucleophilic attack of the P_α?

Most studies have suggested a necessary deprotonation of 3'-OH prior to the nucleophilic attack. However, the acceptor of this proton differs among the DNA polymerases, and so, integral to question 2 is:

3. what is the proton acceptor?

For DNA polymerases three major candidates have been considered by researchers: An oxygen of the α-phosphate group, an adjacent aspartic acid and a water molecule coordinated with Mg^{2+}. Deprotonation by an oxygen of the α-phosphate group is proposed by Schlick and Bojin [13] and Abashkin et al. [11] in their quantum mechanical studies of DNAP β. Schlick and co-workers later also reported an initial proton abstraction by a water molecule in their QM/MM study of the same enzyme [14]. The water-mediated proton transfer has also been supported by Wang et al. in their QM/MM studies of DNAP IV [15]and T7 [16]. Deprotonation by an adjacent aspartic acid is suggested by Cisneros et al. in their QM/MM investigation of DNA polymerase λ[17], by Lin et al. [18] and by Florian et al. [19] in QM/MM studies of DNAP β and T7, respectively. Variation of the proton acceptor possibly stems from the different computational methods as well as structural models employed as pointed out in [14] and [12].

After deprotonation, 3'-O is ready to attack and form a bond with P_α, upon which the bond to $O_{\alpha\beta}$ is broken. One would also be curious if this P_α - $O_{\alpha\beta}$ weakening is also facilitated by a proton transfer to $O_{\alpha\beta}$. If so,

4. what should be the proton donor in this case?

An experimental study conducted by Castro et al. [20] proved the necessity of this proton transfer for the speed/fidelity of DNAPs. They postulated an adjacent lysine as a probable proton

donor in various DNAPs. Lior-Hoffmann et al. found through QM/MM umbrella sampling simulation that in DNAP κ, the second proton transfer to the pyrophosphate could be the result of a proton relay from the 3'-OH [21]. A similar result was also found for DNAP β[14].

The above findings for DNAPs have provided valuable information for the studies of RNAP II. To determine the proton acceptor of 3'-OH, Ramos and co-workers recently conducted a comprehensive computational study of the 3'-OH deprotonation [22] in RNAP II. In this study, they compared an aspartic acid, water molecule, P_α, and a hydroxide ion for proton acceptor suitability. Ultimately they determined a hydroxide ion from the bulk solvent to be most suitable. This deprotonation step was also investigated by Salahub and co-workers with a different starting crystal structure [23, 24]. In regard to the possible proton transfer to the pyrophosphate (PPi), Ramos and co-workers studied an adjacent protonated histidine as the proton donor, and found it to be favorable for the nucleotidyl transfer [22]. However, they showed that the histidine protonated O_β of the β-phosphate instead of the bridge $O_{\alpha\beta}$. In their molecular dynamics study of the PPi release, Huang and co-workers also affirmed the contribution of this histidine to the stabilization of the PPi, although their final leaving form of the PPi was unprotonated due to the positive charges in its exit channel [4]. So far, all the studies of nucleotidyl transfer in RNAP II have assumed an associative mechanism of the 3'O-P attack and the dissociative path has not been considered to our knowledge.

After examining the variety of results of 3'-OH activation in DNAPs, one is prompted to be careful with the choice of starting crystal structure in the case of RNAP II. Two major crystal structures of RNAP II have been employed in computational studies which are 2E2H [25] and 2E2J [25] in PDB code. To obtain a complex with the substrate bound in the active site, chemical modifications were made in both structures. In 2E2H, the 3'-OH of the RNA primer was

removed and in 2E2J, the $O_{\alpha\beta}$ was substituted with a methylene group. The modification in 2E2H results in no coordination between the 3'O and the Mg^{2+} (A) (Fig. 6-2A) while the NTP is in good coordination with both Mg^{2+}. The modification in 2E2J leads to a large gap between the RNA primer and the NTP as a result of the weak interaction between the Mg^{2+} and the triphosphate of the NTP. Unlike in 2E2H though, the 3'O coordinates well with Mg^{2+} (A) (Fig. 6-2B). We have performed molecular dynamics (MD) with these two starting structures; even after a 1-ns-long equilibration, these two defects still cannot be alleviated. Even when a constraint was imposed on the distance between the 3'O and Mg^{2+} (A) the coordination disappeared as soon as the constraint was released, as observed by Ramos and co-workers [22]. To overcome these defects and understand the importance of the Mg coordination, we conducted calculations with models built on both structures and a model combining the two (details in Methodology).

In this paper, we attempt to answer the three questions mentioned above about the nucleotidyl transfer reaction in RNAP II by exploring the 1- and 2-dimensional potential energy using various QM and QM/MM approaches. We also compare the results of three different models to delineate the functions of Mg ions and enzyme surroundings in the two-metal ion mechanism.

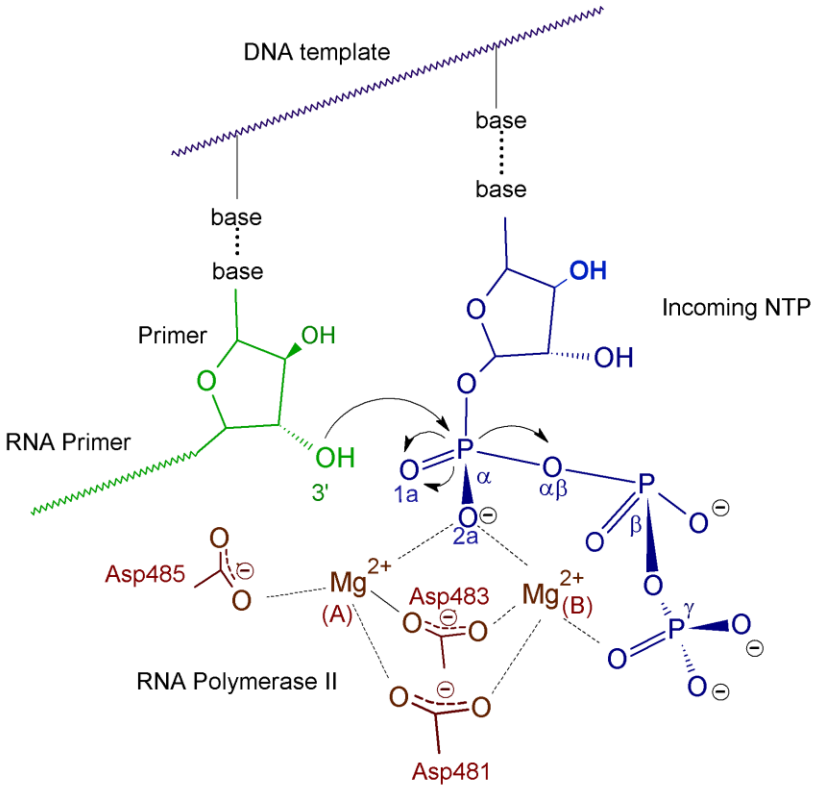


Figure 6-1: The two-metal ion mechanism in RNAP II

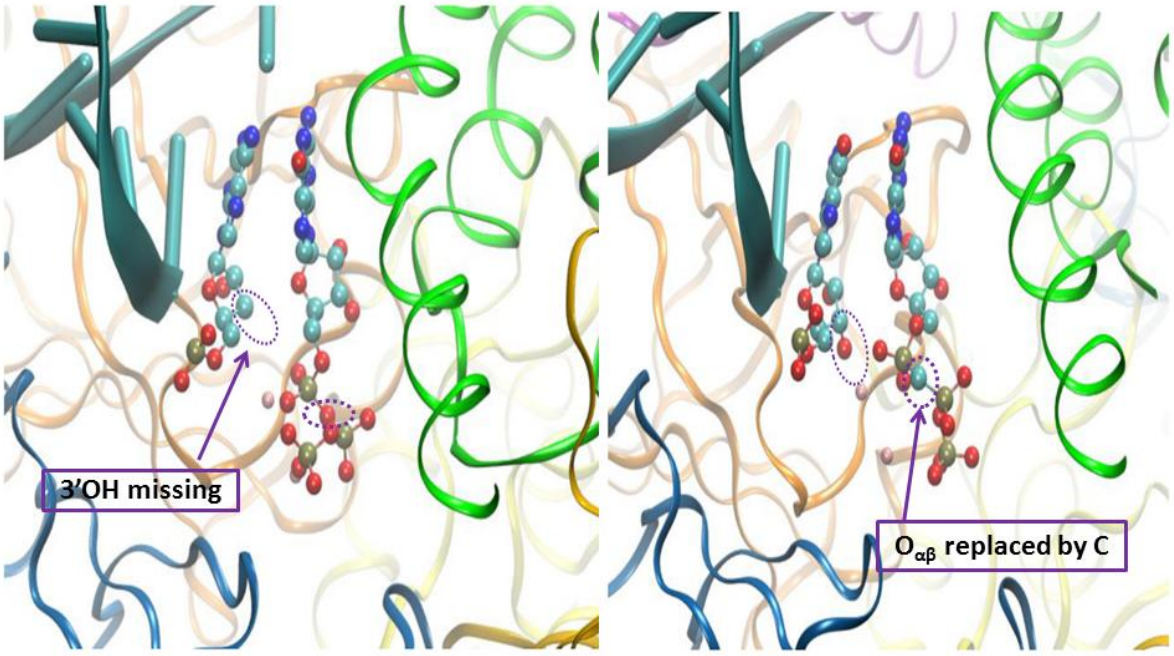


Figure 6-2: Active sites of the crystal structures of 2E2H (A) and 2E2J (B)

A) Active site in 2E2H B) Active site in 2E2J where the 3'-O and $O_{\alpha\beta}$ positions are circled in purple dashes

6.3 Methodology

6.3.1 System setup

6.3.1.1 MM models

Models are constructed based on crystal structures of the ternary elongation complex with a GTP (PDB ID: 2E2H) or a GMPCPP (PDB ID: 2E2J) in the active site [25]. In both structures, a number of residues were not resolved because of structural disorder. In the subunit Rpb1 of 2E2H, missing residues 1446-1733 at the end of the chain were ignored because they are not

important to the core function of RNAP II and modeling of large surface loops is unreliable. Missing non-end residues, 156-160, 186-191, 315-318 and 1232-1235, which are not missing in another crystal structure (2E2I) were added by adopting the same psi and phi angles as in 2E2I. Missing non-end residues, 192-198, 1177-1186 and 1244-1253 were inserted by manually entering the psi and phi angles to fit with other residues. The same protocol was followed for other subunits of 2E2H and 2E2J. After all necessary missing residues were restored, a geometry optimization was performed with non-missing residues constrained. The missing 3'-O atom was added based on the topology in the CHARMM 27 force field [26]. The bridge carbon atom of GMPCPP in 2E2J was replaced with oxygen.

Models built on 2E2H and 2E2J are termed Model-1 and Model-2, respectively. A third model, Model-3, was based on 2E2J where the coordinates of the GMPCPP were replaced by the coordinates of the GTP in 2E2H followed by a restrained optimization when everything except the GTP and Mg ions were restrained.

Protonation states of titratable residues were determined by pKa calculation through the GBMV module [27] in CHARMM [28]. In the case of histidine, the site that has the lowest calculated pKa was protonated. Protonation states were held the same in all the models for consistency. Nucleotide triphosphates were deprotonated in all models and therefore carry a charge of -4 [2, 29].

Each model was fully solvated in a cubic box of explicit water with a length of ~160 Å. To neutralize the system, a total of 88 Na^+ ions were added by randomly replacing the water molecules at the surface of the box. As a result, each model comprises a total of ~350,000 atoms.

6.3.1.2 Hybrid QM/MM models

Since the periphery of the enzyme does not have a significant influence on the phosphotidyl transfer reaction in the active site, a subsystem within 20A (4641 atoms in total) from the P_α of the substrate is selected to reduce the computational cost. During all QM/MM calculations, the boundary of this subsystem is held fixed to maintain the protein structure. For all three models mentioned above, the QM part (Fig. 6-3) includes the entire GTP substrate, the ribose of the last residue of the RNA primer, both Mg cations, Asp481, Asp483, Asp485 and Arg446 of RNAP subunit A, Arg766, Arg 769 and Arg 1020 of subunit B and 3 water molecules. At the boundary between the QM and MM systems, hydrogen link atoms are added to saturate the cutoff bonds, thus resulting in 144 QM atoms with a total charge -1. Initial structures are selected from the respective MD trajectories.

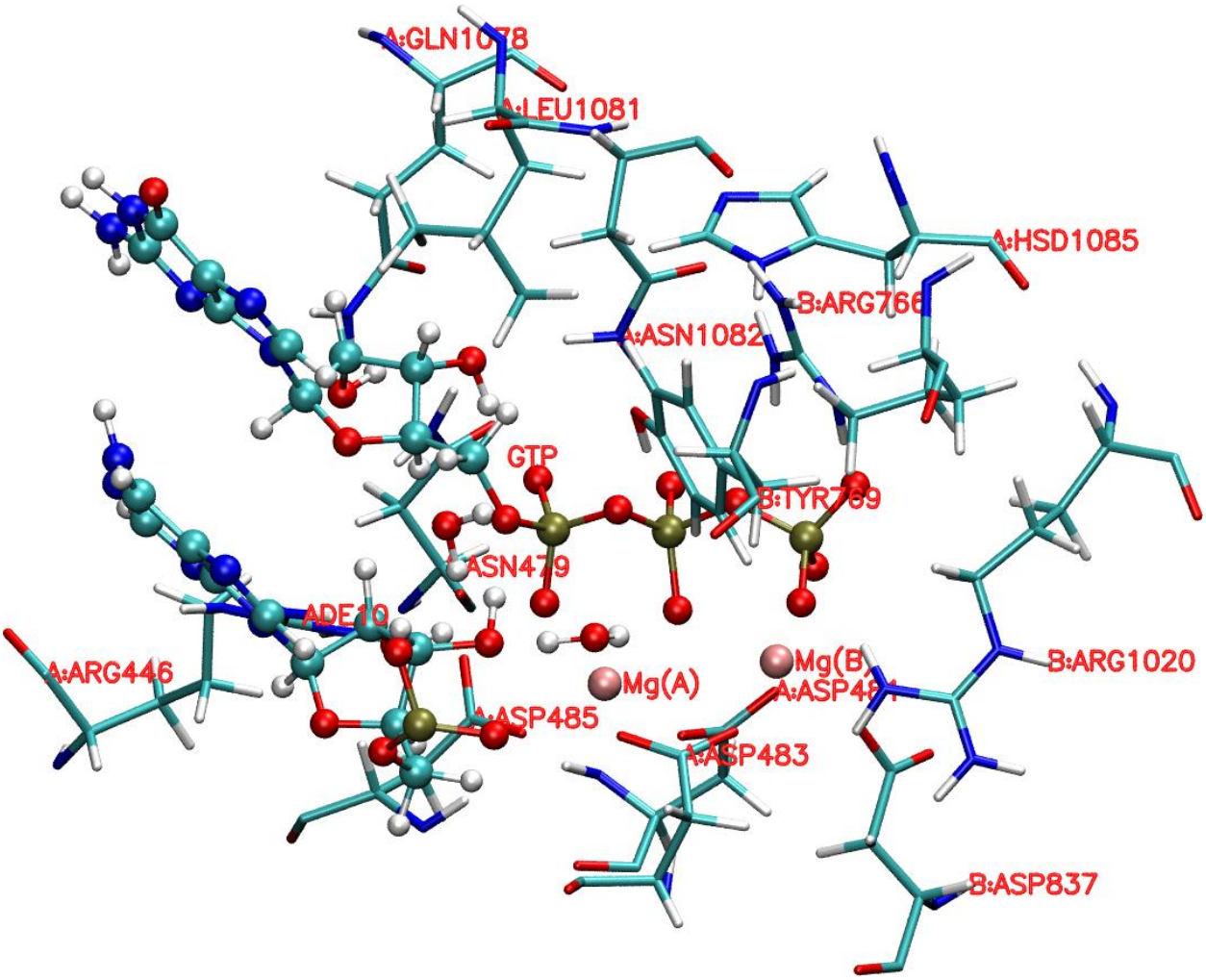


Figure 6-3: The QM section of the QM/MM model

The GTP and the last residue of the RNA primer are shown as balls and sticks, and the rest of the QM atoms as licorice

6.3.1.3 QM model

A QM model was built to benchmark the performance of the semiempirical AM1/d-PhoT method in comparison with other QM methods. This QM model is built based on the QM part of the QM/MM model of Model-1, where aspartic acids are simplified as acetic acids, riboses substituted with methyl groups and arginines are removed, which are comprised of 49 atoms with a total charge of -3.

6.3.2 Simulations

6.3.2.1 Molecular mechanical MD

The CHARMM 27 force field [26, 30] was used to describe the protein and nuclei acids; explicit water was modeled with the TIP3P model [30]; all metal ions were modeled with the CHARMM 27 force field except Mg^{2+} of which the charge remained as 2+ but van der Waals parameters were modified to vdW radius R^* = 1.300 Å, well depth ε = 0.06 to avoid overestimation of Mg-O coordination according to previous studies [19, 31]. Periodic boundary conditions were applied and the particle mesh Ewald summation was used to obtain accurate electrostatic interactions. Langevin-type thermostat and barostat were used to maintain the temperature at 300K and the pressure at 1bar. All systems were subject to an optimization of 10000 steps and an equilibration of 200 ps before production runs with a time step of 1fs. All simulations were performed with NAMD 2.9 [32] and analyzed with VMD 1.9.1 [33].

6.3.2.2 QM and benchmarking of AM1/d-PhoT

AM1/d-PhoT is an Austin Model 1(AM1) method specifically parametrized for phosphoryl transfer reactions [34]. Since this reaction is of the same nature as the phosphotidyl transfer reaction in RNAP II, AM1/d-PhoT is also adopted in this work to expedite the calculation with an adequate accuracy. To validate this method in our study, an extensive benchmarking has been performed. In the benchmarking procedures, a relaxed surface scan (details below) of the 3'O- P_α distance was first performed on the QM model with the PBE functional and double zeta basis sets SVP. And then single point calculations of geometries obtained for the scan were performed with various QM methods including density functional theory (DFT) and semiempirical methods. DFT methods include both pure (PBE/TZVP) and hybrid (B3LYP/TZVP) functionals while semiempirical methods include AM1, PM6 and

AM1/d-PhoT. DFT calculations were conducted with ORCA 2.9.1 [35] and its interface with the pDynamo program library [36]. Semiempirical calculations were performed with the pDynamo program library [36]. The results are plotted in Fig. 6-4 where AM1/d-PhoT, outperforming AM1 and PM6, is in good agreement with B3LYP at all points except the transition state. At the transition state, compared to B3LYP, AM1/d-PhoT tends to overestimate the activation energy by 7kcal/mol whereas PBE underestimates it by 6kcal/mol. This is in line with, if not better than, the mean unsigned error of 8.34kcal/mol for the activation energy of phosphoryl transfer reactions as calculated in [34].

To evaluate the accuracy of AM1/d-PhoT on geometry predictions, geometry optimization was conducted using the structures obtained from the scan with PBE as starting structures. Key distances of the reactants and intermediates are compared between optimizations with AM1/d-PhoT and PBE as summarized in Table S6-1(Supporting information). Most of the distances are reproduced within a difference of 0.1 Å with the highest deviation of 0.187 Å for hydrogen bond length which is a difficulty for most QM methods. This is close to the mean unsigned error of 0.073 Å for hydrogen bond species in [34]. Both energy calculations and geometry optimization have demonstrated that AM1/d-PhoT is an adequate method for the study of the nucleotidyl transfer reaction with a reasonable compromise between accuracy and speed. Moreover, its application has also proven to be successful in the studies of the phosphoryl transfer reaction in hairpin ribozyme [37], hammerhead ribozyme [38], and RNA 2'-O-transesterification [39].

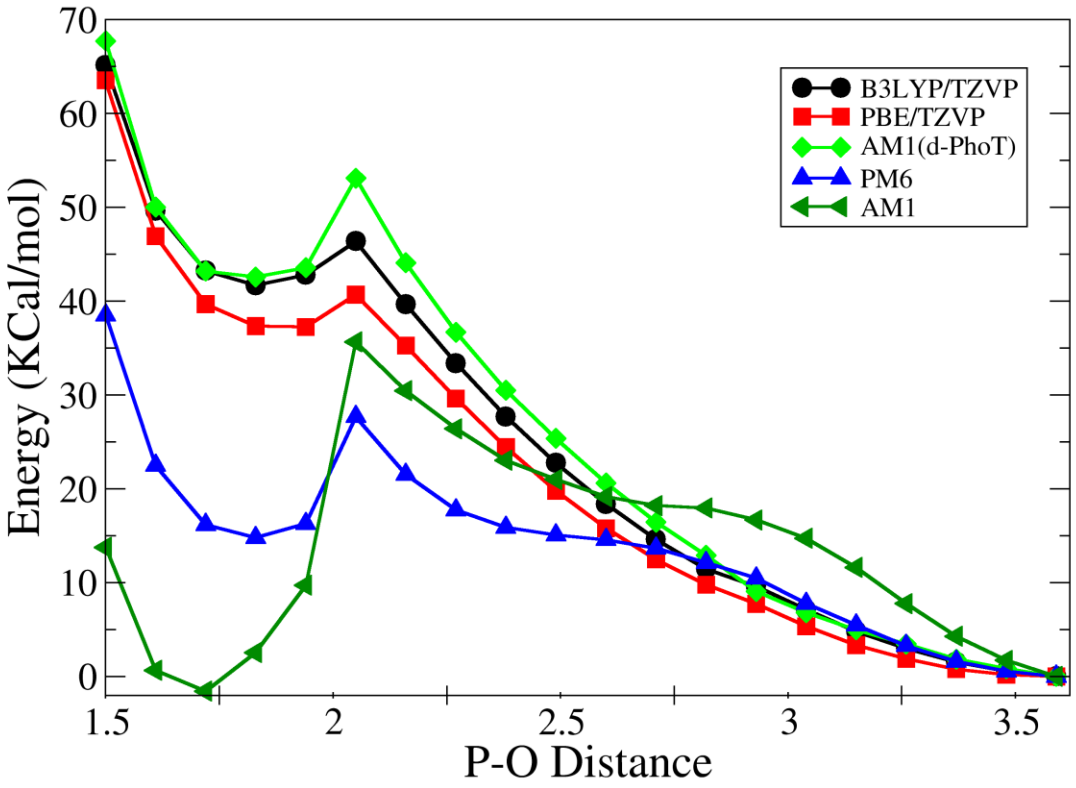


Figure 6-4: Benchmarking of AM1/d-PhoT

6.3.2.3 QM/MM

AM1(d-PhoT) /MM calculations are performed with the pDynamo program library [36] and DFT/MM calculations with the CHARMM-deMon2k interface [40] at the PBE/DZVP level. Parameters for MM and QM/MM interactions are taken from the CHARMM27 force field [26, 30]. In MD simulations with QM/MM, a Langevin-type thermostat and barostat were used to maintain the temperature at 300K and the pressure at 1bar. All systems were subject to an optimization and an equilibration of 2ps before production runs with a time step of 1fs.

6.3.2.4 Relaxed surface scan

A harmonic potential with a coefficient of 2500kJ/(mol·Å^2) (~595kcal/mol·Å^2) was applied to the reaction coordinates during geometry optimizations. There were 20 or 40 steps in 1-D relaxed surface scans and 20 in each dimension in all 2-D scans. The potential energy of each scanned geometry was obtained by single point calculation of the geometry from the scans excluding the harmonic potential.

6.4 Results and discussion

6.4.1 Mg(A) coordination in the active site

Mg(A), structurally defining the active site, is crucial for the function of RNAP II. However, its coordination composition is different between 2E2H and 2E2J, even after long MD equilibration where the chemical defects are corrected. A 1-ns-long MD trajectory of Model-1 (based on 2E2H) with the 3'-OH of the RNA primer added shows that Mg(A) coordinates with Asp481, Asp483, Asp485, the α- and β-phosphate of the incoming GTP, and a solvent water which was not present in the original crystal structure (Fig. 6-5A). This same coordination composition in 2E2H has also been observed by Ramos and colleagues [22]. Notably, the 3'-O remains distant from Mg(A) and thus there is no evident coordination (Fig. 6-6A). After this MD simulation, a 500-ps-long MD trajectory(Fig. 6-6B) with AM1(d-PhoT)/MM and a 3-ps-long one(Fig. 6-6C) with DFT/MM also reveal the same coordination sphere and no direct interaction between 3'-O and Mg(A). A restrained MD simulation at the DFT/MM level with a restraint on the distance between 3'-O and Mg(A) at 2.2Å, resulted in a coordination between them but they became distant once the restraint was removed. This was also found by Ramos and colleagues [22] in their constrained MD simulation.

In Model-2 (based on 2E2J) with the $O_{\alpha\beta}$ restored, a 1-ns-long MD trajectory shows that Mg(A) coordinates with Asp481, Asp483, Asp485, 3'-O of the RNA primer, and two solvent water molecules not present in the original crystal structure (Fig. 6-5B). Compared with trajectories of 2E2H, the 3'-O (Fig. 6-6D) and a water molecule coordinate with the Mg(A) in the place of the triphosphate backbone of the incoming GTP. This is caused by the large gap between Mg(A) and the substrate which creates space for waters to come in since small solvent molecules move much faster than the substrate in MD simulations. And this gap in the initial structure is due to the substitution of a carbon for the $O_{\alpha\beta}$, which resulted in the substrate being trapped slightly outside of the active site.

To investigate if Mg(A) can coordinate with both 3'-O and the triphosphate, Model-3 was built as a combination of Model-1 and Model-2 (details in Methodology). A 1-ns-long trajectory (Fig. 6-6E) of this model demonstrates that Mg(A) can coordinate with both 3'-O of the RNA primer and the α-phosphate of the substrate while Asp481, Asp483, Asp485 and a water molecule remain in contact (Fig. 6-5C). This trajectory was then followed by another 11-ns-long MD to further examine the stability of the coordination sphere. Interestingly, the coordination between Mg(A) and 3'-O broke after 2ns and remained broken afterwards (Fig. 6-6F). The lost coordination is compensated for by Asp485.

To understand the interaction between Mg(A) and the 3'-OH, we performed a relaxed surface scan (RSS) with AM1(d-PhoT)/MM on the distance between them in Model-1. As shown in Fig. 6-7, the first barrier with a significant height is at 3.5Å which is in agreement with Figures 6-3A, 3B and 3C where the accessible distances are always above 3.5Å. The high barrier at 2.1Å is the most difficult to cross which suggests that a strong coordination is almost impossible to form. This explains why the short distance is unable to be maintained even after

the constrained MD at 2.2Å on Model-1. It also explains why the 3'O-Mg(A) coordination is broken after 2ns in the MD of Model-3 as 3'O-Mg distance in Model-3 starts at more than 2.2Å and is easy to fall back and become larger. All our results above suggest that Mg(A) does not coordinate strongly with the 3'-OH when the phosphotidyl transfer reaction does not take place. However, should the reaction happen and the proton be transferred from the 3'-OH, the basicity of 3'-O would be considerably increased and thus a strong coordinate could be formed so as to stabilize the intermediates during the reaction.

It has been pointed out that "the inherent lability of divalent magnesium must be considered in any functional role for the metal ion in the context of enzyme mechanisms, in as much as the structurally most stable configuration may not represent the functionally competent coordination state of the metal during enzyme turnover" [41] . Moreover, studies also show that the reaction path is quite sensitive to the choice of the crystal structure [42, 43]. In this light, the Mg-3'O distance could be an important factor for the reaction. To understand the role of Mg coordination and retrieve a reasonable reaction barrier, we performed reaction path searches with QM and QM/MM methods on all of Model-1, Model-2 and Model-3. In the calculations, one starting structure from Model-1 and Model-2, respectively, was employed, and two starting structures of Model-3 were employed-- one from the 1-ns-long trajectory (Model-3A) and the other from the 12-ns-long trajectory (Model-3B).

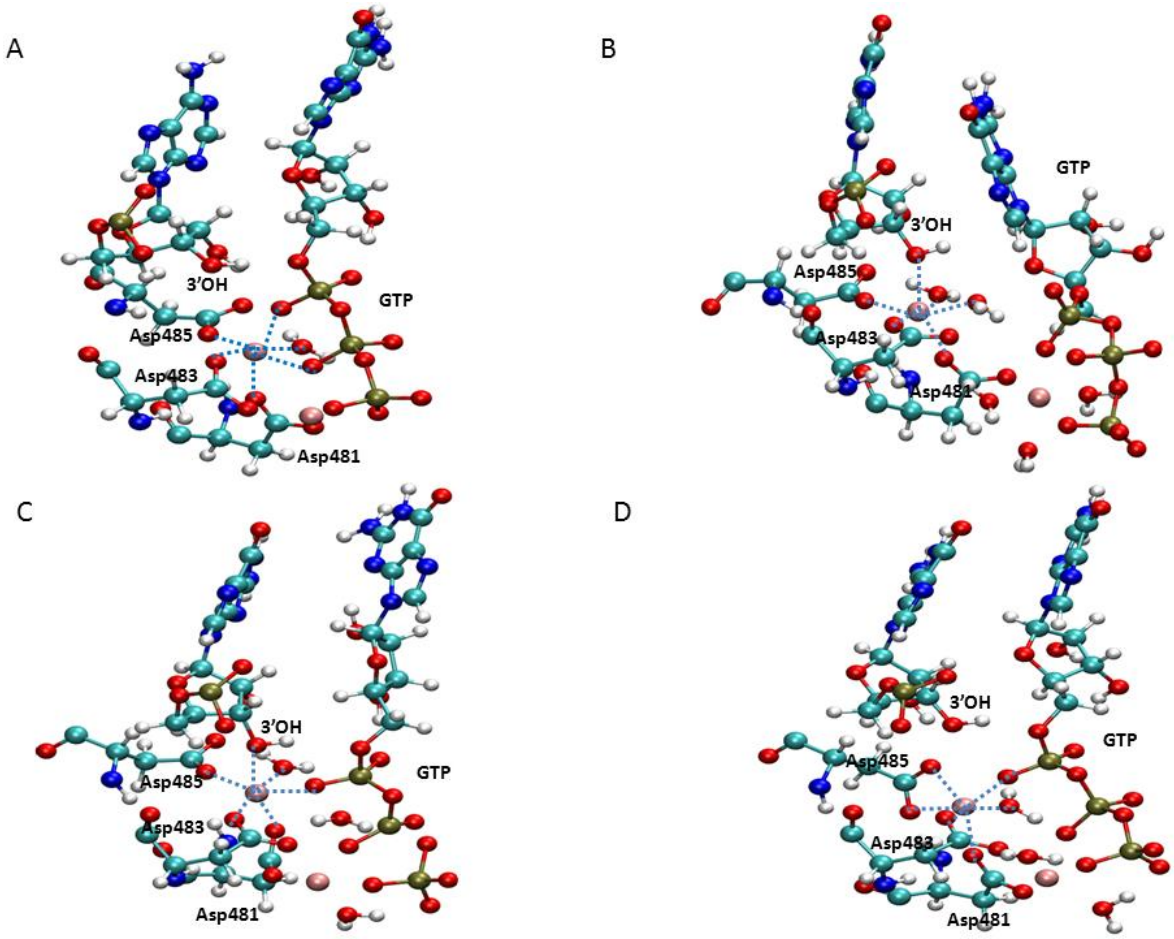


Figure 6-5: Coordination spheres of different models

A) Model-1 B) Model-2 C) Model-3 after 1ns of MD (Model 3-A) D) Model-3 after 12ns of MD (Model 3-B)

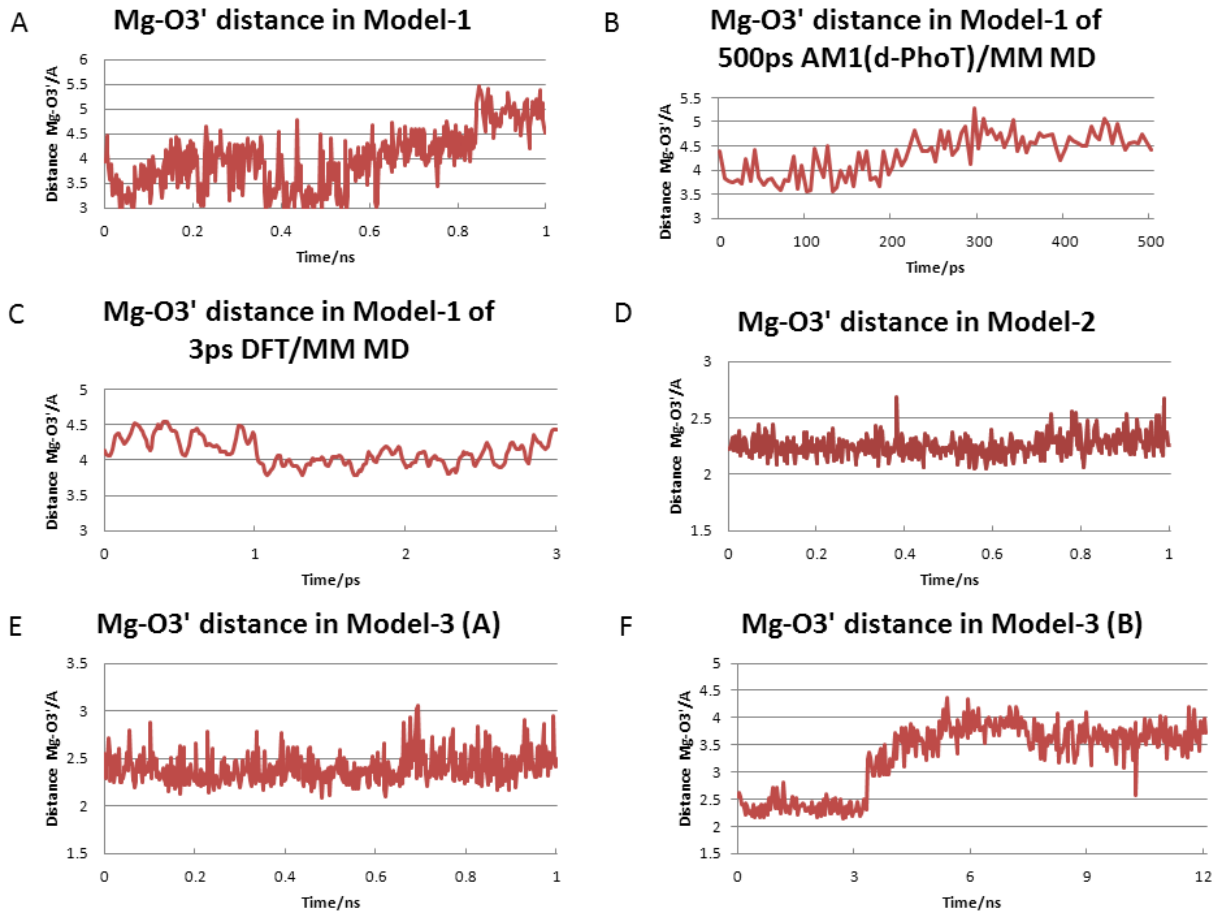


Figure 6-6: Distance between Mg(A) and 3'-O from trajectories of different models

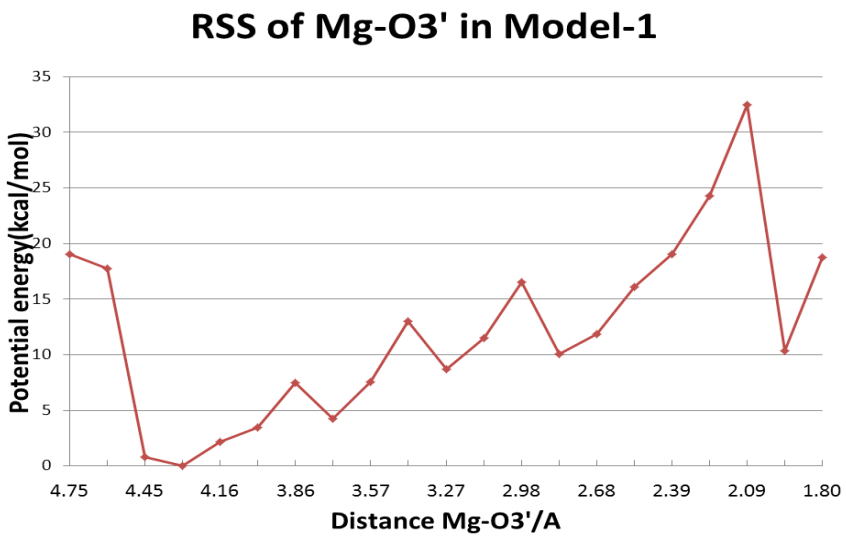


Figure 6-7: Potential energy with respect to the Mg-O3' distance from relaxed surface scan

6.4.2 Nucleophilic attack of 3'-O on P_α

During the nucleotidyl transfer reaction, the 3'-O and P_α approach each other and form a bond while the proton is transferred from the 3'-OH group. Both AM1(d-PhoT)/MM and DFT/MM MD trajectories (Fig. 6-8A and 6-8B) of Model-1 show that the 3'-O remains at a distance of more than 3A from the P_α, whereas a distance of less than 1.9 indicates a bonded interaction. Since this distance is a major index for the progress of the reaction, we first attempted to scan the system along this distance using a relaxed surface scan with AM1(d-PhoT)/MM on Model-1. Geometries along this path show that the 3'-H is still attached to the 3'-O even when the 3'-O is already bonded with the P_α, which results in a high barrier (50kcal/mol) and a high reaction energy (40kcal/mol) as shown in Fig. 6-8C. The same calculations were performed on Model-2 and also produced similar results (Figure S6-1A in Supporting Information). This energy barrier is much higher than the experimental one of 18.1 kcal/mol as estimated from the enzyme turnover rate [22]. This indicates that the proton 3'-H does not

naturally leave when only the 3'O- P_α distance is chosen as the reaction coordinate; another dimension is necessary. Therefore, the proton transfer should be included in the reaction coordinates.

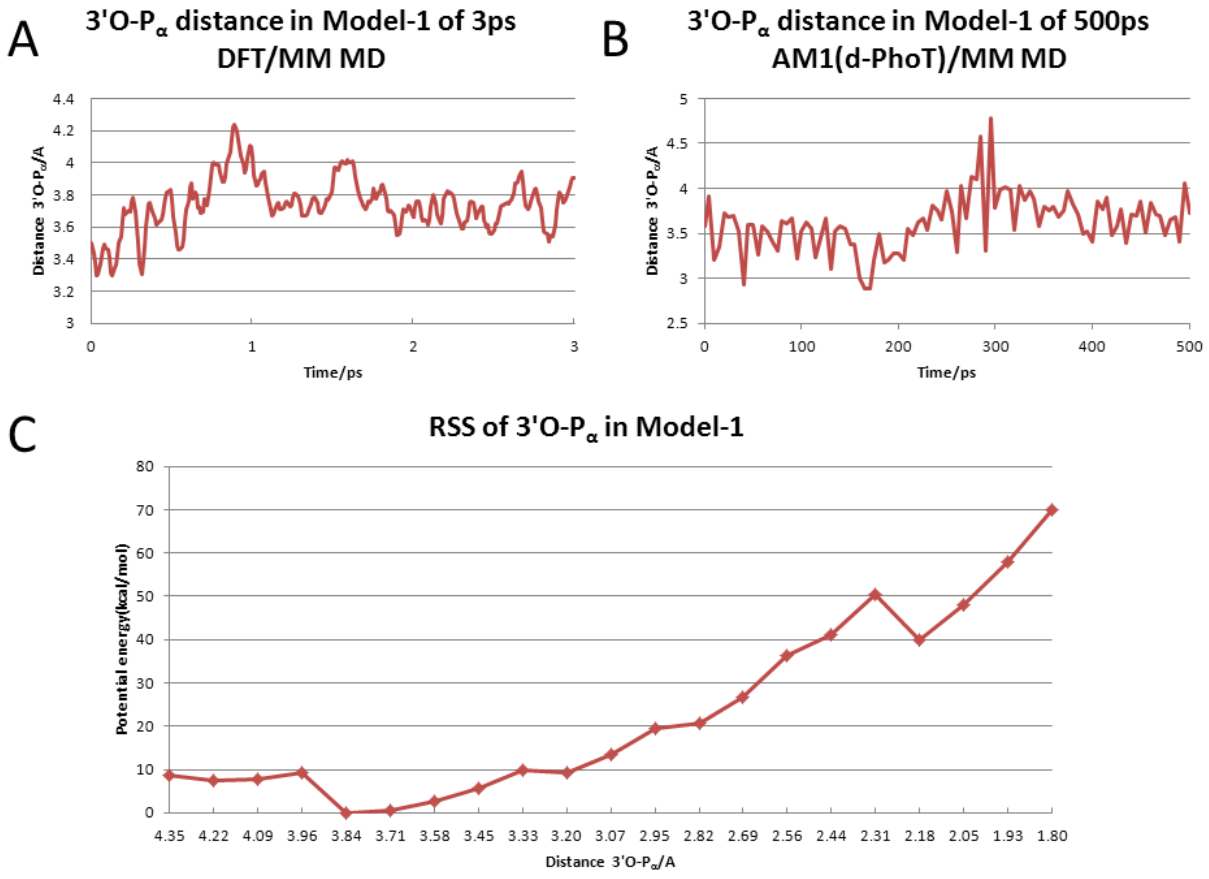


Figure 6-8: Plots of the 3'O- P_α distance in Model 1

A) 3'O- P_α distance in the DFT/MM MD trajectory B) 3'O- P_α distance in the AM1(d-PhoT)/MM MD trajectory C) Relaxed surface scan of the 3'O- P_α distance

6.4.3 Deprotonation of 3'-OH

To include the 3'-H in the reaction coordinates, a proton acceptor will need to be identified. As a first attempt, we chose the O1a atom of the α-phosphate group as the proton acceptor. To ensure the proton transfer, the 3'H-O1a distance is scanned first followed by

scanning of the 3'O- P_α distance in Model-1; thus assuming a step-wise mechanism. The first step of proton transfer produces a barrier of 32.94kcal/mol (Fig. 6-9A) and the second step of nucleophilic attack a barrier of 30.44kcal/mol (Fig. 6-9B), both of which are considerably lower than that of the path with only 3'O- P_α as the reaction coordinate. Similar results are also observed for the scan in Model-2 (Figure S6-1B and S6-1C in Supporting Information). Although our second reaction path includes two dimensions resulting in a lower barrier, each step is still one dimensional and may cause artifacts by not driving the system along the other dimension. Therefore, it is essential to scan both dimensions simultaneously so as to determine whether the reaction is step-wise or concerted, and recover a realistic reaction barrier.

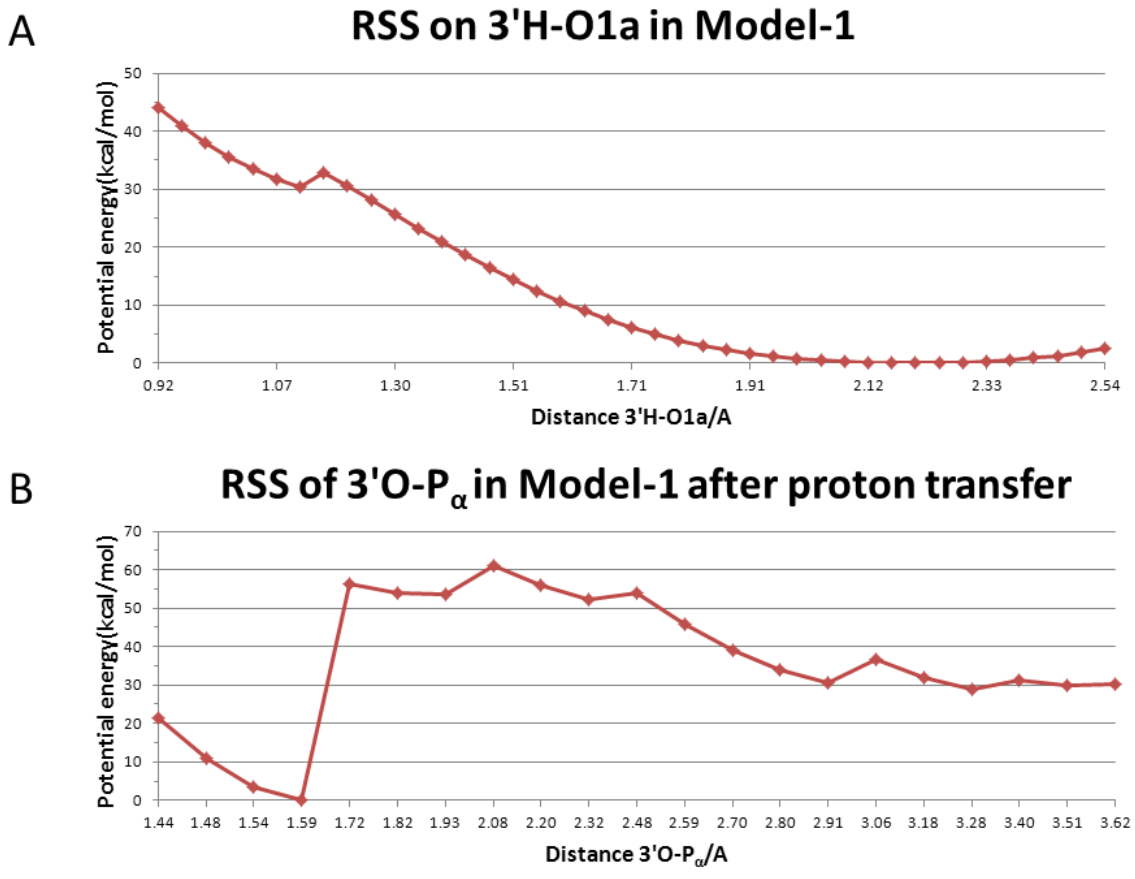


Figure 6-9: Relaxed surface scans in Model 1

A) Scan of the 3'H-O1a distance B) Scan of the 3'O- P_α distance

6.4.4 2-D scans on all models with different proton acceptors

To search all possible reaction paths, potential proton acceptors other than the α-phosphate such as Asp483 and the coordinated water molecule should also be considered (Fig. 10). For each possible proton acceptor, relaxed surface scan using the AM1(d-PhoT)/MM potential has been performed on four starting structures. In each proton transfer scenario, the scanned geometries and the potential energy maps are analyzed and compared. Since the path calculated in this fashion is a minimum energy path, the one with the lowest overall energy barrier will be selected as the final result.

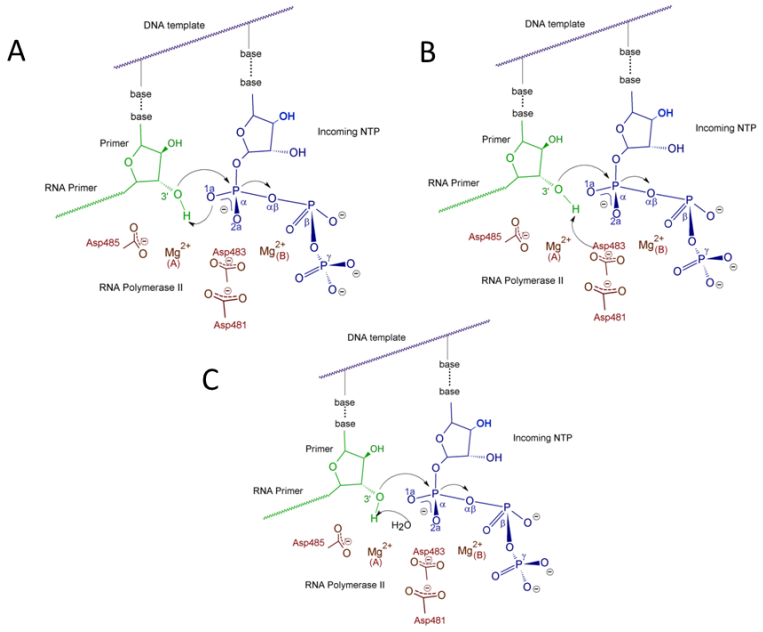


Figure 6-10: 3'-H transfer to different proton acceptors

A) 3'-H transfer to the α-phosphate B) 3'-H transfer to Asp483 C) Proton transfer to the coordinated water

6.4.4.1 Proton transfer to α-phosphate

The 2-D maps of potential energy are shown in Fig. 6-11. The reaction starts at the top-right region where both distances are large and finishes at the bottom-left region where both distances are small, indicating complete proton transfer and nucleophilic attack. Stationary points along the path for Model-3(B) are illustrated in Fig. 6-12, where the intermediate (IM) has the 3'H-O1a hydrogen bond formed, the transition state (TS) has pentavalent organization around the P_α and the product has the 3'O- P_α bond formed. Among all four maps, the general trend is that the reaction first proceeds in the x direction where the 3'H-O1a distance decreases and a hydrogen bond is being formed until the proton is transferred. Following the proton transfer, the reaction is mainly driven by the association of the 3'-O and P_α. A few differences exist between these maps. In Model-1, a steep "ridge" (30-35 kcal/mol) presents an obstacle on the path of proton transfer. This is due to the intervention of Asp483 which in conjuction with O1a "sandwiches" the proton in between and keeps it far from the 3'-O when the 3'H-O1a distance is constrained at 1.4 Å and the O3'- P_α distance larger than 2 Å. In contrast, in other models the proton is kept between the 3'-O and O1a unless the 3'-O and P_α are close. Since the pKa of the 3'-OH is much greater than that of the Asp483, the 3'-O is a better proton stabilizer during the transfer. Ramos and co-workers [22] also found an overall barrier of 35.0 kcal/mol for this pathway using a structure similar to our Model-1. Moreover, in Model-3(A) and Model-3(B), the scan results also show hydrogen bond interactions among the O1a, 3'-H and Asp483, to stabilize the 3'-H when the O3'- P_α distance is less than 1.8 Å after the proton transfer. These observations prompt us to consider Asp483 as a probable alternative proton acceptor, which shall be discussed in the next section. Another difference among the 2-D maps is the singularity of Model-2 which predicts a barrier greater than 50kcal/mol. This is due to the large gap between

the 3'-OH and the α-phosphate of the substrate, as mentioned in 3.1. Additionally, the 3'-O remains tightly bound to the Mg(A) (their distance is between 2 and 2.7 Å for the entire 2D surface) which must first be loosened to approach the P_α, thus resulting in a high barrier for the nucleophilic attack. This implies that the gain of the stabilization by the coordination does not compensate for the loss of mobility of the 3'-O. The third distinction is the minimum, indicating a clear product region in Model-3(B). In contrast, Model-1 and Model-3(A) do not produce an identifiable minimum in the product region. The most significant difference between the starting structure of Model-3(B) and that of Model-1 is that an oxygen of the β-phosphate is not coordinated with Mg(A) in Model-3(B), a direct contrast from Model-1. The difference between the starting structure of Model-3(B) and that of Model-3(A) is that the 3'O-Mg coordination is not present in Model-3(B). These two differences may cause subtle changes in the active site when the product is formed. This suggests that it is important and helpful to select different initial structures for reaction path search.

With regard to the 3'O-Mg coordination, it should be noted that the coordination in Model-3(A) is weakened (2.7-3.3Å) before the proton transfer and broken (>3.8Å) after the proton transfer. As for the dissociation of P_α-$O_{\alpha\beta}$, their distance increases from 1.6Å to 2.2Å during the nucleophilic attack, which suggests an associative mechanism since the P_α-$O_{\alpha\beta}$ bond dissociates spontaneously when the 3'O-P_α is formed. To sum up these four maps, Model-1and Model-2 predict relatively high barriers, while Model-3(A) and Model-3(B) yield reasonable barriers that are comparable and provide a more realistic description of the low-energy reaction pathway. Thus, when the proton is transferred to the α-phosphate, the barrier in terms of potential energy for the nucleotidyl transfer can be estimated as between 15 and 20kcal/mol. The results are summarized in Table 6-1.

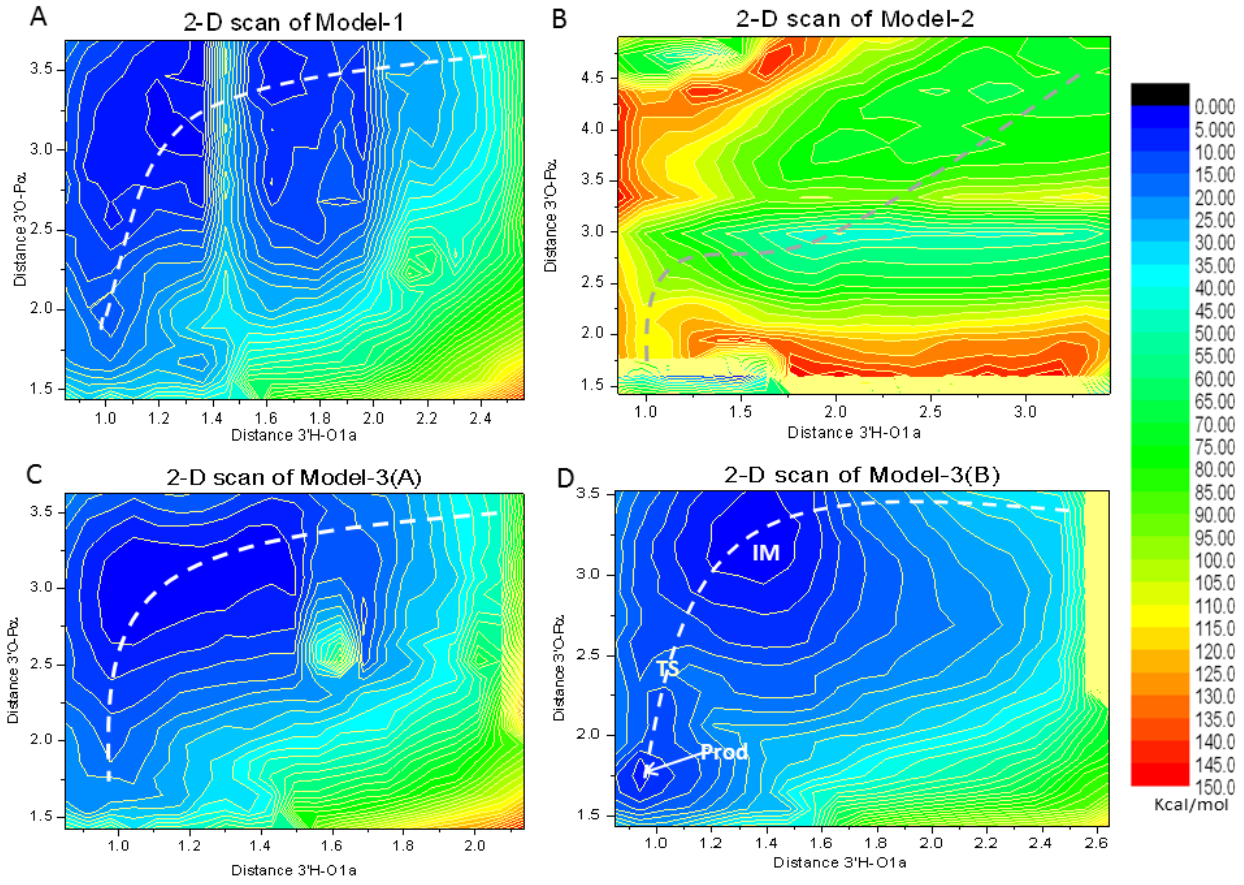


Figure 6-11: 2-D potential energy maps of all models when the 3'-H is transferred to the α-phosphate

The dashed lines indicate approximate paths.

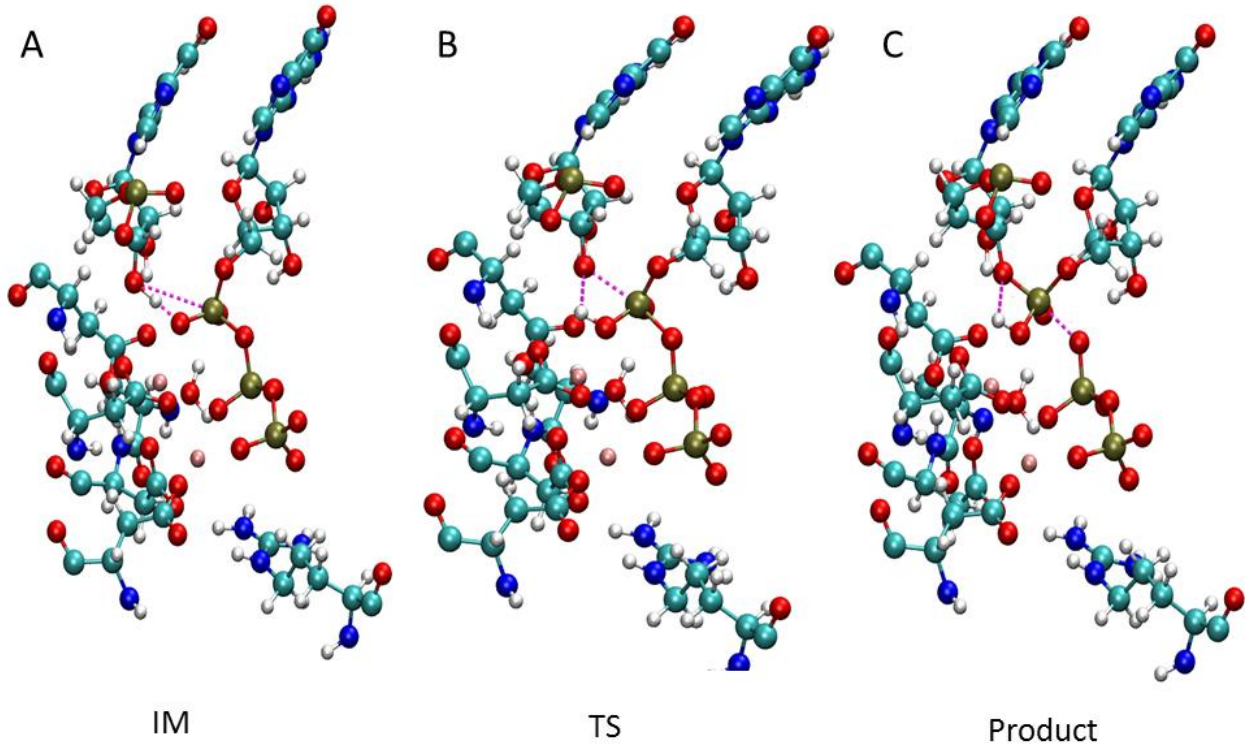


Figure 6-12: Structures of the intermediate, transition state and product from the scan of Model-3(B)

Purple dashed lines indicate forming and breaking bonds.

Table 6-1: Summary of key parameters for all models when the proton is transferred to the α-phosphate

Key parameters	Model-1	Model-2	Model-3(A)	Model-3(B)
Overall barrier(kcal/mol)	35	>50	20	15-20
Proton transfer destination	α-phosphate	α-phosphate	α-phosphate	α-phosphate

3’O-Mg coordination broken	Y	N	Y	Y
P_{α}-$O_{\alpha\beta}$ dissociation	Y	Y	Y	Y

6.4.4.2 Proton transfer to Asp483

In the 2-D potential energy maps in Fig. 6-13, where the proton is instead transferred to Asp483, the general trend is similar to the energy scans where the α-phosphate is the proton acceptor. Explicitly, the 3'-H approaches the proton acceptor, Asp483, and following this, the 3'-O attacks the P_{α}. However, differences also exist among these four maps. The most obvious difference lies between Model-2 and the rest. The unrealistically high barrier produced by Model-2 is due to the strong coordination between the 3'-O and Mg(A) which makes the nucleophilic attack difficult. This is similar to the case where the α-phosphate is the proton acceptor. A minor difference is found between the map of Model-3(A) and those of Model-1 and Model-3(B). In Model-3(A), the reaction progresses through two minima before it reaches the final product minimum, where the first minimum indicates a stable hydrogen bond and the second the complete proton transfer. In the other two models, after the hydrogen bond is formed, the reaction follows a largely concerted path to the product region. When checking the geometries, we found the Asp483-Mg(A) coordination in Model-3(A) is much weaker than that in the other two models while the 3'H-Asp483 distance is short. The Asp483-Mg(A) distance in Model-3(A) increases from 2.3 to 3.3 Å after the proton transfer whereas that distance in the other two models remains between 2.1 and 2.4 Å. Moreover, the energy barrier is more favorable

in Model-3(A) than the barrier in the other two models, explainable by the weaker coordination. This energy barrier difference also derives from the difference in their initial structures. In Model-3(A), the 3'-OH and Asp483 are both coordinated with the Mg(A), thus resulting in an easier proton transfer. Conversely, in the other two models, only Asp483 coordinates with the Mg(A), thus leading to a higher barrier. When the 3'-OH coordinates with the Mg(A), it is very likely that the 3'-H is transferred to Asp483 as yields a lower barrier than when the proton is transferred to the α-phosphate. Under this condition, the potential energy barriers should be 10 kcal/mol for the proton transfer and 10 kcal/mol for the nucleophilic attack.

The lower energy cost to transfer the proton to the aspartic acid than to the α-phosphate can be explained by their pKa difference. The pKa of aspartic acid in solution is 3.86 while the pKa values of phosphate diesters such as dimethyl-, di-n-propyl, and di-n-butyl-phosphate which are analogues to the α-phosphate, is less than 2 [44]. Therefore the aspartic acid is more capable for accepting a proton. This being said, the geometries also show that the α-phosphate helps stabilize the proton when the 3'-O and P_α are within the bonding range. With respect to the breaking of the 3'O-Mg coordination and the dissociation of P_α-$O_{\alpha\beta}$, these two events are also observed in the scanned geometries. The results are summarized in Table 6-2.

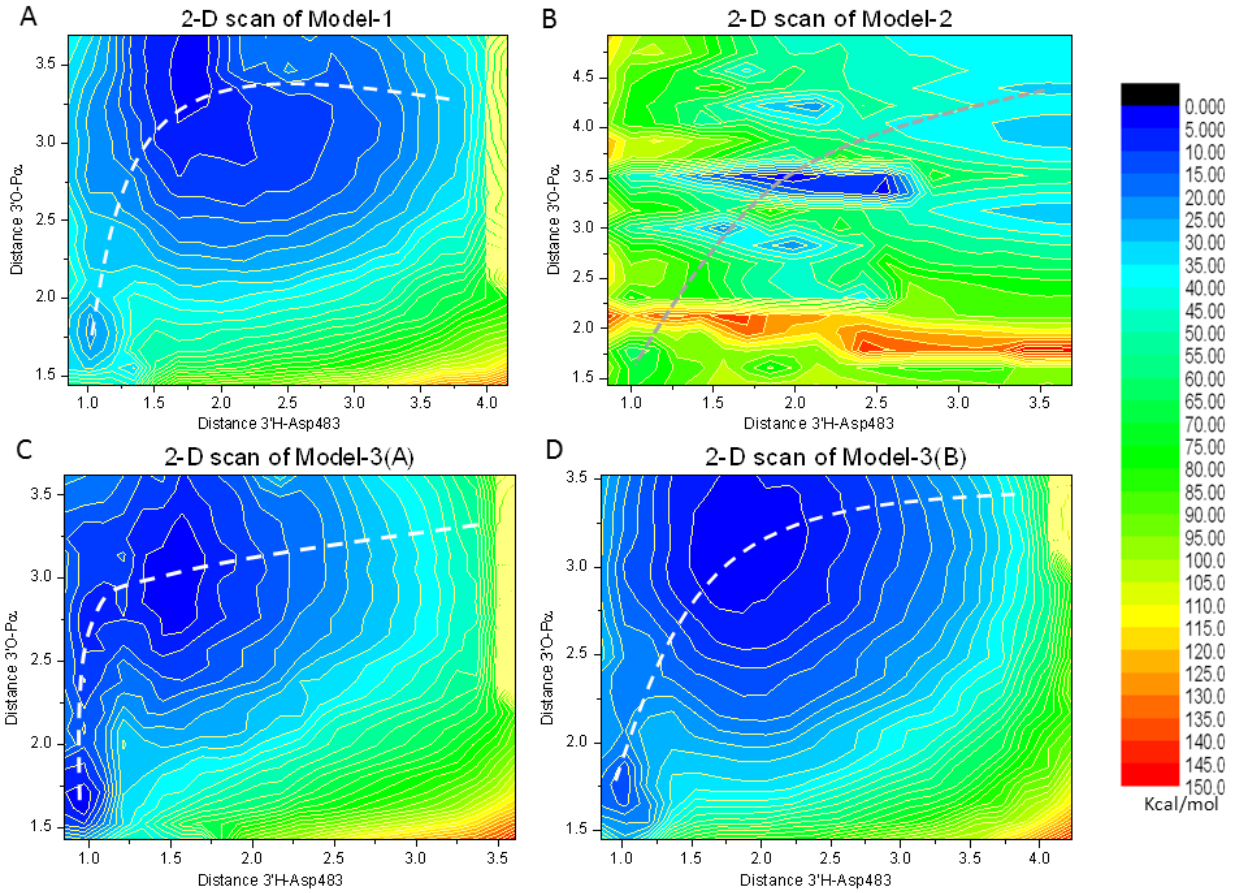


Figure 6-13: 2-D potential energy maps of all models when the 3'-H is transferred to Asp483

The dashed lines indicate approximate paths.

Table 6-2: Summary of key parameters for all models when the proton is transferred to Asp483

Key parameters	Model-1	Model-2	Model-3(A)	Model-3(B)
Overall barrier(kcal/mol)	40	>50	10	25-30
Proton transfer destination	Asp483	Asp483	Asp483	Asp483

3’O-Mg coordination broken	Y	N	Y	Y
P_{α}-$O_{\alpha\beta}$ dissociation	Y	Y	Y	Y

6.4.4.3 Proton transfer to water molecule

The general trend is similar to the previous two cases: Proton transfer precedes nucleophilic attack. Comparing the four maps (Fig. 6-14), Model-1 and Model-3(B) predict a higher barrier in the proton transfer than that in the nucleophilic attack while Model-2 and Model-3(A) predict the opposite. Between the former two models and the latter two, the major difference is the 3'O-Mg coordination. It is conceivable that when the hydronium ion is formed, it causes repulsion between itself and the cationic Mg which then needs to be stabilized by the anionic 3'-O.When checking the geometries, we found that indeed the 3'O-Mg distance shortens when the hydronium ion is formed in the initial structures where the 3'O-Mg coordination exists. This enhanced 3'O-Mg coordination makes the nucleophilic attack difficult, resulting in a higher barrier for this step.

Since the hydronium ion is not stable, the proton is likely to be further transferred to another proton acceptor. In Model-1 and Model-3(A), the proton is eventually transferred to an adjacent aspartic acid – Asp485 via the water, and in Model-3(B), it is transferred to O2a of the α-phosphate whereas in Model-2, the proton adheres to the water. Geometries show that in Model-2, Asp485 hydrogen bonds to the nearby Arg446 and the α-phosphate distantly from the hydronium, which prevents the proton transfer from the hydronium ion. In the other models, both

Asp485 and O2a of the α-phosphate help stabilize the hydronium through a hydrogen bond network. Scans of Model-3(A) and Model-3(B) produce comparable barriers of 20kcal/mol, so the overall barrier height of the 3'-H transfer to the coordinated water is estimated as 20kcal/mol. Therefore, according to the barrier heights, the destination of the proton in this case can be either the α-phosphate or Asp485. The results are summarized in Table 6-3.

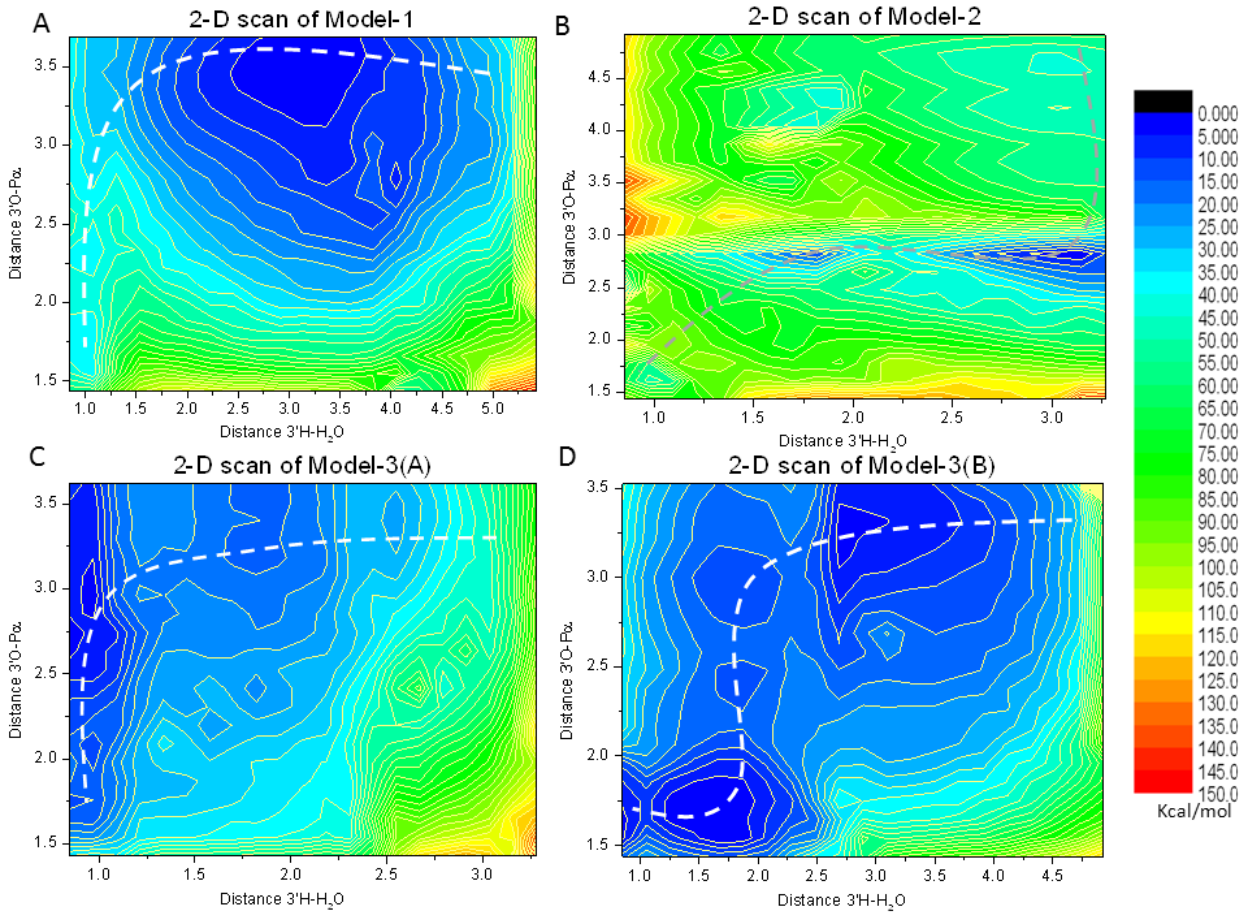


Figure 6-14 2-D potential energy maps of all models when the 3'-H is transferred to the coordinated water. The dashed lines indicate approximate paths.

Table 6-3: Summary of key parameters for all models when the proton is transferred to water

Key parameters	Model-1	Model-2	Model-3(A)	Model-3(B)
Overall barrier(kcal/mol)	50	>50	20	20
Proton transfer destination	Asp485 via water	water	Asp485 via water	α-phosphate via water
3'O-Mg coordination broken	Y	N	N	Y
P_{α}-$O_{\alpha\beta}$ dissociation	Y	Y	Y	Y

6.4.4.4 Summary of the proton transfer

Among all the possible proton transfers, the transfer to Asp483 yields a lower overall barrier of 10 kcal/mol and the other two transfers are comparable as both are in the range of 15 to 20kcal/mol. Of all the proton transfers, the destination varies which could be the α-phosphate, Asp483 or Asp485, as the 2D scans suggest. One would wonder if these different destinations can be connected and what the resulting barrier would be if this proton further transfers from Asp483 to the α-phosphate. A 1-D scan was performed along the distance between the proton acceptor oxygen of Asp483 and the O1a atom of α-phosphate on Model-3(A). The overall barrier of this process turns out to be 23.89 kcal/mol as shown in Fig. 6-15, which makes it comparable to the barriers of the other two transfers. (And this proton can be further transferred to the

pyrophosphate as will be discussed in the following section.) As for Asp485, if the proton is transferred from it to the α-phosphate, the proton will have to pass by the coordinated water or Asp483, and therefore its barrier should be no less than that of Asp483. Reviewing all the proton transfers, if the 3'-H is assumed to be transferred to the α-phosphate (O1a or O2a atom), it can be either a direct transfer or an indirect transfer via Asp483 or a water molecule. In fact, a recent study on DNA polymerase has proposed that the 3'-H proton transfer can take multiple pathways [45].

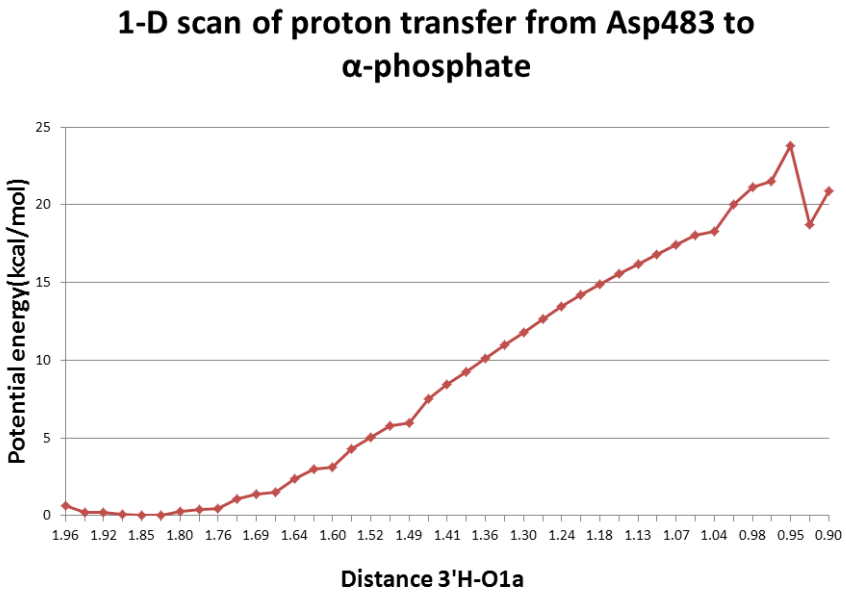


Figure 6-15: Scan of the proton transfer from Asp483 to the α-phosphate

6.4.5 P_{α}-$O_{\alpha\beta}$ dissociation

An integral part of the nucleotidyl transfer reaction is the leaving of the pyrophosphate upon the dissociation of the P_{α}-$O_{\alpha\beta}$ bond. As observed in scans of the proton transfers, this bond is weakened during the nucleophilic attack, indicating an associative mechanism. To further

confirm this finding, a 1-D scan (Fig. 6-16A) was performed on the P_{α}-$O_{\alpha\beta}$ distance after the 3'O- P_{α} bond is formed. This scan produces a barrier of 11.09 kcal/mol for the pyrophosphate dissociation which is considerably lower than that of the proton transfer and nucleophilic attack (approximately 20kcal/mol). To further investigate the connection between the proton transfer and the P_{α}-$O_{\alpha\beta}$ dissociation, a 2-D scan including also the 3'H- $O_{\alpha\beta}$ distance was performed after the 3'H is transferred to the α-phosphate. The 2-D scan (Fig. 6-16B) suggests an even lower barrier of approximately 5 kcal/mol when the proton is finally relayed to the $O_{\alpha\beta}$ atom of the leaving pyrophosphate. These findings all point to the fact that the relayed proton in this process serves as a mediator to lower the barriers and ease the tensions among parties engaged in the reaction. The proton relay from the nucleophilic hydroxyl to a leaving group has been proposed in DNA polymerases [14-16, 21] and hairpin ribozyme [46].

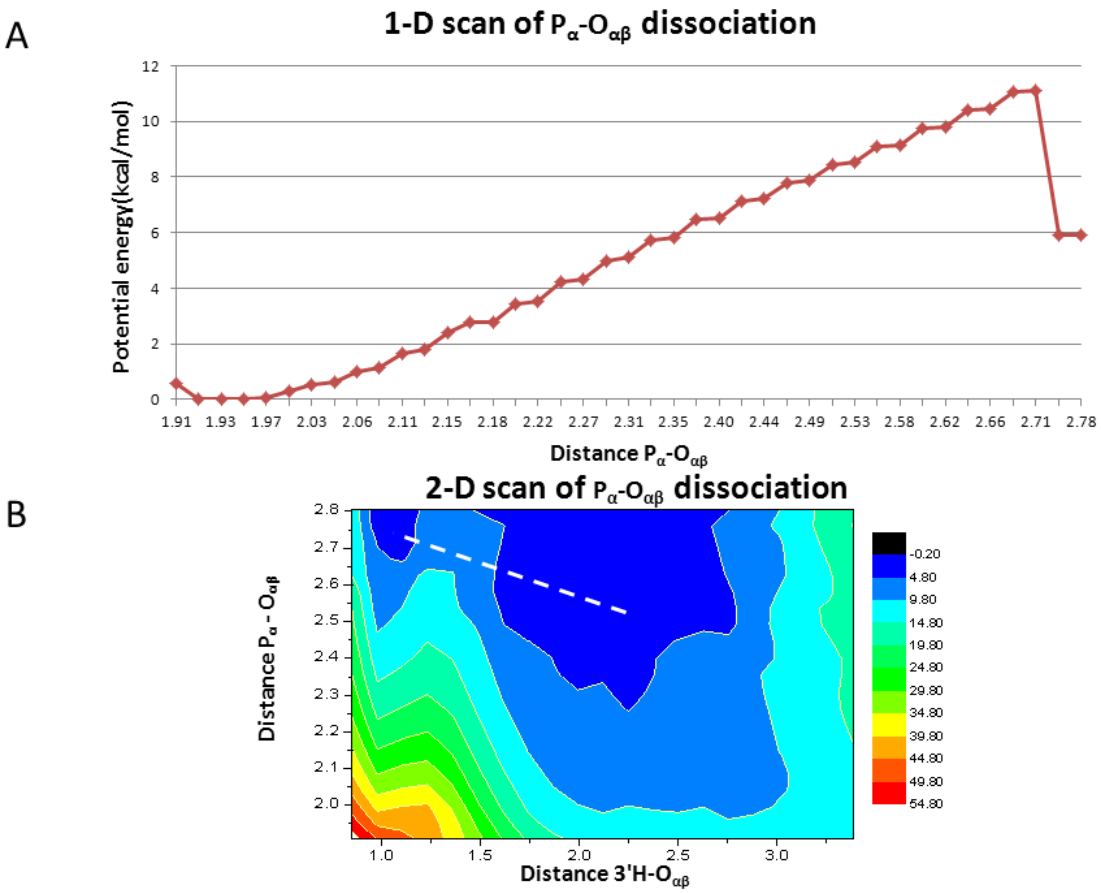


Figure 6-16: Scans for the P_{α}-$O_{\alpha\beta}$ dissociation

6.5 Conclusion

In this study, we have built 3 models and employed molecular dynamics and relaxed surface scan techniques to investigate the nucleotidyl transfer reaction in RNAP II. The results show that the 3'-H is transferred to the α-phosphate either directly or indirectly, facilitating the formation of the 3'O-P_{α} bond and the weakening of the P_{α}-$O_{\alpha\beta}$ bond. Following this, the 3'-H migrates to the $O_{\alpha\beta}$, resulting in the pyrophosphate leaving. The quintessential part of this mechanism is that the proton so efficiently mediates among different parties engaged in the reaction to facilitate the P-O bond forming and breaking. Although the acceptor of the initial

proton transfer may vary depending on the particular conformation of the active site, all possible routes converge to the same destination. A different mechanism of RNAP II was proposed by Ramos and co-workers [22] which requires a hydroxide ion to deprotonate the 3'-OH in the initial proton transfer. The overall barrier of the reaction in the enzyme turned out to be as low as 9.8 kcal/mol. However, it is conceivable that such a high reactivity of the hydroxide ion comes at the price of a low stability. Their thermodynamic integration shows an unfavorable energy difference of +7.5 kcal/mol between the hydroxide ion in the active site and in solution. A recent study of hydroxide ion in solution using infrared spectroscopy finds that it is stable up to ~1.5 ps [47], which suggests an even shorter lifetime of hydroxide ion in the enzyme as it is less stable than in solution. Ramos et al assumed that the hydroxide ion comes from the bulk [22]. However, compared to the turnover number (0.16 s^{-1}) of RNAP II [48], the short lifetime of the hydroxide ion in the enzyme makes it very likely to be reacted during its diffusion into the active site. In addition, since the stability of the hydroxide ion in the active site is quite low, study of the dynamic behavior of the hydroxide ion in the active site should be conducted to test the feasibility of this mechanism. It is more plausible that the enzyme in this case utilizes the structure of its own active site rather than external help to catalyze the reaction.

Regarding the difference in the 3'O-Mg coordination among 3 models, our results show that this coordination is not required for the reaction to proceed as it is evidently broken or weak in most of the scans that produce low energy barriers. Therefore, the role of Mg(A) in RNAP II appears to be more structural than catalytic. In addition, these theoretically built models also allowed us to overcome defects in the crystal structures and explore more potentially interesting regions in the conformational space of the system.

The QM part of the system was described with the semiempirical AM1/d-PhoT method in this study due to practical limitation of resources and the substantial amount of computation that would be required by higher-lever methods. Although the benchmarking results of AM1/d-PhoT show adequate accuracy, higher-level methods such as B3LYP are still more desirable to calculate reaction barriers. For this reason, we took caution to base our conclusions on the ranges of the numbers from calculation instead of comparing them in decimals. (This action is also dictated by the nature of the scan method of which the plotted contour lines identify meaningful regions rather than single points.) To improve the accuracy, adjustments to the results of higher-level methods such as empirical correction should be made. Possible improvements will be explored in our future work.

The 2-D surfaces obtained in this study are only based on potential energies which do not include entropy effects. Free energy surface, as more comparable to experiment results, should be pursued in our future work. This current study nonetheless provides useful insights into the reaction mechanism since its original purpose was to compare different models and identify the minimum energy pathway of the reaction. The knowledge gained from this preliminary work will be helpful for more advanced calculations.

6.6 Supporting information

Tabel S6-1: Comparison between the geometries optimized by DFT and AM1/d-PhoT (Distances are given in Å)

	DFT (reactant)	AM1/d-PhoT (reactant)	AM1/d-PhoT (intermediate)	DFT (intermediate)
O(Acceptor)-H(Nucleophile)	1.750	1.937	1.004	0.989

Mg-OP(2)	2.109	2.168	2.098	2.035
Mg-O(Acceptor)	2.141	2.182	2.300	2.247
P-O(Nucleophile)	3.588	3.642	1.856	1.897
MG-O(acetate)-1	2.179	2.255	2.161	2.138
MG-O(acetate)-2	2.123	2.180	2.172	2.107
MG-O(acetate)-3	2.085	2.130	2.141	2.095

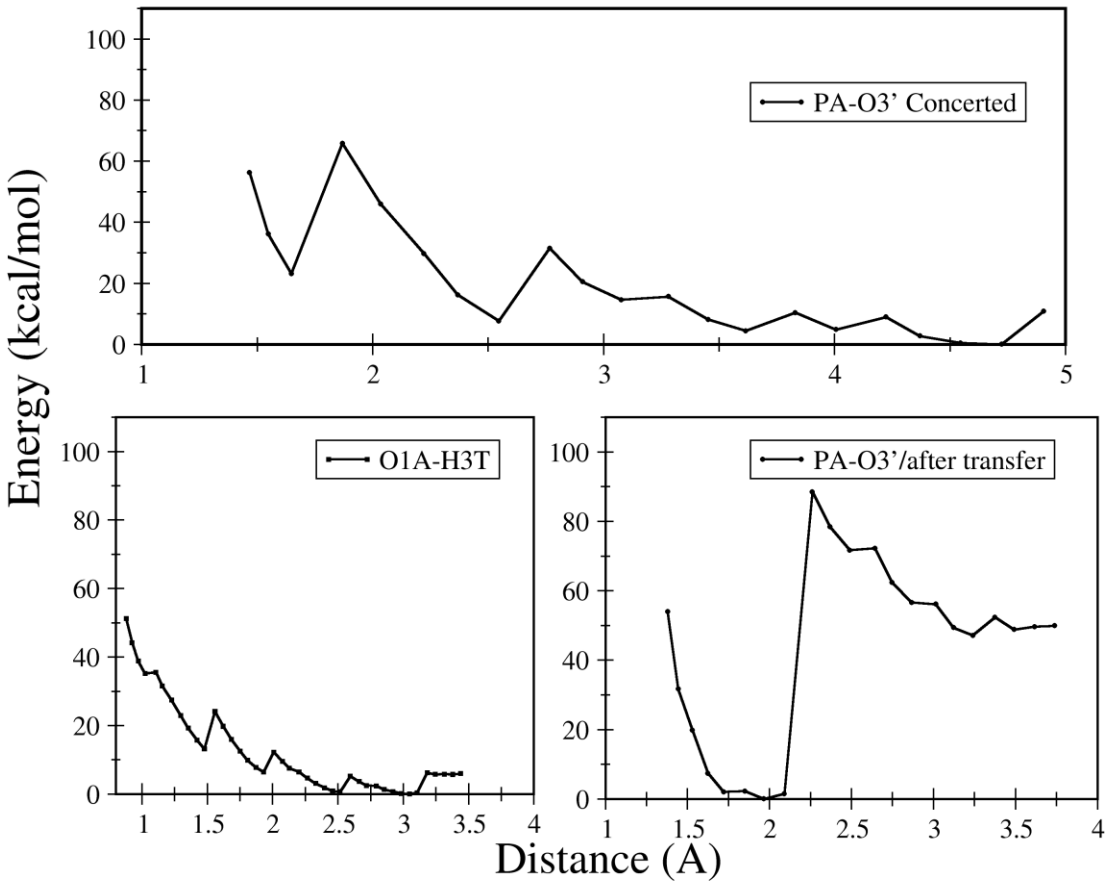


Figure S6-1: Relaxed surface scans in Model-1

6.7 Bibliography

1. Ardehali, MB, Lis, JT. 2009. *Nat Struct Mol Biol* 16: 1123
2. Batada, NN, Westover, KD, Bushnell, DA, Levitt, M, Kornberg, RD. 2004. *P Natl Acad Sci USA* 101: 17361
3. Cramer, P, Bushnell, DA, Fu, JH, Gnatt, AL, Maier-Davis, B, Thompson, NE, Burgess, RR, Edwards, AM, David, PR, Kornberg, RD. 2000. *Science* 288: 640
4. Da, LT, Wang, D, Huang, XH. 2012. *J Am Chem Soc* 134: 2399
5. Feig, M, Burton, ZF. 2010. *Proteins* 78: 434
6. Gnatt, AL, Cramer, P, Fu, JH, Bushnell, DA, Kornberg, RD. 2001. *Science* 292: 1876
7. Larson, MH, Zhou, J, Kaplan, CD, Palangat, M, Kornberg, RD, Landick, R, Block, SM. 2012. *P Natl Acad Sci USA* 109: 6555
8. Tan, L, Wiesler, S, Trzaska, D, Carney, HC, Weinzierl, RO. 2008. *Journal of biology* 7: 40
9. Yuzenkova, Y, Bochkareva, A, Tadigotla, VR, Roghanian, M, Zorov, S, Severinov, K, Zenkin, N. 2010. *Bmc Biol* 8
10. Steitz, TA. 1998. *Nature* 391: 231
11. Abashkin, YG, Erickson, JW, Burt, SK. 2001. *J Phys Chem B* 105: 287
12. Prasad, BR, Warshel, A. 2011. *Proteins-Structure Function and Bioinformatics* 79: 2900
13. Bojin, MD, Schlick, T. 2007. *J Phys Chem B* 111: 11244
14. Alberts, IL, Wang, Y, Schlick, T. 2007. *J Am Chem Soc* 129: 11100
15. Wang, LH, Yu, XY, Hu, P, Broyde, S, Zhang, YK. 2007. *J Am Chem Soc* 129: 4731
16. Wang, LH, Broyde, S, Zhang, YK. 2009. *Journal of molecular biology* 389: 787
17. Cisneros, GA, Perera, L, Garcia-Diaz, M, Bebenek, K, Kunkel, TA, Pedersen, LG. 2008. *DNA Repair* 7: 1824
18. Lin, P, Pedersen, LC, Batra, VK, Beard, WA, Wilson, SH, Pedersen, LG. 2006. *P Natl Acad Sci USA* 103: 13294
19. Florian, J, Goodman, MF, Warshel, A. 2003. *J Am Chem Soc* 125: 8163
20. Castro, C, Smidansky, ED, Arnold, JJ, Maksimchuk, KR, Moustafa, I, Uchida, A, Gotte, M, Konigsberg, W, Cameron, CE. 2009. *Nat Struct Mol Biol* 16: 212
21. Lior-Hoffmann, L, Wang, LH, Wang, SL, Geacintov, NE, Broyde, S, Zhang, YK. 2012. *Nucleic Acids Res* 40: 9193
22. Carvalho, ATP, Fernandes, PA, Ramos, MJ. 2011. *J Chem Theory Comput* 7: 1177
23. Rui, Z, Salahub, DR. 2007. *Aip Conf Proc* 963: 104
24. Zhu, R, Janetzko, F, Zhang, Y, van Duin, ACT, Goddard, WA, Salahub, DR. 2008. *Theor Chem Acc* 120: 479
25. Wang, D, Bushnell, DA, Westover, KD, Kaplan, CD, Kornberg, RD. 2006. *Cell* 127: 941
26. Foloppe, N, MacKerell, AD. 2000. *J Comput Chem* 21: 86
27. Lee, MS, Feig, M, Salsbury, FR, Brooks, CL. 2003. *J Comput Chem* 24: 1348
28. Brooks, BR, Bruccoleri, RE, Olafson, BD, States, DJ, Swaminathan, S, Karplus, M. 1983. *J Comput Chem* 4: 187
29. Feig, M, Burton, ZF. 2010. *Biophysical journal* 99: 2577

30. MacKerell, AD, Bashford, D, Bellott, M, Dunbrack, RL, Evanseck, JD, Field, MJ, Fischer, S, Gao, J, Guo, H, Ha, S, Joseph-McCarthy, D, Kuchnir, L, Kuczera, K, Lau, FTK, Mattos, C, Michnick, S, Ngo, T, Nguyen, DT, Prodhom, B, Reiher, WE, Roux, B, Schlenkrich, M, Smith, JC, Stote, R, Straub, J, Watanabe, M, Wiorkiewicz-Kuczera, J, Yin, D, Karplus, M. 1998. *J Phys Chem B* 102: 3586
31. Oelschlaeger, P, Klahn, M, Beard, WA, Wilson, SH, Warshel, A. 2007. *Journal of molecular biology* 366: 687
32. Phillips, JC, Braun, R, Wang, W, Gumbart, J, Tajkhorshid, E, Villa, E, Chipot, C, Skeel, RD, Kale, L, Schulten, K. 2005. *J Comput Chem* 26: 1781
33. Humphrey, W, Dalke, A, Schulten, K. 1996. *J Mol Graph Model* 14: 33
34. Nam, K, Cui, Q, Gao, JL, York, DM. 2007. *J Chem Theory Comput* 3: 486
35. Neese, F. 2012. *Wires Comput Mol Sci* 2: 73
36. Field, MJ. 2008. *J Chem Theory Comput* 4: 1151
37. Nam, KH, Gaot, JL, York, DM. 2008. *J Am Chem Soc* 130: 4680
38. Wong, KY, Lee, TS, York, DM. 2011. *J Chem Theory Comput* 7: 1
39. Radak, BK, Harris, ME, York, DM. 2013. *J Phys Chem B* 117: 94
40. Lev, B, Zhang, R, De la Lande, A, Salahub, D, Noskov, SY. 2010. *J Comput Chem* 31: 1015
41. Black, CB, Huang, HW, Cowan, JA. 1994. *Coordin Chem Rev* 135: 165
42. Garcia-Viloca, M, Poulsen, TD, Truhlar, DG, Gao, JL. 2004. *Protein Sci* 13: 2341
43. Ruiz-Pernia, JJ, Silla, E, Tunon, I. 2006. *J Phys Chem B* 110: 20686
44. Kumler, WD, Eiler, JJ. 1943. *J Am Chem Soc* 65: 2355
45. Venkatramani, R, Radhakrishnan, R. 2010. *Protein Sci* 19: 815
46. Mlynsky, V, Banas, P, Walter, NG, Sponer, J, Otyepka, M. 2011. *J Phys Chem B* 115: 13911
47. Roberts, ST, Petersen, PB, Ramashesha, K, Tokmakoff, A. 2009. *Springer Series Chem* 92: 481
48. Jin, JP, Dong, W, Guarino, LA. 1998. *J Virol* 72: 10011

CHAPTER SEVEN: CONCLUSIONS AND OUTLOOK

7.1 Conclusions

My thesis is comprised of both methodology implementation and application work towards simulating mRNA synthesis by RNA polymerase II. It first studies the kinetics of mRNA synthesis and identifies important steps for selecting the cognate NTP. It then focuses on the NTP binding process involving protein domain motions described at the MM level. Thermodynamics of this process are calculated and analyzed for a cognate NTP while the stability of cognate and non-cognate NTPs is also compared quantitatively. Before presenting results on the catalytic reaction, this thesis reviews the state-of-the-art QM/MM techniques and introduces an implementation of an interface between deMon2k and CHARMM. With an in-depth understanding of QM/MM methods gained from the review and the implementation, the nulceotidyl transfer reaction is simulated using QM/MM methodology. Overall, this work provides novel results and conclusions on mRNA synthesis by RNAP II on multiple scales, as outlined below. It elucidates the binding and catalytic processes in great detail by analyzing the surrounding protein residues. The energetic calculations provide insights into the selection mechanism and possible catalytic reaction pathways. In addition to the simulation results, the review and the implementation represent methodological contributions.

In Chapter 2, we build a kinetic model and successfully recover the rate of the nucleotide addition cycle based on empirical reaction parameters. We find that the selection for the matched base is achieved in the binding process while the discrimination of the 2'-OH should be fulfilled in the catalytic reaction.

In Chapter 3, we identify from MD trajectories the important residues in the NTP transfer and binding process. The free energy profile of the cognate NTP transfer from the entry site to

the addition site suggests that the trigger loop and the bridge helix actively participate, and that this process is not rate-limiting with the participation of the trigger loop and the bridge helix. The free energy difference between different types of NTPs demonstrates that the cognate NTP is the most stable NTP in the addition site. The 2'-dNTP is slightly less favored, by 1.91 kcal/mol, and the unmatched NTP is the least stable by 16.80 kcal/mol. The MD simulation results indicate that the instability of the 2'-dNTP is due to its twisted ribose, a direct result of the absent 2'-OH, which results in less interaction with surrounding residues. The instability of the unmatched NTP lies in the lack of contacts between its misplaced base and surrounding residues, stemming from mismatching between the base and the DNA template. The results of thermodynamic stability and kinetic transfer calculations suggest that the NTP is mainly discriminated in the addition site. In the case of unmatched NTPs, this is directly the result of thermodynamic instability. 2'-dNTPs, however, are likely discriminated through catalytic inefficiency.

Chapter 4 presents a comprehensive review of the QM/MM methodology as a backdrop for our own QM/MM implementation. We first introduce the QM/MM potential and illustrate various expressions of the effective QM/MM Hamiltonian and methods of partitioning the system. We then summarize geometry optimization and transition state search techniques with a QM/MM potential. Subsequently, special sampling methods for QM/MM are highlighted for free energy calculations such as free energy perturbation and umbrella sampling. Lastly, application of QM/MM methods to the study of DNA polymerases is reviewed to set a background for our study on RNAP II.

Chapter 5 presents an implementation of a QM/MM interface between deMon2k and CHARMM. In this work, the QM-MM coupling is on the QM level calculated by a one-electron operator in the QM Hamiltonian. The QM potential is based on DFT methods with auxiliary

DFT treatments available. The MM potential can be either a classical non-polarizable force field or a polarizable one such as the Drude model. Besides geometry optimization and MD simulation, this interface is also capable of performing free energy perturbation using the dual topology method. Free energy perturbation calculations on potassium and sodium ions in solution have demonstrated excellent performance of this interface.

In Chapter 6, we build three models to correct the defects in the RNAP II crystal structures and conduct MD simulations and QM/MM relaxed surface scans on each of them. Regarding the difference in the 3'O-Mg coordination among the three models, our results show that this coordination is not required for the reaction to proceed as it is evidently broken or weak in most of the scans that produce low energy barriers. Therefore, the role of Mg(A) in RNAP II appears to be more structural than catalytic. Regarding the nucleotidyl transfer reaction by RNAP II, the results show that the 3'-H is transferred to the α-phosphate either directly or indirectly, facilitating the formation of the 3'O-P_α bond and the weakening of the P_α-$O_{\alpha\beta}$ bond. Following this, the 3'-H migrates to the $O_{\alpha\beta}$, resulting in the pyrophosphate leaving. The quintessential part of this mechanism is that the proton so efficiently mediates among different parties engaged in the reaction to facilitate the P-O bond forming and breaking. Although the acceptor of the initial proton transfer may vary depending on the particular conformation of the active site, all possible routes converge to the same destination.

7.2 Outlook

Given my results, there are a large number of potential avenues for future research. Both methodology development and application work is needed to better understand the functions of RNAP II. A more in-depth understanding of this enzyme will enhance our general view of biological systems, and pave the way for a multitude of computational applications.

Regarding the MM study on the NTP transfer and binding, our MD trajectories from the interpolated structures for the NTP transfer produce an estimated pathway for this process. Further refinement of this pathway could be done using other enhanced sampling techniques such as targeted MD. On the other hand, for any of these techniques, an inevitable issue is that of choosing an effective reaction coordinate. The reaction coordinate must be characteristic of the reaction pathway and be capable of driving the system effectively. On this note, more specific reaction coordinates based on the interactions between the enzyme and the NTP should be explored. In case the reaction coordinate is not unique throughout the reaction, variation of the reaction coordinate for different stages might be necessary. Furthermore, this NTP transfer could also be simulated with an unmatched NTP and a 2'-deoxyNTP to check for consistency.

For the free energy perturbation calculations on the stability of NTPs, we found that the energy change of each window has converged. It might still be worthwhile to change the substrate reversely to double-check on the convergence. Lastly, for the MD of the cognate GTP in the addition site, a longer simulation could be performed to study the translocation of the RNAP II along the DNA strand. This can be done more efficiently with a large step size or more advanced techniques such as replica exchange.

Regarding the QM/MM study on the nucleotidyl transfer reaction, entropy effects need to be added to the calculated potential energies. Currently, I am running umbrella sampling calculations based on the geometries obtained from relaxed surface scans. A key factor in this calculation is the proper choice of the spring constant *k*, which should be varied for different windows to ensure adequate overlaps between neighboring windows. Another possible improvement is higher-level QM methods such as DFT to calculate the energy for the most favorable model. Since we have identified the best model for each proton transfer pathway using

AM1/d-PhoT with MM, DFT can now be employed to check on the energy. Lastly, a more rigorous transition state search could be conducted to refine the pathway obtained from relaxed surface scans.

With respect to methodology development, the greatest impediment for efficient QM/MM calculations is the speed of the QM method. Currently, with deMon2k using the PBE functional, DZVP basis set and the medium-sized grid, a single point calculation for a QM system of 124 atoms from RNAP II takes around 5 minutes with 64 processors. Each step of the relaxed surface scan could take thousands of steps to fully optimize the system. This QM computation speed problem may be mediated by more efficient algorithms or simply faster processors. In the case of semiempirical methods such as AM1/d-PhoT, better parametrization with more complete training sets would be necessary to improve the accuracy.

When considerable improvements on the accuracy of all above methods are achieved, the empirical rate constants in the kinetic model can be replaced by calculated ones. Thus, a true multi-scale approach can be realized. The weakest link currently is the calculated energies which could instigate orders-of-magnitude differences in the rate constant. However, I am hopeful that, with continuous efforts and improvements on the methodology, reliable theoretical prediction of energetics for real-life systems will be accomplished in the future.

Appendix 1: Copyright Permissions

By The Publishers

ELSEVIER LICENSE
TERMS AND CONDITIONS
Dec 12, 2013

This is a License Agreement between Rui Zhang ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Rui Zhang
Customer address	BI 547 2500 University Dr. NW Calgary, AB t2n1n4
License number	3285730472211
License date	Dec 11, 2013
Licensed content publisher	Elsevier
Licensed content publication	Elsevier Books
Licensed content title	Advances in Quantum Chemistry, Volume 59
Licensed content author	Rui Zhang,Bogdan Lev,Javier Eduardo Cuervo,Sergei Yu Noskov,Dennis R. Salahub
Licensed content date	2010
Number of pages	48
Start Page	353
End Page	400
Type of Use	reuse in a thesis/dissertation
Intended publisher of new work	other
Portion	full chapter
Format	electronic
Are you the author of this Elsevier chapter?	Yes
How many pages did you author in this Elsevier book?	50
Will you be translating?	No

Title of your thesis/dissertation	Multiscale simulation of mRNA synthesis by RNA polymerase II
Expected completion date	Dec 2013
Estimated size (number of pages)	
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.00 USD / 0.00 GBP
Total	0.00 USD

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier

Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and

conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.

16. **Website**: The following terms and conditions apply to electronic reserve and author websites:
Electronic reserve: If licensed material is to be posted to website, the web site is to be password-protected and made available only to bona fide students registered on a relevant course if:
This license was made in connection with a course,
This permission is granted for 1 year only. You may obtain a license for future website posting,
All content posted to the web site must maintain the copyright information line on the bottom of each image,
A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com , and
Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

17. **Author website** for journals with the following additional clauses:

All content posted to the web site must maintain the copyright information line on the bottom of each image, and the permission granted is limited to the personal version of your paper. You are not allowed to download and post the published electronic version of your article (whether PDF or HTML, proof or final version), nor may you scan the printed

edition to create an electronic version. A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx . As part of our normal production process, you will receive an e-mail notice when your article appears on Elsevier's online service ScienceDirect (www.sciencedirect.com). That e-mail will include the article's Digital Object Identifier (DOI). This number provides the electronic link to the published article and should be included in the posting of your personal version. We ask that you wait until you receive this e-mail and have the DOI to do any posting.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

18. **Author website** for books with the following additional clauses:
Authors are permitted to place a brief summary of their work online only.
A hyper-text must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

19. **Website** (regular and for author): A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx. or for books to the Elsevier homepage at http://www.elsevier.com

20. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

21. **Other Conditions**: Permission is granted to submit your article in electronic format. This license permits you to post this Elsevier article online on your Institution's website if the content is embedded within your thesis.

v1.6

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you

---

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS
Dec 12, 2013

This is a License Agreement between Rui Zhang ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3277970427817
License date	Nov 28, 2013
Licensed content publisher	John Wiley and Sons
Licensed content publication	Journal of Computational Chemistry
Licensed content title	The QM-MM interface for CHARMM-deMon
Licensed copyright line	Copyright © 2009 Wiley Periodicals, Inc.
Licensed content author	Bogdan Lev,Rui Zhang,Aurélien de la Lande,Dennis Salahub,Sergei Yu Noskov
Licensed content date	Dec 21, 2009
Start page	1015
End page	1023
Type of use	Dissertation/Thesis
Requestor type	Author of this Wiley article
Format	Electronic
Portion	Full article
Will you be translating?	No
Total	0.00 USD
Terms and Conditions	

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a "Wiley Company") or a society for whom a Wiley Company has exclusive publishing rights in relation to a particular journal (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that you opened your

RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

1. The materials you have requested permission to reproduce (the "Materials") are protected by copyright.

2.You are hereby granted a personal, non-exclusive, non-sublicensable, non-transferable, worldwide, limited license to reproduce the Materials for the purpose specified in the licensing process. This license is for a one-time use only with a maximum distribution equal to the number that you identified in the licensing process. Any form of republication granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before may be distributed thereafter). The Materials shall not be used in any other manner or for any other purpose. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Material. Any third party material is expressly excluded from this permission.

3. With respect to the Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Materials without the prior permission of the respective copyright owner. You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Materials, or any of the rights granted to you hereunder to any other person.

4. The Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc or one of its related companies (WILEY) or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto.

5. NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS,

IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY, INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

6. WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.

7. You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.

8. IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

9. Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.

10. The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.

11. This Agreement may not be assigned (including by operation of law or otherwise) by

you without WILEY's prior written consent.

12. Any fee required for this permission shall be non-refundable after thirty (30) days from receipt

13. These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.

14. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall prevail.

15. WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

16. This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

17. This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

Wiley Open Access Terms and Conditions

Wiley publishes Open Access articles in both its Wiley Open Access Journals program [http://www.wileyopenaccess.com/view/index.html] and as Online Open articles in its subscription journals. The majority of Wiley Open Access Journals have adopted the Creative Commons Attribution License (CC BY) which permits the unrestricted use, distribution, reproduction, adaptation and commercial exploitation of the article in any medium. No permission is required to use the article in this way provided that the article is properly cited and other license terms are observed. A small number of Wiley Open Access journals have retained the Creative Commons Attribution Non Commercial License (CC BY-NC), which permits use, distribution and reproduction in any medium, provided the

original work is properly cited and is not used for commercial purposes.

Online Open articles - Authors selecting Online Open are, unless particular exceptions apply, offered a choice of Creative Commons licenses. They may therefore select from the CC BY, the CC BY-NC and the Attribution-NoDerivatives (CC BY-NC-ND). The CC BY-NC-ND is more restrictive than the CC BY-NC as it does not permit adaptations or modifications without rights holder consent.

Wiley Open Access articles are protected by copyright and are posted to repositories and websites in accordance with the terms of the applicable Creative Commons license referenced on the article. At the time of deposit, Wiley Open Access articles include all changes made during peer review, copyediting, and publishing. Repositories and websites that host the article are responsible for incorporating any publisher-supplied amendments or retractions issued subsequently.
Wiley Open Access articles are also available without charge on Wiley's publishing platform, **Wiley Online Library** or any successor sites.

Conditions applicable to all Wiley Open Access articles:

- The authors' moral rights must not be compromised. These rights include the right of "paternity" (also known as "attribution" - the right for the author to be identified as such) and "integrity" (the right for the author not to have the work altered in such a way that the author's reputation or integrity may be damaged).

- Where content in the article is identified as belonging to a third party, it is the obligation of the user to ensure that any reuse complies with the copyright policies of the owner of that content.

- If article content is copied, downloaded or otherwise reused for research and other purposes as permitted, a link to the appropriate bibliographic citation (authors, journal, article title, volume, issue, page numbers, DOI and the link to the definitive published version on Wiley Online Library) should be maintained. Copyright notices and disclaimers must not be deleted.
    - Creative Commons licenses are copyright licenses and do not confer any other rights, including but not limited to trademark or patent rights.

- Any translations, for which a prior translation agreement with Wiley has not been agreed, must prominently display the statement: "This is an unofficial translation of an article that appeared in a Wiley publication. The publisher has not endorsed this translation."

Conditions applicable to non-commercial licenses (CC BY-NC and CC BY-NC-

ND)

For non-commercial and non-promotional purposes individual non-commercial users may access, download, copy, display and redistribute to colleagues Wiley Open Access articles. In addition, articles adopting the CC BY-NC may be adapted, translated, and text- and data-mined subject to the conditions above.

Use by commercial "for-profit" organizations

Use of non-commercial Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee. Commercial purposes include:

- Copying or downloading of articles, or linking to such articles for further redistribution, sale or licensing;
- Copying, downloading or posting by a site or service that incorporates advertising with such content;
- The inclusion or incorporation of article content in other works or services (other than normal quotations with an appropriate citation) that is then available for sale or licensing, for a fee (for example, a compilation produced for marketing purposes, inclusion in a sales pack)
- Use of article content (other than normal quotations with appropriate citation) by for-profit organizations for promotional purposes
- Linking to article content in e-mails redistributed for promotional, marketing or educational purposes;
- Use for the purposes of monetary reward by means of sale, resale, license, loan, transfer or other form of commercial exploitation such as marketing products
- Print reprints of Wiley Open Access articles can be purchased from: corporatesales@wiley.com

  The modification or adaptation for any purpose of an article referencing the CC BY-NC-ND License requires consent which can be requested from RightsLink@wiley.com .

  Other Terms and Conditions:

BY CLICKING ON THE "I AGREE..." BOX, YOU ACKNOWLEDGE THAT YOU HAVE READ AND FULLY UNDERSTAND EACH OF THE SECTIONS OF AND PROVISIONS SET FORTH IN THIS AGREEMENT AND THAT YOU ARE IN AGREEMENT WITH AND ARE WILLING TO ACCEPT ALL OF YOUR OBLIGATIONS AS SET FORTH IN THIS AGREEMENT.

v1.8

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK501170365.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

SPRINGER LICENSE
TERMS AND CONDITIONS
Dec 12, 2013

This is a License Agreement between Rui Zhang ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3277950914853
License date	Nov 28, 2013
Licensed content publisher	Springer
Licensed content publication	Interdisciplinary Sciences: Computational Life Sciences
Licensed content title	Exploring the molecular origin of the high selectivity of multisubunit RNA polymerases by stochastic kinetic models
Licensed content author	Rui Zhu
Licensed content date	Jan 1, 2009
Volume number	1
Issue number	2
Type of Use	Thesis/Dissertation
Portion	Full text
Number of copies	1
Author of this Springer article	Yes and you are the sole author of the new work
Order reference number	
Title of your thesis / dissertation	Multiscale simulation of mRNA synthesis by RNA polymerase II
Expected completion date	Dec 2013
Estimated size(pages)	220
Total	0.00 USD
Terms and Conditions	

Introduction

The publisher for this copyrighted material is Springer Science + Business Media. By clicking "accept" in connection with completing this licensing transaction, you agree that

the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

Limited License
With reference to your request to reprint in your thesis material on which Springer Science and Business Media control the copyright, permission is granted, free of charge, for the use indicated in your enquiry.

Licenses are for one-time use only with a maximum distribution equal to the number that you identified in the licensing process.

This License includes use in an electronic form, provided its password protected or on the university's intranet or repository, including UMI (according to the definition at the Sherpa website: http://www.sherpa.ac.uk/romeo/). For any other electronic use, please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com).

The material can only be used for the purpose of defending your thesis, and with a maximum of 100 extra copies in paper.

Although Springer holds copyright to the material and is entitled to negotiate on rights, this license is only valid, subject to a courtesy information to the author (address is given with the article/chapter) and provided it concerns original material which does not carry references to other sources (if material in question appears with credit to another source, authorization from that source is required as well).

Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

Altering/Modifying Material: Not Permitted
You may not alter or modify the material in any manner. Abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author(s) and/or Springer Science + Business Media. (Please contact Springer at (permissions.dordrecht@springer.com or permissions.heidelberg@springer.com)

Reservation of Rights
Springer Science + Business Media reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

Copyright Notice:Disclaimer
You must include the following copyright and permission notice in connection with any

reproduction of the licensed material: "Springer and the original publisher /journal title, volume, year of publication, page, chapter/article title, name(s) of author(s), figure number(s), original copyright notice) is given to the publication in which the material was originally published, by adding; with kind permission from Springer Science and Business Media"

Warranties: None

Example 1: Springer Science + Business Media makes no representations or warranties with respect to the licensed material.

Example 2: Springer Science + Business Media makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.

Indemnity
You hereby indemnify and agree to hold harmless Springer Science + Business Media and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

No Transfer of License
This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer Science + Business Media's written permission.

No Amendment Except in Writing
This license may not be amended except in a writing signed by both parties (or, in the case of Springer Science + Business Media, by CCC on Springer Science + Business Media's behalf).

Objection to Contrary Terms
Springer Science + Business Media hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and Springer Science + Business Media (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

Jurisdiction
All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in The Netherlands, in accordance with

Dutch law, and to be conducted under the Rules of the 'Netherlands Arbitrage Instituut' (Netherlands Institute of Arbitration).***OR:***

All disputes that may arise in connection with this present License, or the breach thereof, shall be settled exclusively by arbitration, to be held in the Federal Republic of Germany, in accordance with German law.

Other terms and conditions:

v1.3

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK501170352.
Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

Appendix 2: Copyright Permissions

By The Co-authors

Gmail by Google

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Bogdan Lev <lev.bogdan@gmail.com> Thu, Nov 28, 2013 at 12:52 AM
To: Rui Zhang <zhangrui1002@gmail.com>

Hi Rui,

It's fine with me, good luck on the defence.

Best regards,
Bogdan

On Thu, Nov 28, 2013 at 6:47 PM, Rui Zhang <zhangrui1002@gmail.com> wrote:

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Lev, B., **Zhang, R.**, De la Lande, A., Salahub, D., and Noskov, S.Y., *The QM-MM Interface for CHARMM-deMon.* Journal of Computational Chemistry, 2010. **31**(5): p. 1015-1023.

Thank you very much in advance!

Sincerely,

Rui Zhang

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Bogdan Lev <lev.bogdan@gmail.com> Thu, Nov 28, 2013 at 12:52 AM
To: Rui Zhang <zhangrui1002@gmail.com>

Hi Rui,

It's fine with me, good luck on the defence.

Best regards,
Bogdan

On Thu, Nov 28, 2013 at 6:46 PM, Rui Zhang <zhangrui1002@gmail.com> wrote:

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Zhang, R., Lev, B., Cuervo, J.E., Noskov, S.Y., and Salahub, D.R., *A Guide to QM/MM Methodology and Applications.* Advances in Quantum Chemistry, Vol 59, 2010. **59**: p. 353-400.

Thank you very much in advance!

Sincerely,

Rui Zhang

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Aurélien de la Lande <aurelien.de-la-lande@u-psud.fr> Thu, Nov 28, 2013 at 1:52 AM

To: Rui Zhang <zhangrui1002@gmail.com>

Dear Rui,

as co-author of the two papers you mention I give you my persimission to include them in you PhD thesis.
Very best wishes
Aurélien
Le 28/11/2013 08:44, Rui Zhang a écrit :

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Zhu, R., de la Lande, A., **Zhang, R.**, and Salahub, D.R., *Exploring the Molecular Origin of the High Selectivity of Multisubunit RNA Polymerases by Stochastic Kinetic Models.* Interdisciplinary Sciences-Computational Life Sciences, 2009. **1**(2): p. 91-98.

Thank you very much in advance!

Sincerely,

Rui Zhang

--
Aurélien de la Lande
Chargé de recherche CNRS
Laboratoire de Chimie Physique, Université Paris Sud
15, avenue Jean Perrin
91405 Orsay Cedex, FRANCE
Tel: +33 (0)1 69 15 73 98
Fax: +33 (0)1 69 15 61 88

Rui Zhang
<zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Dennis Salahub <dennis.salahub@ucalgary.ca> Thu, Nov 28, 2013 at 7:03 AM
To: Rui Zhang <zhangrui1002@gmail.com>
Cc: Dennis Salahub <dennis.salahub@ucalgary.ca>

Permission granted.

Dennis Salahub
--
Dennis R. Salahub Ph.D., FRSC, FAAAS
Professor of Chemistry
IQST – Inst. for Quantum Science and Technology
CMS – Centre for Molecular Simulation
ISEEE – Inst. for Sustainable Energy, Environment and Economy
BI 556
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada T2N 1N4
Tel: +403 220 3720
Fax: +403 210 8655
Email: dennis.salahub@ucalgary.ca

From: Rui Zhang <zhangrui1002@gmail.com>
Date: Thu, 28 Nov 2013 00:47:28 -0700
To: Bogdan <lev.bogdan@gmail.com>, "aurelien.de-la-lande@u-psud.fr" <aurelien.de-la-lande@u-psud.fr>, Dennis Salahub <dennis.salahub@ucalgary.ca>, Sergei Noskov <snoskov@ucalgary.ca>
Subject: permission to re-use our co-authored paper

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Lev, B., **Zhang, R.**, De la Lande, A., Salahub, D., and Noskov, S.Y., *The QM-MM Interface for CHARMM-deMon.* Journal of Computational Chemistry, 2010. **31**(5): p. 1015-1023.

Thank you very much in advance!

Sincerely,

Rui Zhang

Gmail by Google

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Dennis Salahub <dennis.salahub@ucalgary.ca> Thu, Nov 28, 2013 at 7:03 AM
To: Rui Zhang <zhangrui1002@gmail.com>
Cc: Dennis Salahub <dennis.salahub@ucalgary.ca>

Permission granted.

Dennis Salahub
--
Dennis R. Salahub Ph.D., FRSC, FAAAS
Professor of Chemistry
IQST – Inst. for Quantum Science and Technology
CMS – Centre for Molecular Simulation
ISEEE – Inst. for Sustainable Energy, Environment and Economy
BI 556
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada T2N 1N4
Tel: +403 220 3720
Fax: +403 210 8655
Email: dennis.salahub@ucalgary.ca

From: Rui Zhang <zhangrui1002@gmail.com>
Date: Thu, 28 Nov 2013 00:46:08 -0700
To: Bogdan <lev.bogdan@gmail.com>, Dennis Salahub <dennis.salahub@ucalgary.ca>, Sergei Noskov <snoskov@ucalgary.ca>, Javier Cuervo <javier.cuervo@gmail.com>
Subject: permission to re-use our co-authored paper

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Zhang, R., Lev, B., Cuervo, J.E., Noskov, S.Y., and Salahub, D.R., *A Guide to QM/MM Methodology and Applications.* Advances in Quantum Chemistry, Vol 59, 2010. **59**: p. 353-

400.

Thank you very much in advance!

Sincerely,

Rui Zhang

Gmail by Google

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Dennis Salahub <dennis.salahub@ucalgary.ca> Thu, Nov 28, 2013 at 7:03 AM
To: Rui Zhang <zhangrui1002@gmail.com>
Cc: Dennis Salahub <dennis.salahub@ucalgary.ca>

Permission granted.

Dennis Salahub
--
Dennis R. Salahub Ph.D., FRSC, FAAAS
Professor of Chemistry
IQST – Inst. for Quantum Science and Technology
CMS – Centre for Molecular Simulation
ISEEE – Inst. for Sustainable Energy, Environment and Economy
BI 556
University of Calgary
2500 University Drive NW
Calgary, Alberta, Canada T2N 1N4
Tel: +403 220 3720
Fax: +403 210 8655
Email: dennis.salahub@ucalgary.ca

From: Rui Zhang <zhangrui1002@gmail.com>
Date: Thu, 28 Nov 2013 00:44:11 -0700
To: "aurelien.de-la-lande@u-psud.fr" <aurelien.de-la-lande@u-psud.fr>, Dennis Salahub <dennis.salahub@ucalgary.ca>, Rui Bijuan Zhu <zhurhot@hotmail.com>
Subject: permission to re-use our co-authored paper

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Zhu, R., de la Lande, A., **Zhang, R.**, and Salahub, D.R., *Exploring the Molecular Origin of the High Selectivity of Multisubunit RNA Polymerases by Stochastic Kinetic Models.*

Interdisciplinary Sciences-Computational Life Sciences,2009. **1**(2): p. 91-98.

Thank you very much in advance!

Sincerely,

Rui Zhang

Gmail by Google

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Sergei Noskov <noskovsy@gmail.com> Thu, Nov 28, 2013 at 9:59 AM
To: Rui Zhang <zhangrui1002@gmail.com>
Cc: Bogdan <lev.bogdan@gmail.com>, Aurélien de la Lande <aurelien.de-la-lande@u-psud.fr>, Dennis Salahub <dennis.salahub@ucalgary.ca>, Sergei Noskov <snoskov@ucalgary.ca>

You have my permission to use it.

Sincerely

Sergei

On Wed, Nov 27, 2013 at 11:47 PM, Rui Zhang <zhangrui1002@gmail.com> wrote:

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Lev, B., **Zhang, R.**, De la Lande, A., Salahub, D., and Noskov, S.Y., *The QM-MM Interface for CHARMM-deMon.* Journal of Computational Chemistry, 2010. **31**(5): p. 1015-1023.

Thank you very much in advance!

Sincerely,

Rui Zhang

Gmail by Google

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Rui Zhu <zhurhot@hotmail.com> Thu, Nov 28, 2013 at 4:20 PM
To: Rui Zhang <zhangrui1002@gmail.com>, aurelien.de-la-lande@u-psud.fr, Dennis Salahub <dennis.salahub@ucalgary.ca>

Hi Rui,

I am glad you would include our work in your Ph.D. thesis.

Cheers,
Rui

From: Rui Zhang
Sent: Thursday, November 28, 2013 12:44 AM
To: aurelien.de-la-lande@u-psud.fr ; Dennis Salahub ; Rui Bijuan Zhu
Subject: permission to re-use our co-authored paper

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Zhu, R., de la Lande, A., **Zhang, R.**, and Salahub, D.R., *Exploring the Molecular Origin of the High Selectivity of Multisubunit RNA Polymerases by Stochastic Kinetic Models.* Interdisciplinary Sciences-Computational Life Sciences, 2009. **1**(2): p. 91-98.

Thank you very much in advance!

Sincerely,

Rui Zhang

Gmail by Google

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Javier Cuervo <javier.cuervo@gmail.com> Mon, Dec 2, 2013 at 1:51 PM
To: Rui Zhang <zhangrui1002@gmail.com>

Dear Rui,

Please take this email as my formal permission for you to use as part of your thesis the coauthored manuscript:

"**Zhang, R.**, Lev, B., Cuervo, J.E., Noskov, S.Y., and Salahub, D.R., *A Guide to QM/MM Methodology and Applications.* Advances in Quantum Chemistry, Vol 59, 2010. **59**: p. 353-400."

If you need anything else, please do let me know

Best Regards,

Javier Cuervo

On Thu, Nov 28, 2013 at 12:46 AM, Rui Zhang <zhangrui1002@gmail.com> wrote:

Dear all,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Zhang, R., Lev, B., Cuervo, J.E., Noskov, S.Y., and Salahub, D.R., *A Guide to QM/MM Methodology and Applications.* Advances in Quantum Chemistry, Vol 59, 2010. **59**: p. 353-400.

Thank you very much in advance!

Sincerely,

Rui Zhang

Gmail by Google

Rui Zhang <zhangrui1002@gmail.com>

permission to re-use our co-authored paper

Sergei Noskov <noskovsy@gmail.com> Thu, Dec 12, 2013 at 12:05 PM
To: Rui Zhang <zhangrui1002@gmail.com>

You can use it!

Sergei

Dec 12, 2013, в 12:00 PM, Rui Zhang <zhangrui1002@gmail.com> написал(а):

Dear Sergei,

I'd like to request your permission to include in my PhD thesis the following our co-authored manuscript:

Zhang, R., Lev, B., Cuervo, J.E., Noskov, S.Y., and Salahub, D.R., *A Guide to QM/MM Methodology and Applications.* Advances in Quantum Chemistry, Vol 59, 2010. **59**: p. 353-400.

Thank you very much in advance!

Sincerely,

Rui Zhang