Strategies in the Computational Modelling of Biological Systems: Case Studies with Radical Enzymes

Der Naturwissenschaftlichen Fakultät

der Friedrich-Alexander-Universität Erlangen-Nürnberg

zur

Erlangung des Doktorgrades Dr. rer. nat

Vorgelegt von

Karmen Čondić-Jurkić

aus Zagreb, Kroatien

Als Dissertation genehmigt von der Naturwissenschaftlichen Fakultät der Friedrich-Alexander Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung: *2. Juli 2013*

Vorsitzender des Promotionsorgans: *Prof. Dr. Johannes Barth*

Gutachter/in: *Prof. Dr. Ana-Sunčana Smith*

Prof. Dr. Tim Clark

ACKNOWLEDGEMENTS

In the beginning, I would like to express my gratitude to my supervisors, Dr. David M. Smith and Prof. Ana-Sunčana Smith, whose patient and competent guidance has led me to the end of this path. Thank you for your practical advices and tips, interesting discussions about science and life, all the provided opportunities, for your endless support and everything you have taught me and thank you for waving with that light at the end of the tunnel. Thank you, David, for finding the way to motivate me even in the pitch dark moments of the way and thank you, Ana, for your rare ability to combine a funky personality with a strong leadership. I respect and admire you both. I honestly think that blockbuster versions of Mr and Mrs Smith are mere shadows of you two – you are my supersupervisors. .

I am also grateful to the entire staff of the Cluster of Excellence: Engineering of Advanced Materials and Friedrichs Alexander University for giving me this opportunity to defend my thesis in Erlangen. I am especially grateful to Waltraud Meinecke for her kind help on many occasions.

Almost all of my calculations were done on the supercomputing facilities maintained by University Computing Centre (SRCE) in Zagreb and Regionalen Rechenzentrums Erlangen (RRZE). Running jobs on these supercomputers was smooth thanks to the excellent staff in both centres, especially Emir Imamagić (SRCE) and Thomas Zeiser (RRZE).

Having a pleasant working environment is very important aspect of every job, including PhD training and I had the opportunity to switch between different, but always welcoming and friendly places. My colleagues from the Group of quantum organic chemistry at Ruđer Bošković Institute in Zagreb have always provided a pleasent working environment, and for this I am particulary grateful to Danijela, Boris and Robert for tolerating my presence and absence in the shared office, which was in roughly equal amounts. My days at RBI were more interesting and amusing thanks to many serious and not so serious discussions over coffee by the machine with Nives, Fran and Matko. The list for Matko is a bit longer and here it goes - thank you for: Nutella and the spoon; for a batch of the episodes of the stupid show in the middle of the night; for all the bike rides, cups of lousy coffee, Skype chats, true friendship and the kindest of hearts. The insightful and often cynical remarks made by Boris Zimmerman in

our random or deliberate encounters at RBI were always something refreshing and useful. Finally, I have to thank to Pavle Trošelj for acting as my pillar for so many years. I could not have asked for a better supporting system.

I owe much of my gratitude to Prof. Hendrik Zipse and his group at te Ludwig Maximillians University in Munich. Working with the entire group was a great experience and I appreciate all the given opportunities. Often being the only lady in the group, I am thankful to the boys for their warm welcome and hospitality during my every visit. That misfortunate football match and the broken arm are forgotten and forgiven.

It is impossible not to mention my Erlangen group and must thank all of them for making me feel like home. Zoran, thanks for all the chocolate, creating disk space, papers, and most of all, thank you for taking care of our golden boy! Jayant, you are living proof that an embodiment of the incessant gloominess can be a source of laughter with all your witty and funny remarks. Divine Zlatko, there are no words for you, just prayers and a secret admiration...

Sara, my dearest bitch, you get a separate line because from all the people in the world, you know me the best at this moment, despite the distance. Thank you for being everything you are and for being „my person“.

All these scientfic years would be much more boring experience without entire Bioinformatic group at Faculty of Science and their ability to throw some really good parties and barbecues. The conferences would not be the same without you, and I will never forget (or remember) Vedran's birthdays, Tina's creative interventions in Photoshop and all hilarious Facebook communication.

Thanks to my girlfriends Mirna, Branka, Olga, and Anita, I even had a social life during this long journey and I am forever grateful to have you in my life. Without you, my world would be a boring and lonely place. The excitement and the adventures in the recent years I owe almost exclusively to my climbing partner and a good friend, Rene, who took me there and back again. I would like to thank him for always bringing me back; for patient ear to my nagging; for taking the leading climb when I was too scared and for keeping me safe when I wasn't.

Finally, I have to acknowledge my entire family for bearing with me in the moments when I was unbearable, or to be more precise, for being there for me in any given moment. I must mention my dearest cousins – Katarina, Ana, Antica and Josipa for ensuring that every family reunion is so much fun since we were little girls. I must admit that I would be often lost in

everyday life, if not for my explosive, yelling, loud-laughing, but reliable and simply the best sister Iva. I really do not know what I would do without you. Dad, I will always stay your „malo zlo“. And Mom, I do not exaggarate when I say that you are a little miracle. I love you all.

ZUSAMMENFASSUNG

Fortschritte in der Theorie und Entwicklung der Informationstechnologie haben in den letzten Jahrzehnten gemeinsam dazu beigetragen, dass computergestützte Ansätze in der Molekülmodellierung in den Vordergrund der Wissenschaftsforschung gekommen sind. Die moderne Molekümodellierung kann als ein entscheidendes Instrument in der Erforschung verschiedener Prozesse, die sich auf atomarer Ebene abspielen, angesehen werden – von einfachen chemischen Reaktionen bis zur Assoziation großer Moleküle zu komplexen Strukturen. Als Beispiel für letztere repräsentieren biologische Systeme eine besondere Herausforderung für die Modellierung, und zwar wegen ihrer Komplexität und den komplizierten Verbindungen zwischen den Quantenereignissen und dem makroskopischen Verhalten lebender Systeme. Chemische Reaktionen stellen auf der anderen Seite elementare Schritte unzähliger Zellenprozesse dar, in denen Enzyme, als biologische Katalysatoren, eine wichtige Rolle spielen.

Die in dieser Doktorarbeit durchgeführte Forschung beruht auf der Anwendung moderner computergestützten Techniken für die Modellierung der Enzymkatalyse. Um chemische Reaktionen und die begleitenden Veränderungen in der Elektronendichte angemessen beschreiben zu können, ist es notwendig die Methode anzuwenden, die sich auf Quantenmechanik (QM) stützt. Die Anwendbarkeit der QM – Methoden ist auf kleine Systeme, die sich aus bis zu einpaar hundert Atomen zusammensetzen, beschränkt. Deswegen ist für die Modellierung großer Molekülsysteme, wie z.B. Proteine, ein alternativer Ansatz, der sich typischerweise auf klassischer Mechanik basiert, vonnöten. Dieser Ansatz ist als Molekülmechanik (MM) bekannt und die potentielle Energie des Systems ist nur dann als Funktion der Kernkkoordinaten geben, wenn die elektronischen Freiheitsgrade in die potentielle Energie implizit miteingeschlossen sind. Wegen dieser Approximation ist die MM - Technik nicht für die Modellierung chemischer und/oder elektronischer Prozesse, die sich in einem Enzym ereignen, angemessen. Dessen ungeachtet gibt sie aber einen Einblick in die Dynamik großer Systeme.

Verschiedene Strategien wurden entwickelt, um die enzymatischen Reaktionen zu beschreiben. Einer von den breit akzeptierten Ansätze verwendet kleine Molekülmodelle, welche das aktive Zentrum der Enzyme repräsentieren und die mit den quantenmechanischen Berechnungen behandelt werden können. Ein besser ausgearbeiteter Ansatz, der die Protein-Umgebung berücksichtigt, ist die QM/MM – Technik, die die Präzision der QM – Methode mit

der Anwendbarkeit des MM – Ansatzes kombiniert. Dies wird durch das Identifizieren des entscheidenden Enzymteils, der in die Katalyse involviert ist, und das Behandeln mit der QM-Methode erreicht, während der Rest des Systems mit der weniger kostspieligen MM - Methode beschrieben wird.

Im Fokus dieser Doktorarbeit war das Erforschen auserwählter Radikal-Enzyme, und zwar durch eine bedachte Kombination der oben genannten computergestützten Ansätze. Diese Katalysatoren benutzen die freien radikalen Arten auf eine spezifische Art und Weise – sie enthalten ungepaarte Elektronen, mithilfe der sie bestimmte chemische Reaktionen in den Zellen beschleunigen. Wegen ihrer hohen Reaktivität, besonders in Kombination mit Sauerstoff, werden die freien radikalen Arten gewöhnlich nicht in den Katalyselabors benutzt. Radikal-Enzyme können auch ziemlich empfindlich auf die Anwesendheit von Sauerstoff reagieren, obwohl sie eine bedeutendere Rolle in frühen Lebensformen hatten, als Sauerstoff mangelnd war. Ungeachtet der aeroben oder anaeroben Natur der Umwelt involviert die biologische radikale Katalyse üblicherweise ein metallisches Ion oder einen anderen Kofaktor als radikalen Initiator. Im Falle der (6-4) Photolyase, eines der in dieser Doktorarbeit studierten Enzyme, ist der Kofaktor das Flavinadenin-Dinukleotid (FAD). Die (6-4) Photolyase ist ein lichtabhängiges Enzym, welches die Fähigkeit besitzt, die Verletzungen am DNA zu reparieren, und zwar mittels eines radikalen Mechanismus mit der Beteiligung zweier katalytischen Histidinen. Der Protonierungszustand dieser beiden Histidinen spielt eine wichtige Rolle im Reparatur-Mechanismus, wobei dieses Problem mittels einer Multi-Skalen-Molekülmodellierung festgestellt wurde. Dieser Ansatz schloss die QM/MM – Berechnungen der spektroskopischen Parameter, pK_a Berechnungen mittels impliziter Lösungsmittel-Modelle und Simulationen der molekularen Dynamik mit ein, um die strukturellen Auswirkungen auf verschiedene Protonierungszustände zu untersuchen. Die Kombination aller Resultate gab einen koherenten Einblick in die wahrscheinlichen Protonierungszustände der zwei Histidinreste.

Zusätzlich zu den Radikal-Enzymen mit einer spezifischen Kofaktor-Abhängigkeit, wie bei der (6-4) Photolyase, gibt es noch unzählige Enzyme, die ein aktives Zentrum auf einem Proteinrückstand tragen. Ein wichtiges Beispiel solch einer katalytischen Maschinerie ist die Pyruvat-Formiat-Lyase (PFL), die zur Klasse der Glycyl-Radikal-Enzyme gehört und in der das aktive Zentrum durch ein zusätzliches Enzym zur Aktivierung repräsentiert wird. PFL katalysiert einen wichtigen Schritt im Glukose-Metabolismus unter anaerobischen Verhältnissen in vielen Mikroorganismen, wo das Pyruvat und Coenzym A (CoA) zu Formiat und Acetyl-CoA konvertiert wurden. Die Katalyse verfährt in zwei Schritten. Im ersten Schritt

wird das Pyruvat zu Formiat und einer Acetyl-Gruppe gespalten, folgend einem Angriff des radikalen Cysteinrestes, welches dann als vorübergehender Acetyl-Träger fungiert. Der zweite Schritt involviert den Transfer der Acetyl-Gruppe zum CoA. In dieser Doktorarbeit wurde die erste Halbreaktion zusammen mit dem dazugehörigen Inhibitor des PFL vom Oxamat, mit kleinen Modellen, die sehr genauen QM – Berechnungen unterstanden, analysiert. Es wurden genauso alternative Reaktionswege für Pyruvat und Oxamat erforscht, die auf Wasserstof-Abstraktion anstelle der radikalischen Addition beruhten. Abschließend wurden die Resultate dieser computergestützten Studie benutzt, um die QM/MM – Ansätze im Gegensatz zu den reinen QM – Resultaten zu bewerten.

Die Studie der zweiten Halbreaktion, die die Acetylierung des CoA involviert, setzt ihren Fokus auf mögliche Wege, mittels der das Ko-Substrat (CoA) das aktive Zentrum erreichen kann. Die Notwendigkeit solche Wege zu finden kommt daher, dass die Verbindungsseite des CoA sich auf der Proteinoberfläche befindet, und zwar 30 Å weit vom aktiven Zentrum entfernt. Deswegen sollten sich in der zweiten Halbreaktion bestimmte konformative Veränderungen, die in der Lage sind das CoA zum aktiven Zentrum zu bringen, ereignen. In dem Bestreben die Natur der erforderlichen strukturellen Neuanordnungen zu ermitteln, wurde eine Serie von Simulationen der Moleküldynamik ausgetragen. Ein wichtiger Aspekt dieser Studie war das Bewerten der Veränderungen der freien Energie, die den Ansatz des CoA zum aktiven Zentrum begleiteten. Um bestimmen zu können, ob die chemischen Veränderungen, zu denen es in der ersten Halbreaktion gekommen ist, den CoA Ansatz beeinflussen, wurden Simulationen von Systemen, die das verbindende Arrangement reflektieren, vor und nach dem initialen Acetyl-Transfer ausgetragen.

Zusammenfassend wurde eine Reihe von computergestützten Techniken auf Radikal-Enzyme angewandt. Die spezifischen Ansätze, die in dieser Doktorarbeit verwendet wurden, reichen von Rechnungen der spektroskopischen Parameter und pK_a – Werte bis zur Konstruktion der Potentialenergieflächen für chemische Reaktionen und Veränderungen der freien Energie mit strukturellen Neuanordnungen. Mit solch einem facettenreichen Ansatz war es möglich eine Serie von ungelösten mechanistischen Problemen, die relevant für die Systeme von Interesse sind, anzusprechen. Der Gebrauch vielfältiger Techniken ermöglichte im Besonderen die Untersuchung der Vernetzung verschiedener Prozesse, die sich in verschiedenen Zeit- und Längenskalen abspielen. Soweit sich die molekulare Modellierung auch weiterhin entwickelt, kann man davon ausgehen, dass die Anzahl solcher Anwendungen, in denen man Techniken kombiniert, um Antworten auf Fragen von biologischer Bedeutung zu beantworten, steigen wird.

ABSTRACT

Theoretical advances and the development of information technology over the last decades have combined to place computational approaches to molecular modelling at the forefront of scientific research. As such, modern molecular modelling can be considered to be a powerful tool in the investigations of various processes taking place at the atomic level - from simple chemical reactions to the association of large molecules into complex structures. As an example of the latter, biological systems represent a special challenge for modelling due to their complexity and the intricate liaisons between quantum events and the macroscopic behaviour of living systems. Chemical reactions, on the other hand, are the elementary steps of numerous cellular processes, in which enzymes, as biological catalysts, play a crucial role.

The research conducted in this thesis is based on the application of modern computational techniques to model enzyme catalysis. To properly describe chemical reactions and the accompanying changes in the electron densities, it is necessary to employ methods based on quantum mechanics (QM). The applicability of the QM methods is limited to small systems that contain up to a few hundreds of atoms. Hence, the modelling of large molecular systems, such as proteins, requires an alternative approach, typically based on classical mechanics. This approach is known as molecular mechanics (MM) and the potential energy of the system is given as a function of the nuclear coordinates only, while the electronic degrees of freedom are included in the potential energy implicitly. Due to these approximations, the MM technique is not an appropriate tool for modelling the chemical and/or electronic processes taking place in an enzyme, but it does provide an insight into dynamics of large systems.

Different strategies have been developed to describe enzymatic reactions and one of the widely accepted approaches employs small molecular models representing the active site subjected to the QM calculations. A more elaborate approach that takes into account the protein environment is the QM/MM technique that combines the accuracy of QM methods with the practicality of the MM approach. This is achieved by identifying the crucial part of the enzyme that is involved in catalysis and treating it with QM methods, while the rest of the system is described with the less expensive MM method.

Using a judicious combination of the aforementioned computational approaches, the applicative focus of this thesis was the investigation of selected radical enzymes. These catalysts are specific in their use of radical species, containing unpaired electrons, to accelerate

certain chemical reactions in the cells. Due to their high reactivity, especially with oxygen, radical species are not commonly employed in laboratory catalysis. Radical enzymes can also be quite sensitive to the presence of oxygen, although they might have had a more prominent role in early life forms, when oxygen was scarce. Irrespective of the aerobic or anaerobic nature of the environment, biological radical catalysis usually involves a metal ion or some other cofactor as a radical initiator. In the case of (6-4) photolyase, one of the enzymes studied in this thesis, the cofactor is flavin adenine dinucleotide (FAD). (6-4) photolyase is a light-dependent enzyme capable of repairing DNA lesions via a radical mechanism with the participation of two catalytic histidines. The protonation states of these two histidines play an important role in the repair mechanism and this issue was assessed by a multiscale molecular modelling approach. This approach included QM/MM calculations of spectroscopic parameters, pK_a calculations using implicit solvation models and molecular dynamics simulations to examine the structural impact of different protonation states. The combination of all of the results provided a coherent insight into the likely protonation states of the two histidine residues.

In addition to the radical enzymes with a specific cofactor dependence, like (6-4) photolyase, there are numerous enzymes that carry a radical centre on a protein residue. An important example of this kind of catalytic machinery is pyruvate formate-lyase (PFL), which belongs to a class known as glycyl radical enzymes and in which the radical site in introduced by means of an additional activation enzyme. PFL catalyzes an important step in glucose metabolism under anaerobic conditions in many microorganisms, where pyruvate and coenzyme A (CoA) are converted to formate and acetyl-CoA. The catalysis proceeds in two steps. In the first step, pyruvate is cleaved into formate and an acetyl group following the attack of radical cysteine residue, which then acts as a temporary acetyl carrier. The second step involves transfer of the acetyl moiety to CoA. In this thesis, the first half-reaction, including the related inhibition of PFL by oxamate, was analyzed with small models subjected to very accurate QM calculations. Alternative reaction pathways for both pyruvate and oxamate, involving hydrogen abstraction instead of radical addition, were also investigated. Finally, the results from this computational study were used to validate QM/MM approaches against the pure QM results.

The study of the second half-reaction, which involves the acetylation of CoA, is focused on the possible pathways by which the co-substrate, (CoA) can reach the active site. The need to establish such pathways stems from the fact that the binding site of CoA is located at the protein surface, some 30 Å away from the active site. Thus, in order to complete the second

half-reaction, certain conformational changes, capable of bringing CoA to the active site, should take place. In an effort to identify the nature of the requisite structural rearrangements, a series of molecular dynamics simulations was carried out. An important aspect of this study was the evaluation of the free-energy changes accompanying the approach of CoA to the active site. To determine if the chemical changes that take place in the first half reaction influence the CoA approach, simulations were carried out on systems reflecting the bonding arrangement both before and after the initial acetyl transfer.

In summary, a variety of computational techniques has been applied to selected radical enzymes. The specific approaches that were used in the current thesis range from the calculation of spectroscopic parameters and pK_a values to the construction of potential energy surfaces for chemical reactions and free energy changes associated with structural rearrangements. With such a multifaceted approach, it was possible to address a series of unresolved mechanistic issues relevant to the systems of interest. In particular, the use of a range of techniques enabled the investigation of the interconnectedness of processes taking place at different time and length scales. As molecular modelling continues to develop as an independent discipline, one can expect an increased number of similar applications, in which targeted techniques are combined to provide answers to questions of biological relevance.

CONTENTS

1. Computational Methods

1.1 Introduction

Over the years, many experimental and theoretical tools have been developed to gain better understanding of natural phenomena at a molecular level and, finally, to apply the acquired knowledge in the improvement of existing and the development of new technologies. One of the crucial moments for in science was the development of quantum mechanics (QM) during the last century, providing an excellent theoretical framework for studying the world of subatomic particles. Theoretically, any atomic or molecular property can be predicted by quantum mechanics, but in real life, the exact solutions exist only for systems with one electron. To get around this problem, different methods were developed for many-electron systems giving approximate answers, most of them being very computationally demanding. The development of computer science enabled modelling of molecular systems and their properties by combining theoretical principles of quantum chemistry with the rapidly growing computing power, giving rise to a new discipline, known as computational chemistry. Computational chemistry had a major uptake in the last few decades, primarily due to the appearance of the high performance computing facilities and constant improvements in methods and algorithms in commonly used software packages.

Nevertheless, only a very small fraction of systems can be treated with QM based methods, with limitations mostly posed by the size of the system. To model large systems, such as biomacromolecules, even more approximate methods are necessary, in which molecular interactions are not described anymore by quantum mechanics, but rather within a classical formalism. This concept is known as *molecular mechanics* (MM), where the potential energy of the system is defined with the so-called *force field*, a set of parameters describing the interacting particles, but with no ability to reproduce electronic processes, such as formation and cleavage of chemical bonds. Having forces defined with the given force field, propagation of the system in time is achieved by numerically solving Newton's equations, leading to *molecular dynamics* (MD) simulations. From MD simulations, information about conformational changes and fluctuations of the system can be retrieved.

This chapter is an overview of the different computational methods used in this thesis, with a glimpse at the basic theoretical background. The methods are grouped by the underlying

theory (QM, MM) and the calculated properties (enthalpy or free energy). More details about listed methods and theory can be found in standard textbooks in quantum chemistry and molecular modelling.[1-5]

1.2 Quantum Mechanics Methods

To describe the electron motion in atoms and molecules and the rearrangement of electron density during chemical reactions, it is necessary to employ quantum mechanics. Methods based on the equations derived from quantum mechanical laws are known as *ab initio* methods, i.e. all molecular properties can be calculated from the first principles. These principles will be introduced in this section, together with the selected computational techniques developed from them.

1.2.1 Schrödinger Equation

Two equivalent mathematical formulations of quantum mechanics were independently developed by W. Heisenberg and E. Schrödinger, but most of the computational chemistry methods today rely upon Schrödinger's wave function. The time-dependent Schrödinger's equation corresponds to Newton's second law in classical mechanics:

$$i\hbar \frac{\partial \Psi(q,t)}{\partial t} = \boldsymbol{H}\Psi(q,t) \tag{1}$$

where the symbol $\boldsymbol{H}$ stands for Hamiltonian operator, which acts upon a wave function $\Psi(q,t)$, dependent on space (q) and time (t) coordinates. The wave function $\Psi(q,t)$contains all the information about the state of the system, as does its complex conjugate $\Psi^*(q,t)$. The wave function is a complex function in general and has no physical meaning by itself. According to the Born interpretation, the product $\Psi^*(q,t)\Psi(q,t)\mathrm{d}q$ is a real number corresponding to the probability of finding a particle at time t.

If $\boldsymbol{H}$ is independent of time, the wave function can be written as a product of two separate functions $\Psi(q,t)=\psi(q)\phi(t)$, where $\psi(q)$ depends only on the spatial coordinates and $\phi(t)$ depends only on time. In that case, we can write time-independent Schrödinger equation:

$$\boldsymbol{H}\psi(q) = \mathrm{E}\psi(q) \tag{2}$$

$$i\hbar \frac{\mathrm{d}\varphi(t)}{dt} = \mathrm{E}\varphi(t) \tag{3}$$

The general definition of the Hamiltonian operator for charged particles is given by the following expression:

$$\boldsymbol{H} = -\sum_{i}^{\text{particles}} \frac{1}{2m_i}\nabla_i^2 + \sum_{i<j}^{\text{particles}} \sum \frac{q_i q_j}{r_{ij}} \tag{4}$$

$$\nabla_i^2 = \frac{\partial^2}{\partial x_i^2} + \frac{\partial^2}{\partial y_i^2} + \frac{\partial^2}{\partial z_i^2} \tag{5}$$

where ∇_i^2 is Laplace operator acting upon particle *i* with the mass m_i and charge q_i, and r_{ij} is the distance between the particles. The first term in the expression describes the kinetic energy of the particle within the wave mechanics formulation, while the second term is the result of coulombic interactions between the particles. This is time-independent non-relativistic formulation of the Schrödinger equation (additional interactions can be included in the expression) and it cannot be solved analytically for most of the systems, exceptions being the one-electron systems. However, through the use of several rigorous mathematical simplifications, approximate solutions can be obtained for a wide range of chemical problems.

One of the most important simplifications is the separation of the nuclear and electronic contributions, known as the Born-Oppenheimer approximation. This approximation relies on the fact that nuclei in atoms are considerably heavier and slower than the electrons and the electrons can be considered to be moving in a field of fixed nuclei. The overall wave function ψ_{tot} describing the state of the molecules can be written in the following way:

$$\psi_{tot}(q_{el}, q_N) = \psi_{el}(q_{el}, q_N)\psi_N(q_N) \tag{6}$$

$$(H_{el} + V_N)\psi_{el} = U\psi_{el} \tag{7}$$

The electronic Hamiltonian H_{el} does not contain nuclear kinetic energy contribution, and nuclear potential energy remains constant for any particular configuration of the nuclei.

The electronic Schrödinger equation is an eigenvalue equation, whose solutions are eigenfunctions $\psi(q)$ and they correspond to the different stationary states of the system, each having its own eigenvalue E (energy). The solution with the lowest eigenvalue (E_0) is defined as the electronic ground-state ψ_0, and it can be approximated as a product of one-electron molecular orbitals $\phi_\mu(q)$, i.e. functions describing spatial coordinates of a single electron. For correct description, it is necessary to include dependence on the spin coordinates, described by spin functions, $\alpha(s)$ and $\beta(s)$. The complete wave function of a single electron is then given by

a product of the molecular orbital and a spin function, resulting with a function known as spin orbital $\chi(q,s)$.

These one-electron spin orbitals are then used to construct a wave function describing the *n*-electron system and its simplest form would be a product of *n* spin orbitals. Unfortunately, this product is not antisymmetric, which is a very important property that every wave function describing electrons (or fermions, in general) should possess. Antisymmetry requires that the total wave function change sign with respect to the exchange of the particles. To satisfy this condition, it is common to use Slater determinants for construction of the wave function:

$$\psi(q) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_n(1) \\ \chi_1(2) & \chi_2(2) & \cdots & \chi_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(n) & \chi_2(n) & \cdots & \chi_n(n) \end{vmatrix} \tag{8}$$

Each row in the determinant represents one of the electrons in every one of the *n* possible configurations. Interchanging two of the electrons is equivalent to interchanging two of the rows in the determinant. This interchange alters the sign of the overall wave function, and thus ensures that the antisymmetry condition is satisfied. The Slater determinant also satisfies the Pauli exclusion principle, in that it is not possible for two electrons to occupy the same molecular orbital, while having the same spin (in determinant, two identical columns vanish!). The factor $(n!)^{-1/2}$ is to ensure normalization.

In practice, molecular orbitals are expressed as a linear combinations of a finite set of one-electron functions:

$$\varphi_i(q) = \sum_{\mu} c_{\mu i} \rho_{\mu} \tag{9}$$

In the general mathematical treatment of the problem, any set of appropriately defined basis functions may be used. It is most convenient for each atom in the molecule to be the centre for a set of basis functions. When the basis functions are taken to be the atomic orbitals of the constituent atoms, Eq. 9 is known as a linear combination of atomic orbitals (LCAO).

With all the introduced approximations, solving of the Schrödinger equation has been reduced to determining molecular-orbital expansion of coefficients $c_{\mu i}$. These coefficients will completely specify the ground-state electronic wave function and to find the most appropriate ones, it is necessary to employ the variational principle. According to the variational principle, any energy obtained with the approximated wave function, like above, will be greater than the

energy calculated from the exact solution of the Schrödinger equation. The variational method sets an upper limit for the exact energy and the best possible wave function is found by minimizing the energy with respect to the molecular orbital expansion coefficients $c_{\mu i}$.

1.2.2 HARTREE-FOCK THEORY

Hartree-Fock (HF) theory[6,7] is the basis for many modern *ab initio* calculation methods, known as *post-Hatree-Fock* methods. One of the fundamental approximations of HF procedure is the assumption that motion of an electron does not depend explicitly on the instantaneous motion of the other electron, rather it feels the Coulomb repulsion due to average position of all electrons (*mean field* theory). In the frames of the Hartree-Fock theory, the many-electron Schrödinger equation can be written in a form of many one-electron HF equations:

$$\boldsymbol{F}_i \varphi_i = \varepsilon_i \varphi_i \tag{10}$$

where $\boldsymbol{F}_i$ is Fock operator, and ε_i is the energy of the one-electron molecular orbital φ_i .In order to solve Hartree-Fock equations, it is necessary to make an additional approximation by expanding φ_i with a set of one-electron functions (atomic orbitals):

$$\varphi_i(q) = \sum_{\mu} c_{\mu i} \rho_{\mu} \tag{11}$$

Finally, the Roothan-Hall[8,9] equations are employed to obtain coefficients $c_{\mu i}$ and orbital energies ε_i :

$$\sum_{\mu=1}^{N} c_{\mu i} \left(F_{\mu\nu} - \varepsilon_i S_{\mu\nu} \right) = 0 \tag{12}$$

In a system with N basis functions, $F_{\mu\nu}$ are the elements of an N by N matrix called the Fock matrix, $S_{\mu\nu}$ are the elements of another N by N matrix known as the overlap matrix and ε_i is the one-electron energy of the molecular orbital φ_i . The equation above can be written in the matrix notation, that is, as a generalized eigenvalue problem (**C** is the N-order matrix consisting of elements $c_{\mu i}$, **ε** is the diagonal matrix with elements ε_i):

$$\mathbf{FC} = \mathbf{SC}\,\boldsymbol{\varepsilon} \tag{13}$$

These equations are being solved in the iterative fashion, due to dependence of Fock operator and its matrix elements on the occupied molecular orbitals (MO) φ_i, which depend on the unknown coefficients $c_{\mu i}$. An initial guess to the orbitals is made and each electron is treated as being described by a potential energy due to the nucleus, shielded by some average field of all the other electrons. Each iteration corresponds to a better guess of all the orbitals and the process ceases when no change in the $c_{\mu i}$ is observed from one cycle to the next. Since the resulting orbitals are derived from their own effective potential, the procedure is commonly referred to as the *self consistent field* (SCF) technique. This method is variational and its resulting energies are always higher than the exact values. In addition, HF method has properties of *size-consistency* and *size-extensivity*.

1.2.2.1 Open Shell Systems

Construction of the molecular orbitals by following HF procedure will depend on the number of α and β electrons in the system. Namely, an additional constraint is introduced in Roothan-Hall equations for the *closed-shell* systems (α = β), that enforces two electrons of the opposite spin into one spatial orbital. This approach results with so called *restricted Hartree-Fock* wave function (RHF).

For the molecules with the unpaired electrons, i.e. *open-shell* systems, there are two available techniques for building the HF wave function. The *restricted open-shell Hartree-Fock* (ROHF) includes similar restraint as in RHF, where the electron pairs are required to occupy the same spatial orbital. This ensures that the wave function is a pure spin state and thus an eigenfunction of the spin squared ($\boldsymbol{S^2}$) operator, but because any spin polarization of the electrons (in doubly-occupied orbitals) can only be in the direction of the unpaired electrons, ROHF spin densities are often incorrect.

The other approach for the open-shell systems employs two separate sets of Roothan-Hall equations for α and β electrons, known as the Pople-Nesbet equations.[10] This technique is gives *unrestricted Hartree-Fock* (UHF) wave function, where the paired electrons do not share the same spatial distribution. The UHF solution will collapse to the RHF wave function for majority of the closed-shell systems, but for the systems with the unpaired electrons, UHF energies will always be lower than alternative ROHF, due to greater flexibility of the wave function. Spin polarization within UHF framework can occur in both directions.

Unfortunately, the UHF wave function is not an eigenfunction of $\boldsymbol{S}^2$ anymore, meaning that it can contain some spurious contributions of the higher-lying spin states. This phenomenon is known as spin contamination and severity with which it can affect the final result depends highly on the observed system. In general, the UHF wave functions provide a better description for the spin polarization and dissociation of the chemical bonds, but for singlet spin molecules, it is common to use RHF to speed up the calculations.

1.2.3 Basis Sets

A basis set is a set of functions used for description of the orbital shapes in the atoms, most often being expressed as a product of a radial and angular function. In the approximation that uses linear combination of atomic orbitals (LCAO), the angular functions have a form of spherical harmonics, associated to the *s, p, d, f* etc atomic orbitals. These functions originate from the exact solutions of Schrödinger equation for hydrogen atom.[2] The solutions for the hydrogen atom result with radial functions exhibiting exponential dependence [exp(-*kr*)] on the distance between the electron and the nucleus. Basis functions with this type of radial dependence are commonly called Slater-type orbitals (STOs), but most modern algorithms use less accurate Gaussian-type functions [exp(-ζr^2)]due to computational efficiency. The Gaussian-type orbitals (GTOs) have certain shortcomings in providing the proper radial shape of the orbital in comparison to more realistic STOs, but computing of the important integrals in the HF procedure is significantly faster with GTOs, for which the analytical solutions are available, while the same integrals with STOs need to be solved via numerical methods. To keep the best features of the both function types, it is a common practice to combine several Gaussian-tpye functions (*primitives*) into the so called *contracted* Gaussian basis functions, approximating STOs.[5]

The desired properties of the basis sets are achieved by adjusting the exponents in the primitives and the contraction coefficients, and of course, the number of primitives. The *split-valence* basis sets have one basis function per core orbital and two or more basis functions per valence orbital, characterized with different exponents ζ. For this reason, the valence functions are known as *double-zeta, triple-zeta* etc. Such basis sets permit the size of a valence atomic orbital to be adjusted (through optimization of the contraction coefficients) to allow for variable atom size and anisotropy. Addition of the *diffuse* functions with low values of ζ allows

greater flexibility of the weakly bound electrons, that is, the electron distribution can extend farther away from nuclei than given by the conventional split-valence basis sets. This is usually necessary for the species carrying negative charge or lone electron pairs. The presence of the many nuclei in the molecules polarizes the atomic orbitals and to account for that effect, basis functions of higher angular momentum (usually for one quantum number) may be added to the split-valence sets and increase the flexibility of the molecular orbital description. More functions describing the orbitals usually mean more accurate final results, but in the everyday life it is necessary to make compromise between the accuracy and computational efficiency.

Over the time, many basis sets were developed and optimized for different systems and computational methods, but most commonly are used Pople basis sets, denoted as *k-lmnG*, and those developed by Dunning and co-workers (cc-pVDZ, cc-pVTZ etc). This thesis primarily contains applications of the Pople basis sets, or to be more precise, the basis sets 6-31G(d), 6-31+G(d), and G3MP2Large. This abbreviation will be used throughout for practical reasons and it stands for 6-311++G(2df,2p) for the first row elements and 6-311++G(3df,2p) for the second row elements.

1.2.4 Electron Correlation

One of the major limitations of the HF method is the lack of the proper description of the electron correlation, that is, HF method includes only the average effect of interelectronic repulsion, but not the instantaneous interaction among the electrons. Although the error resulting from this approximation roughly corresponds to 1% of the total energy of the system for the given basis set, it is often the case that this 1% plays an important role in describing chemical phenomena. Incorporating the electron correlation generally improves accuracy of the computed energies and corresponding geometries. Selected methods that include this correction, based on HF procedure, will be presented in this section.

1.2.4.1 Møller-Plesset Perturbation Theory

Møller-Plesset (MP) perturbation theory finds its basis in Rayleigh-Schrödinger perturbation theory.[1] The initial hypothesis is that Hamiltonian could be written as a sum of

two parts; the first part ($\boldsymbol{H_o}$) is represented by an operator for which solutions are already known. In the context of the molecular orbital theory, a sum over Fock operators was chosen as $\boldsymbol{H_o}$, whose solutions correspond to the HF wave functions. The other part ($\lambda \boldsymbol{V}$) introduces a small perturbation to the first part, and in this particular case, it approximates the electronic repulsion potential. The new solutions can be written in the form of the Taylor expansion for λ:

$$\Psi_\lambda = \Psi^{(0)} + \lambda\Psi^{(1)} + \lambda^2\Psi^{(2)} + \ldots \tag{14}$$

$$E_\lambda = E^{(0)} + \lambda E^{(1)} + \lambda^2 E^{(2)} + \ldots \tag{15}$$

$\Psi^{(0)}$ and $E^{(0)}$ are the already known solutions of $\boldsymbol{H_o}$, while $\Psi^{(n)}$ and $E^{(n)}$ are the *n*th order corrections to the wave function and energy, respectively. The various orders of Møller-Plesset perturbation theory are obtained by setting λ=1 and truncating equations above accordingly. The methods are named to denote after which term the expansion (9) has been truncated, like MP2, where the second term is of the highest order, in MP3 is the third, etc. The improved wave functions obtained with MPn methods include linear combinations of the determinants, and energy corrections contain contributions from the excited states, but their impact depends on the order of the expansion.

The formulation of the reference HF wave function is also very important and affects the overall results. The wave functions with imposed constraint that forces electrons in doubly-occupied orbitals to share the same spatial distributions (RHF or ROHF) are basis for restricted Møller-Plesset theory (RMP). For a closed-shell system where the RHF reference is well-behaved, the perturbation expansion generally converges relatively smoothly and MP theory proves to be a cost-effective method to recover the correlation energy. An open-shell system may be treated with an unrestricted reference wave function (UHF), upon which a UMP expansion is applied. The advantage of this treatment is correct qualitative prediction of the spin polarization, but UMP methods have been found to be very sensitive to spin contamination. One possible solution to avoid this issue is employing ROHF as a reference function. This approach has certain deficiencies (similar to those described above in Section 1.2.2.1), but they are often outweighed by the benefit of lack of spin contamination in the reference wave function. The other way to solve this problem is to project out the various contaminants and re-evaluate the energy, but in some cases it is computationally expensive and complicated. This approach is known as *projected* UMP (PUMP).

Although MP perturbation theory is size-consistent and size-extensive, it is not a variational method. Therefore, sometimes the calculated energies can be lower than the exact energy of the ground state. In addition, MP methods are quite computationally demanding and mostly are used in the *single-point* calculations on the geometries obtained at the lower level of theory.

1.2.4.2 Coupled Cluster Theory

A more elegant approach to deal with electron correlation problem is coupled-cluster (CC) theory.[1] The basic equation of coupled-cluster theory is given below:

$$\Psi_{cc} = e^{T}\Psi_{HF} \tag{16}$$

In Eq. 16, Ψ_{cc} is the exact non-relativistic wave function of the molecule in its ground state, Ψ_{HF} is a normalized HF wave function for the ground state, and e^{T} is an operator defined through a Taylor series expansion in $\boldsymbol{T}$:

$$e^{T} = 1 + \boldsymbol{T} + \frac{\boldsymbol{T}^{2}}{2!} + \frac{\boldsymbol{T}^{3}}{3!} + \ldots = \sum_{k=0}^{\infty}\frac{\boldsymbol{T}^{k}}{k!} \tag{17}$$

The cluster operator $\boldsymbol{T}$ is a sum of the *i*-particle operators $\boldsymbol{T}_i$ (i=1,2,3,...,*n*), where *n* is the total number of electrons.

$$\boldsymbol{T} = \boldsymbol{T}_1 + \boldsymbol{T}_2 + \boldsymbol{T}_3 + \cdots + \boldsymbol{T}_n \tag{18}$$

The effect of $\boldsymbol{T}_i$ on the HF wave function is to generate a sum of all excited determinants in which *i* electrons have been promoted. The coefficients of the generated determinants are known as amplitudes. The role of the operator e^{T} is to express Ψ_{cc} as a linear combination of the Slater determinants that include Ψ_{HF} and all possible electron transitions from the occupied into the virtual orbitals. Incorporation of these excited states in the total wave function accounts for the electron correlation by enabling them to better avoid each other.

Common practice is to approximate the operator $\boldsymbol{T}$ with the inclusion of the limited number of the operators $\boldsymbol{T}_i$. The simplest approximation includes only $\boldsymbol{T}_2$, since double excitations have the strongest contribution to the operator $\boldsymbol{T}$, according to theory. The resulting method is known as *coupled-cluster doubles* (CCD). The next improvement in CCD approach is made by including the operator $\boldsymbol{T}_1$ ($\boldsymbol{T}= \boldsymbol{T}_1+ \boldsymbol{T}_2$), and this method is known as

coupled-cluster singles and doubles (CCSD), because it additionally includes single excitations. Method that includes triple excitations utilizes the operator $\boldsymbol{T}= \boldsymbol{T}_1+ \boldsymbol{T}_2+\boldsymbol{T}_3$ and it is called CCSDT (*coupled-cluster singles doubles and triples*). CCSDT gives quite accurate results for correlation energies, but it is very computationally demanding and usually available only to very small systems, in the combination with rather small basis sets. More common approach is to treat triple excitations with perturbation theory, using a result from MP4 expansion. This method, termed CCSD(T), has been shown to be relatively affordable and it was used in this thesis to calculate accurate energies.

The coupled-cluster wave function, like the MP series, is built upon an HF reference determinant. For the open-shell systems an unrestricted version of the theory (UCCSD(T)) is usually used. Again, the main arising problem is spin contamination, like with the reference UHF wave function, although in a lesser extent than with the UMP wave functions. Here is also possible to introduce restrictions, but lack of uniqueness associated with the ROHF orbital energies, means that several alternatives of RCCSD(T) are possible. This problem can be solved by introducing further restrictions.

.

1.2.5 DENSITY FUNCTIONAL THEORY

Over the years, density functional theory (DFT) has become one of the most popular tools in computational molecular modelling, from chemical problems to material science. DFT is based on the Hohenberg-Kohn theorem,[11] published in 1964., which states that exact ground-state electronic energy is completely determined by the electron density (ρ). Unfortunately, the theorem does not provide any information about the functional relationship of these two quantities and in practice are applied approximate functionals. A crucial step for development of these approximate functionals was made by Kohn and Sham, showing that the ground-state energy can be written in the following form:[12]

$$E_{exact} = E^{T}(\rho) + E^{V}(\rho) + E^{J}(\rho) + E^{XC}(\rho) \qquad (19)$$

E^T describes the kinetic energy in the hypothetical system of non-interacting electrons, E^V is a term describing the Coulombic attraction between electrons and nuclei, while E^J represents the classical repulsion between electron charge distributions. E^{XC} is known as the exchange-

correlation term and, by definition, contains all contributions not accounted for by the first three terms, including the kinetic energy arising from the interacting nature of the electrons.

In practice, the density is written in terms of a set of auxiliary one-electron functions or orbitals:

$$\rho(q) = \sum_{i=1}^{N} \left|\psi_i(q)\right|^2 \tag{20}$$

with the orbitals themselves expanded as a linear combination of basis functions (using a conventional basis set), also known as Kohn-Sham orbitals. With these orbitals it is possible to calculate the exact electron density ρ from Eq. 19, but they bear no physical meaning. If density is expressed like in Eq. 20, the functional relationship between energy and density is known for all the terms in Eq. 19, except for E^{XC}. This small, but vital contribution to the overall energy in practice is calculated by using approximations. In the simplest approximation, E^{XC} can be written as a sum of terms describing electron exchange (E^X) and electron correlation (E^C), both of them being density functionals:

$$E^{XC}(\rho) = E^{X}(\rho) + E^{C}(\rho) \tag{21}$$

Both components can be represented by functionals dependent only on the electron density. This so called local density approximation (LDA) assumes that the electron density can be treated as a homogenous electron gas, where the value of the potential at some point depends only on the value of the density at that point. When the total electron density is expressed as a sum of α and β densities, the approach is known as the local spin density approximation (LSDA). These functionals can be improved by introducing a correction in the form of the gradient of electron density (generalized gradient approximation, GGA). Becke formulated the exchange functional (B) with the gradient correction in 1988.[13] The B exchange functional has been widely used with a range of gradient-corrected correlation functionals. A popular one is that developed by Lee, Yang, and Parr (LYP).[14]

The alternative way to express exchange-correlation energy is to interpolate between the E^{XC} of a system of non-interacting electrons and the E^{XC} of a similar, but interacting system (using the adiabatic correction formula).[15] Since HF theory actually contains the exact exchange energy in the limit of non-interacting electrons, this exchange has been included in the model. A most successful attempt in including the HF exchange was made by Becke with his three parameter model (B3).[16] The combination of this hybrid functional with LYP correlation yields one of the most popular and most widely used functional in chemistry –

B3LYP. This functional was used in this thesis for geometry optimizations, as well as for single-point calculations. Material science and application of DFT methods to extended systems generally prefer PBE GGA approximation,[17] and involve plane-wave description of electron density that satisfies periodic boundary conditions.

Low computational cost and a possibility of modelling large systems important in material science have brought immense popularity to DFT among computational scientists; however, this approach also suffers from some prominent deficiencies. Known difficulties include poor treatment of van der Waals interactions and strongly correlated systems, underestimated gaps in bulk solids and lack of proper scheme for excitations.[18] Most of these failures are due to failures of approximations. In the recent years, numerous functionals were developed and designed to tackle various problems and systems, resulting with wide variety of the available functionals specifically designed for a given problem. Very good example of this trend is given with MO6 family of functionals developed by Truhlar *et al.*, where large number of parameters is used to improve their accuracy.[19] Grimme has developed DFT-D method in which he has successfully introduced empirical corrections to accurately account for weak interactions,[20] while development of time-dependent DFT (TD-DFT) provided tool for exploring the world of excitations.[21]

All this paved the way for DFT to become indispensable tool in molecular modelling from solid-state physics to chemistry, despite the lack of single all-purpose and accurate functional. The existence of so many different functional solved certain problems found in DFT approach, but it also made life of a computational chemist a bit complicated with a continuous search of the optimal functional to assess the problem of interest.

1.2.6 Multilevel Methods

1.2.6.1 Composite Methods (Gaussian-N methods)

Gaussian methods (G1, G2, G3, and G4) were developed after it has been noticed that certain *ab initio* methods are inclined to systematic errors in estimating the ground-state energies for organic molecules. These methods use a correction expression that reaches very accurate results by extrapolation from the ground-state energies calculated at different (*ab*

initio) levels. Lower levels of theory are used for geometry optimizations and frequency calculations, followed by more computationally demanding *single-point* calculations to provide more accurate energies. This approach assumes "basis set additivity", and according to this approximation, the effect of increasing the size of the basis set is not strongly dependent upon the level of correlation involved in the calculation:

$$E[\text{High level / Large basis set}] = E[\text{High level / Small basis set}] + E[\text{Low level / Large basis set}] - E[\text{Low level / Small basis set}] \quad (22)$$

G3(MP2) theory[22] starts from the molecular geometries optimized at UMP2/6-31G(d) level of theory, followed by the *single-point* calculations that include only valence electrons in their correlation treatment (frozen core approximation). The total energy E_o is defined as the energy obtained at QCISD(T)/6-31G(d) level of theory, improved with several corrections:

$$E_0[\text{G3(MP2)}] = E[\text{QCISD(T) / 6-31G(d)}] + \Delta E_{MP2} + \Delta E(\text{SO}) + E(\text{HLC}) + E(\text{ZPVE}) \quad (23)$$

In the expression above, the correction for MP2 method is given by the following term:

$$\Delta E_{\text{MP2}} = E[\text{UMP2/G3MP2 Large}] - E[\text{UMP2/6-31G(d)}] \quad (24)$$

The corrected (scaled) zero-point energies, *E*(ZPVE), are calculated at HF/6-31G(d) level of theory. The remaining corrections in Eq. 23 are the spin-orbital correction, Δ*E*(SO), which refers to the atoms only, and so called high level correction, *E*(HLC). This correction accounts for all the other deficiencies present in this energy calculation by relying on the empirical parameters, adjusted to give the best possible agreement with the experimental data.

For radical species, the G3-RAD method has been specially developed.[23] This method, in contrast to G3, uses RMP instead of UMP energies and RCCSD(T) methods substitutes the UQCISD(T) for taking the electron correlation into account. Geometry optimization is performed with DFT methods. In this work was used B3LYP/6-31+G(d) both for geometry optimization and for frequency calculations. This choice of methods ensures that the final result will not suffer from the spin contamination, but sacrifices the correct prediction of the spin polarization, being the lesser evil, from the energy point of view.

The energies presented in the thesis are usually not absolute electronic energies or total atomization energies. Instead we are largely interested in comparisons between species on a single electronic surface, in which the number of paired and unpaired electrons is maintained. In such cases, the empirical higher level correction (HLC) and spin-orbital (SO) corrections normally associated with the G3 procedures cancels, and the resulting energy differences are purely *ab initio*. Therefore, we referred only to the relative energies or energy differences (Δ*E*):

$$\Delta E = \Delta E\left[\text{RCCSD(T)} \quad / 6-31\,\text{G}+(\text{d})\right]+\Delta E\left[\text{ROMP} \quad 2/\text{G}\,3\,\text{MP}\,2\,\text{Large}\right]- \Delta E\left[\text{ROMP} \quad 2/6-31\,\text{G}+(\text{d})\right]+\Delta E\,(\text{ZPVE}) \tag{25}$$

In case where the atomization energies are important, it is necessary to incorporate the corrections mentioned above, including the scaled zero-point vibrational energies *E*(ZPVE). In this thesis unscaled values of the calculated zero-point energies were generally used.

1.2.6.2 ONIOM

The acronym ONIOM stands for *Our Own N-layered Integrated MO and MM method*, originally developed by Morokuma *et al.*[24] Since its first implementation in Gaussian 98 software package, it has received significant improvements in the upgraded versions of Gaussian 03 and 09.[25-27] This computational technique allows treatment of the (usually large) systems with two or three levels of theory in successive layers. The general approach involves selecting a small model system out of the entire system and treating it with the higher accuracy methods, while the complete system is described at the lower level of theory. In the case of three layers, the middle layer is treated with a method of the intermediate accuracy compared to the high and low level method. A general expression for energy obtained in the following way:

$$E = E^{\text{low level}}_{\text{complete system}} + E^{\text{high level}}_{\text{model}} - E^{\text{low level}}_{\text{model}} \tag{26}$$

ONIOM is able to combine different quantum mechanical (molecular orbital, MO) and semi-empirical methods, but it can also serve as a QM/MM method, where molecular mechanics (MM) is used as low level method (see more about molecular mechanics later in the chapter). Of course, the model system is then treated with the QM method and the expression above can be re-written as:

$$E = E^{\text{MM}}_{\text{complete system}} + E^{\text{QM}}_{\text{model}} - E^{\text{MM}}_{\text{model}} \tag{27}$$

$$E^{\text{QM}}_{\text{model}} = \left\langle \Psi \left| \boldsymbol{H}^{\text{QM}}_{\text{model}} \right| \Psi \right\rangle \tag{28}$$

ONIOM methods are termed according to the QM and MM methods used in the calculation, as ONIOM[QM:MM].

The electrostatic interactions between the QM and the MM part of the system can be treated in different ways, depending on the method used and the aims of the calculation. The

simpler approach is called *mechanical embedding*, where the QM system feels its MM environment only sterically. All the electrostatic interactions of the QM subsystem with the MM part are completely described with the molecular mechanics (interaction between partial charges on the MM part and the partial charges assigned to the QM atoms). The other possibility is to explicitly include those partial charges into the QM Hamiltonian, and incorporation of this term enables polarization of the QM wave function in the response to its environment. This approach is known as *electrostatic embedding*.[26]

$$E^{\text{QM}}_{\text{model}} = \left\langle \Psi \left| \boldsymbol{H}^{\text{QM}}_{\text{model}} - \sum_{i}^{\text{electron}} \sum_{N}^{\text{point charge}} \frac{q_N}{r_{iN}} + \sum_{J}^{\text{nuclei}} \sum_{N}^{\text{point charge}} \frac{Z_J q_N}{r_{JN}} \right| \Psi \right\rangle \tag{29}$$

The first term in the upper expression is equivalent to the Hamiltonian from the Eq. 28, while the second term gives the energy of the interaction between all the electrons present in the model (QM) system and all the point charges in the MM layer. The third term describes the interaction between all the nuclei in the QM subsystem with the outside point charges.

To avoid double counting of the equivalent energies in the calculation of the total energy, it was necessary to introduce a correction for the energy of the interactions between point charges in the MM part and the partial charges located at the centres of the atoms in the model system. The partial charges are commonly derived from the electrostatic potential (ESP), calculated on a grid of points in space.

$$E = E^{\text{MM}}_{\text{complete system}} + E^{\text{QM}}_{\text{model}} - E^{\text{MM}}_{\text{model}} - \sum_{J}^{\text{nuclei}} \sum_{N}^{\text{point charge}} \frac{q_J q_N}{r_{JN}} \tag{30}$$

In this thesis a variation of the ONIOM method, where the model system was treated at the B3LYP/6-31+G(d) level of theory, while AMBER force field was used to describe the MM layer, with the electrostatic embedding (ONIOM[B3LYP/6-31+G(d):AMBER]), was used. Due to its flexibility, ONIOM technique can be combined with composite methods for calculating the accurate energies for systems to large to be explicitly treated with such computationally expensive methods. Of course, in this thesis we used G3(MP2)-RAD and the expression for obtaining the total energy of the system is given below:

$$\begin{aligned} \Delta E(\text{ONIOM}\ [\text{G3(MP2)}\ \text{- URAD}\ :\text{AMBER}\ ]) &= \Delta E(\text{ONIOM}\ [\text{UCCSD(T)/6}\ \text{- 31 + G(d)}\ :\text{AMBER}\ ]) + \\ + \Delta E(\text{ONIOM}\ [\text{ROMP/G3MP2}\ \text{Large}\ :\text{AMBER}\ ]) &- \Delta E(\text{ONIOM}\ [\text{ROMP/6}\ \text{- 31 + G(d)}\ :\text{AMBER}\ ]) + \Delta(\text{ZPVE}) \end{aligned} \tag{31}$$

1.3 Classical Mechanics

An important issue of simulation studies is the accessible time- and length-scale which can be covered by different computational methods. QM methods are appropriate for the description of processes that occur at small length scales, such as bond making and breaking, while QM/MM were developed to address this issue in the systems too large to compute the wave function for the entire system. This is a bridging method toward longer simulations of the processes taking place at nano- and microsecond scale, usually involving conformational changes in the large molecules like proteins. Actually, most of the biological systems are too large to be treated with quantum mechanics – they are rather described with classical mechanics, also known as molecular mechanics. In this approach, which is also based on the Born-Oppenheimer approximation, the potential energy of the system is given as a function of nuclear coordinates only, while the electron motion is included only implicitly. The motion of the particles in these systems is governed by Newton's laws. To gain an insight in the behaviour of the system and possibly estimate important physical quantities describing the system and related processes, certain knowledge of statistical thermodynamics and its fundamental postulates is necessary.

1.3.1 Statistical Mechanics

Statistical mechanics provides a theoretical framework that relates microscopic dynamics or fluctuations with the observed properties of a large system by combining probability theory with the laws of mechanics. All macroscopic systems consist of a large number of particles (N), whose motion is governed by Schrödinger equation or Newton's laws, resulting with the many-body problem. Knowledge of the positions and momenta of these particles define a microstate in which the system resides at a given moment, giving rise to an enormous number of variables required for description of the system. A set of all possible microstates defines the phase space of the given system:

$$(p^N, r^N) \equiv (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N; \boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$$

The total energy of the system is represented by the Hamiltonian, which depends on the coordinate and momenta $\boldsymbol{H}(\mathbf{p},\mathbf{r})$. In a typical case, the potential energy is not a function of

momenta, which allows us to write the Hamiltonian as a sum of separate terms describing potential ($\boldsymbol{U}(\mathbf{r})$) and kinetic energy ($\boldsymbol{T}(\mathbf{p})$) of the system, each being a function of the position and momenta only:

$$\boldsymbol{H}(\mathbf{p},\mathbf{r}) = \boldsymbol{T}(\mathbf{p}) + \boldsymbol{U}(\mathbf{r}) \tag{32}$$

If the initial microstate is specified, the future states of the system can be obtained by solving the equations of motion for each particle. This task for a large number of particles can be achieved only through the usage of the numerical methods and powerful computers, since the analytical solution is not available for systems with more than two particles. Although we witness the exponential growth of the computer power, many of the important (biological) systems are still not accessible to this type of the investigation due to their size, as the complexity of the system rises with the increasing number of particles. It would seem that prediction of the bulk properties from the microscopic behaviour of the system is an impossible quest, but surprisingly, it takes only a few thermodynamic variables to describe a macroscopic system in equilibrium. Namely, with the assumption that the observed properties of the macroscopic systems are the result of the underlying statistical laws, the necessity for precise knowledge of the dynamics of *N* particles vanishes.

The fundamental postulate of statistical mechanics states that all the microscopic states are equally probable when the system resides in the state of thermodynamic equilibrium. If *N* independent measurements were made on a system, a certain value could be attributed to the observed property *Q* based on these measurements based on the following expression:

$$Q_{obs} = \sum_x \frac{n_x}{\mathcal{N}} Q_x = \sum_x P_x Q_x \equiv \langle Q \rangle \tag{33}$$

where Q_x is the expected value of Q when the system finds itself in the state x, n_x denotes the number of times the system has visited the state x, while P_x gives the probability of finding the system in state x. The observed value of the chosen variable $\langle Q \rangle$ corresponds to the average value for the given ensemble, where the brackets $\langle \ldots \rangle$ denote the average value of the variable enclosed. The ensemble is defined as an assembly of all the possible microstates that system can occupy under given constraints that characterize the system macroscopically. The ergodic hypothesis, another important concept in statistical mechanics, assumes that if given enough time, the system will visit all the microstates accessible to the system under the given constraints, depending on the choice of the ensemble. The observed time average would be equivalent to the ensemble average:

$$\langle Q \rangle \equiv \bar{Q} = \lim_{t\to\infty} \frac{1}{\tau} \int_0^\tau Q(\boldsymbol{r}(t))dt \tag{34}$$

The systems that obey this hypothesis are called ergodic. It is believed that majority of many-body systems are ergodic, but problems with the proper sampling of the entire phase space are often encountered during the computer simulations. In that case, only the fragments of the phase space are explored, while the other regions remain inaccessible due to slow diffusion or high barriers between the states. Therefore, the calculated averages are strongly influenced by the initial conditions and systems exhibit so called quasi-nonergodicity.[28]

As mentioned earlier, the macroscopic state of the system is characterized by keeping a handful of the natural variables fixed, such as volume, energy, temperature etc. Choice of these constraints results with definition of the different ensembles. The most commonly used ensembles are listed below:

- *microcanonical ensemble* (N, V, E) – the assembly of all states where number of particles N, volume V, and total energy E of the system are kept fixed. The system is completely isolated and there is no exchange of matter and energy with the environment.
- *canonical ensemble* (N, V, T) – the assembly of all states where number of particles N, volume V, and temperature T of the system are kept fixed. The system is able to exchange the energy with its surroundings.
- *grand canonical ensemble* (μ, V, T) – the assembly of all states where chemical potential μ, volume V, and temperature T are fixed. This system is allowed to exchange energy and matter with the environment, in order to keep the chemical potential constant.

For the purpose of this thesis, it is enough to describe the canonical ensemble in greater detail, where the more information about other ensembles and statistical mechanics in general can be found in the standard textbooks on statistical mechanics.[29] In the canonical ensemble, the number of particles in the system N and its volume V are kept constant, while the energy is allowed to fluctuate. The system is kept in equilibrium by coupling with the heat bath at constant temperature T. The stationary probability distribution of the canonical ensemble is given with the following expression:

$$P = \frac{e^{-\beta E_x}}{\sum_x e^{-\beta E_x}} = \frac{e^{-\beta E_x}}{Z} \tag{35}$$

In the equation above, β denotes the inverse temperature (β=1/$k_B T$), E_x is the energy of the state x, while $Z = \sum_x e^{-\beta E_x}$is canonical partition function. The sum in the partition function goes over all the accessible states, serving as the normalization factor in the probability

distribution. This distribution gives the probability of finding the system in the state x with respect to the temperature T and energies of all the possible states of the system.

For a classical system of N indistinguishable particles, the energy of the system is given with the Hamiltonian $\boldsymbol{H(p,r)}$, which is a function of $3N$ momenta $\boldsymbol{p}$ and position coordinates $\boldsymbol{r}$ describing a given state $x(\boldsymbol{p,r})$. In that case, partition function Z can be written as:

$$Z = \sum_x e^{-\beta E_x} = \frac{1}{N!h^{3N}} \iint e^{-\beta H(\boldsymbol{p,r})}\, d\boldsymbol{p}d\boldsymbol{r} \tag{36}$$

Knowledge of the partition function Z allows calculation of various thermodynamic properties $Q(\boldsymbol{p,r})$ of the system as an expected value for the given probability distribution:

$$\langle Q \rangle = \frac{\iint Q(\boldsymbol{p,r}) e^{-\beta H(p,r)}}{\iint e^{-\beta H(p,r)} d\boldsymbol{p}d\boldsymbol{r}} \tag{37}$$

Relationships between the partition function and some important state functions are given below:

- internal energy: $E = -\frac{dlnZ}{d\beta}$;
- entropy: $S = -k_B(lnZ + \beta E)$;
- Helmholtz free energy: $F = -\frac{lnZ}{\beta}$.

However, analytical formulation of partition function is available only for the simplest systems, while this is generally not possible for large systems with complex interactions. This aggravates calculation of the absolute values of the quantities mentioned above. In order to avoid this issue, often the calculated value corresponds to difference between two states of interest.

Computer simulations have become indispensable tool in modern science, including physics, chemistry and molecular biology. To be able to simulate large molecular systems and their complex properties, it was necessary to construct appropriate theoretical framework. One of the fundamental steps is the definition of the system Hamiltonian and development of the algorithms that are capable of generating statistical ensembles of the systems of interest. This resulted with a number of different approaches to address different types of problems.

1.3.2 FORCE FIELDS

As mentioned earlier, most of the biological systems are too large to be treated with quantum mechanics and these systems are rather described with classical mechanics, also known as molecular mechanics. In this approach based on Born-Oppenheimer approximation, the potential energy of the system is given as a function of nuclear coordinates only, while the electron motion is implict. Definition of Hamiltonian in these "classical" systems relies on the simplified description of bonding and non-bonding interactions between atoms and molecules. The interactions are represented by classical terms and parameterized to reproduce experimental data and/or results of QM calculations, providing a classical description of the potential energy of the system, known as a *force field*. A rather general functional form of the force fields is given with the following expression:

$$E_{FF} = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,o})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,o})^2 + \sum_{torsions} \frac{V_n}{2}(1 + \cos(n\omega - \gamma)) + \sum_{i=1}^{N}\sum_{j=1}^{N}\left(4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon_o r_{ij}}\right)$$

Namely, different functional terms describing interactions between atoms and molecules are used in different force fields. Bonds and angles are commonly approximated with harmonic oscillators; the electrostatic interactions are described with Coulomb law, and van der Waals forces with Lennard-Jones potential. These can, however, be replaced with other functions to improve the quality of the potential energy description. This can also be achieved by including additional terms in the force field expression. Force field parameters can be fitted to the experimental or theoretical data, computed with high-level QM methods.

In the present day, different force fields and software packages for modelling various systems are available. The most popular examples of force fields developed for biological systems such as proteins, nucleic acids, sugars and lipids include AMBER, GROMACS, CHARMM and OPLS, with a long list of the available parameters. The mentioned force fields use different functional forms to describe the potential energy, but also different procedures to devise required parameters. These differences complicate the exchange of the parameters between the force fields and the parameters are generally not considered to be transferable between the force fields. In the thesis, we used AMBER force fields (ff99SB, ff03, gaff) implemented in the AMBER software package.

1.3.3 Molecular Dynamics

Once the Hamiltonian has been defined, the appropriate statistical ensemble needs to be generated to extract the desired properties of the system. The traditional simulation methods for many-body systems can be divided basically into two classes, stochastic and deterministic simulations, which are largely represented by the Monte Carlo (MC) method and the molecular dynamics (MD), respectively.

To generate an ensemble of configurations, Monte Carlo simulations explore the configurational space by trial moves of particles, usually within the so-called Metropolis algorithm. In this approach, the energy change between two successive moves is used as a criterion to accept or reject a new configuration. Namely, if a move results with a configuration that is lower in energy than its predecessor, it is automatically accepted, while those with higher energy are accepted with a probability governed by Boltzmann statistics. This algorithm ensures the correct limiting distribution and properties of a given system can be calculated by averaging over all Monte Carlo moves within a given statistical ensemble (where one move means that every degree of freedom is probed once on average).

In contrast to this „static“ approach, molecular dynamics propagates system in time by numerically solving Newton's equations of motion, resulting with moving particles to new positions and assigning new velocities at these new positions. The generated trajectories provide additional information about system dynamics by probing the entire phase space, not only configurational space as done in MC approach. Both methods are complementary in nature but they lead to the same averages of static quantities, given that the system under consideration is ergodic and the same statistical ensemble is used.

In order to properly simulate the behaviour of the complex systems, such as biological macromolecules, it is important to build a realistic model that is able to reproduce experimental findings, such as distribution functions, but also obey the theoretical constraints imposed onto the system to keep in accordance with the underlying physical laws like energy or momentum conservation. To run a successful molecular dynamics simulation, there are several important steps in the preparation of simulation that one has to account for:

(i) The choice of the appropriate force field for a given system and problem type, which is able to correctly describe the interactions between the particles in the system. AMBER offers parameters developed for proteins, nucleic acids and common sugars.

(ii) An integration algorithm that will propagate particle positions and velocities from time t to $t + \Delta t$. It is a finite difference scheme which propagates trajectories discretely in time. AMBER uses the so-called *leap frog* algorithm for integration. The adequate time step Δt has to be chosen to guarantee stability of the integrator, i.e. there should be no drift in the system's energy.

(iii) The choice of the statistical ensemble and control methods of the corresponding thermodynamic quantities like pressure, temperature or the number of particles. The natural choice of an ensemble in MD simulations is the microcanonical ensemble (NVE), since the system's Hamiltonian without external potentials is a conserved quantity. Nevertheless, there are extensions to the Hamiltonian which also allow simulating different statistical ensembles, such as NVT and NPT. In AMBER, there are several available thermostats (Berendsen, Andersen, Langevin), while the pressure is controlled by using weak-coupling scheme.

The steps listed above provide the general framework of an MD simulation. The quality of simulation results will depend on the quality of the force field used to build a model system and accompanying algorithms that control the system behaviour under given constraints. The obtained results should be verified against the available experimental data and theoretical predictions. If there is significant deviation between simulation results and measured properties, the model needs to be improved until satisfactory agreement has been achieved for the properties of interest.

1.3.4 Free Energy Calculation

Free energy is one of the most important state functions in chemistry, as it governs the spontaneity of a process and gives the probability of the system adopting a given state in the canonical ensemble:

$$P(\boldsymbol{p},\boldsymbol{r}) = \frac{e^{-\beta H(\boldsymbol{p},\boldsymbol{r})}}{\iint e^{-\beta H(\boldsymbol{p},\boldsymbol{r})} d\boldsymbol{p}d\boldsymbol{r}} = \frac{e^{-\beta E_x}}{Z} = \exp\{\beta F - \beta E_x\} \tag{38}$$

In cases where Hamiltonian is separable in its variables $\boldsymbol{p}$ and $\boldsymbol{r}$, the kinetic contribution can be integrated out and we obtain only the configurational probability density depending on the potential energy $U(\boldsymbol{r})$:[30]

$$P(\boldsymbol{r}) = \frac{e^{-\beta U(r)}}{\int e^{-\beta U(r)} d\boldsymbol{r}} \tag{39}$$

Due to the complexity and high-dimensionality of the phase space, it is common to choose a set of variables that play a crucial role in the process of interest, known also as a reaction or generalized coordinate. In this way, we observe the system evolution only as a function of the chosen subset of the coordinates (ξ), instead of the entire phase space. The reduced probability distribution $P(\xi)$and corresponding free energy can be expressed as:

$$P(\xi) = P(\boldsymbol{r})\delta\big(\xi - \xi(\boldsymbol{r})\big) = \frac{\int e^{-\beta U(r)} \delta\big(\xi - \xi(\boldsymbol{r})\big) d\boldsymbol{r}}{\int e^{-\beta U(r)} d\boldsymbol{r}} = \frac{Z(\xi)}{Z} \tag{40}$$

$$F(\xi) = -\frac{lnZ(\xi)}{\beta} \tag{41}$$

As mentioned earlier, we are interested in the free energy difference when going from state A to state B:

$$\Delta F = F_{\mathrm{B}} - F_{\mathrm{A}} = -\frac{1}{\beta} ln \frac{Z_{\mathrm{B}}}{Z_{\mathrm{A}}} \tag{42}$$

However, on many occasions it is also important to know the free energy profile of the given process and the barriers between those two states. In that case, the so-called potential of mean force (PMF) is calculated. The origin of this term, first introduced by Kirkwood in 1935,[31] comes from differentiating the free energy with respect to the selected reaction coordinate, e.g. an atomic coordinate $\boldsymbol{r}_1$:[32]

$$-\frac{dF}{dr_1} = -\left\langle \frac{dH}{dr_1} \right\rangle_{r_1} = -\left\langle \frac{dU}{dr_1} \right\rangle_{r_1} = \langle \boldsymbol{F}_1 \rangle_{r_1} \tag{43}$$

where $\langle \dots \rangle_{r_1}$ denotes the average computed with the fixed value of $\boldsymbol{r}_1$. The obtained $\boldsymbol{F}_1$ is the force acting on the chosen coordinate $\boldsymbol{r}_1$ averaged over all the other variables. Hence, $A(\boldsymbol{r}_1)$ can be thought of as the mean potential or PMF for $\boldsymbol{r}_1$. Since the generalized coordinate ξ can be any function of atomic positions, $-dF/d\xi$ is not necessarily a force, but the interpretation remains the same. Specifically, $-dF/d\xi$ is the mean force exerted on the generalized particle ξ.

$$\frac{dF}{d\xi} = \left\langle \frac{\partial H}{\partial \xi} \right\rangle_{\xi} = \left\langle \frac{\partial U}{\partial \xi} - \frac{1}{\beta} \frac{\partial \ln |J|}{\partial \xi} \right\rangle_{\xi} \tag{44}$$

The term $|J|$ is the determinant of the Jacobian matrix upon changing from Cartesian to generalized coordinates. It measures the change in the volume element due to change in the coordinates and it is effectively an entropic contribution. Choice of the generalized coordinate is somewhat arbitrary, as it depends greatly on the problem of interest and type of the system.

It can be a distance, an angle or some other function of the Cartesian coordinates. Definition of the reaction coordinate ranges from the straightforward to more obscure cases, where more effort has to be invested to define the variable that properly describes the given process in terms of free energy.

To get a better insight in the procedures behind the free energy calculations, a simple system of two methane molecules enclosed in a box of water will be used to illustrate the free energy methods presented in this section. The goal is to compute potential of mean force (PMF) between these two molecules as a function of distance.

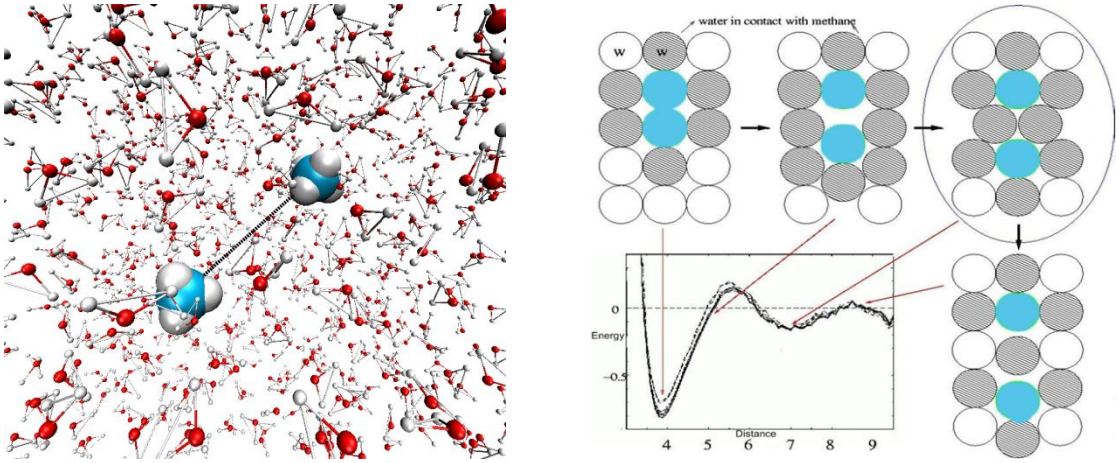


Figure 1.1 Simple system of two methane molecules in the box of water is taken as an illustrative example for free energy methods presented in this section (panel left). Potential mean force is calculated as a function of distance between these two molecules (r). The configurations methanes and solvent molecules that correspond to the minima of the PMF curve obtained for the described system (panel right, originally from Andy Hsu's thesis).[33]

According to the expressions listed above, it should be possible to estimate the PMF from statistical data sampled along the chosen reaction coordinate during the molecular dynamics simulation. This MD simulation should be long enough to exhibit ergodic behaviour and to allow the system to visit all the configurations of importance in the reduced phase space. In this particular case, that is the distance between two methanes in the interval 3Å – 10 Å and sampling is done in that phase space (Figure 1.1a). The collected data is then binned in a histogram (Figure 1.1b) to obtain the reduced probability distribution $P(\xi)$, from which obtaining PMF is straightforward (Figure 1.1c). However, to obtain the correct PMF (Figure 1.1d) it is necessary to add a term that accounts for changing Jacobian matrix from Cartesian to generalized coordinate.

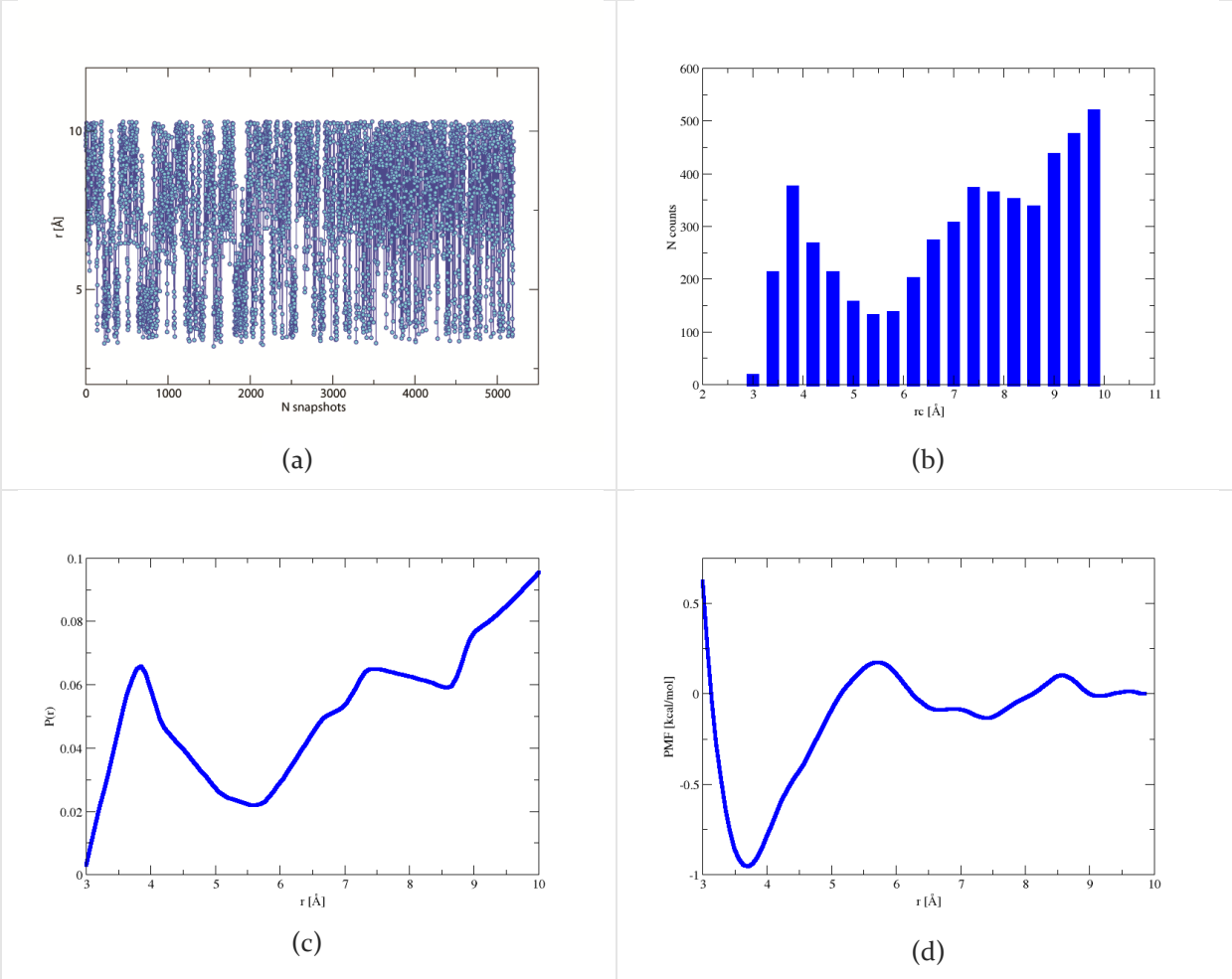


Figure 1.2 (a) Data collected during the unbiased NVT simulation of two methanes in box of water at 298 K. The chosen reaction coordinate (r) is distance between two methanes. **(b)** Histogram of distances between 3-10 Å collected from NVT simulation. (c) Probability distribution P(r) derived from the histogram. (d) Final PMF obtained from probability distribution and corrected for the Jacobian term (2RTln(r)).

One of the important steps in the free energy calculation regards the generation of configurations that obey desired probability distribution $P(\boldsymbol{r})$ for the given system. These configurations are sampled from molecular dynamics. Adequate sampling can be a quite challenging task in case of the complex systems, due to quasi-nonergodic behaviour of the systems. To overcome this issue, various strategies have been developed to efficiently sample configurations along the chosen coordinate and retrieve the free energy information from the sampled data. Very popular approaches in free energy calculation include free energy perturbation, thermodynamic integration, umbrella sampling, steered molecular dynamics etc. Important methods for enhanced sampling used in this thesis and the corresponding PMF estimators from the collected data are described in more detail in the following text. To enable a better familiarity with these methods, they are illustrated with the example of two methane molecules in water and the calculation of the PMF as a function of their separation.

1.3.4.1 Umbrella Sampling

One of the major challenges in free energy calculations is overcoming the inability to visit all important configurations during molecular dynamics simulation and to obtain correct probability distribution for the given system. If proper sampling can be achieved with an unbiased simulation at a given set of conditions, it is possible to count directly the visited states and use histograms to estimate probability distribution for the given reaction coordinate, as shown in the example above.

However, even in very long simulations, some states are rarely visited due to their high energy. Umbrella sampling is a method that uses modified Hamiltonian to improve sampling in those rarely visited regions by introducing suitably chosen biasing potential or umbrella (V^B) as a function of the chosen collective variable $\xi(\boldsymbol{r})$:

$$U^B = U(\boldsymbol{r}) + V^B(\xi(\boldsymbol{r})) \tag{45}$$

The probability distribution (P^B) for system evolved with this altered potential energy is then given with:

$$P^B(\xi) = \frac{Z}{Z^B}\exp\left(-\beta\left(V^B(\xi) + A(\xi)\right)\right) \tag{46}$$

where Z^B is the canonical partition function for the potential U^B. From this biased probability distribution P^B it is possible to extract the unbiased free energy according to the following expression:

$$F(\xi) = -\frac{1}{\beta}\ln\left(P^B(\xi)\right) - V^B(\xi) - f^B \tag{47}$$

In the equation above, the factor f^B is a constant independent of ξ:

$$f^B = \frac{1}{\beta}\ln\frac{Z}{Z^B} \tag{48}$$

The biased probability distribution is estimated from the as a normalized histogram from n independent data points collected in the simulations:

$$P^B(\xi) \sim \frac{1}{n\Delta\xi}\sum_{t=1}^{n} X(\xi_t) \tag{49}$$

where $X(\xi_t)$=1 if $\xi_t \in [\xi, \xi + \Delta\xi]$ and zero otherwise.[34] To reduce the statistical error to minimum, the optimal choice of the biasing potential would equal $V^B(\xi) = -A(\xi)$. However, this is exactly the quantity that we are trying to determine and it cannot be used in practical

applications. The usual strategy is to divide the coordinate ξ in small intervals, or windows, and sample around a predefined value (ξ_i). The common choice of the function that confines the system to sample the values of ξ in the defined window is a harmonic potential centred on successive values of ξ_i:

$$V^{Bi}(\xi) = k_i(\xi - \xi_i)^2 \tag{50}$$

This stratification helps to achieve a more efficient configurational sampling in the given region, but each window provides only a fraction of the desired information about PMF along the chosen coordinate.

In the case study of two methanes, a series of umbrella potentials were used to achieve proper sampling along the reaction coordinate. The experiment was taken over from Trzesniek at al.[35] and it was designed to have more closely spaced windows at shorter distances, while the spacing between the windows was increased at larger distances between the methanes.

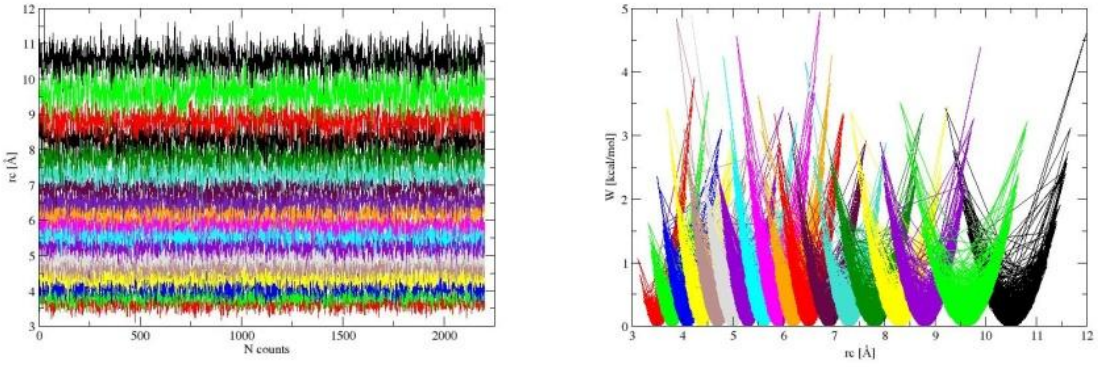


Figure 1.3 Umbrella sampling of distances between two methane molecules in a box of waters in an interval 3Å – 10.5Å. The spacing between the windows used in this experiment increases with the increasing distance between two methanes.

The spacing of the windows depends sensitively on the strength of the harmonic potential used to restrain the system in the given window. The usage of a stronger potential usually implies closer spacing of the windows to achieve overlap between distributions from neighbouring windows. The choices of the force constant and window spacing are quite system dependent and rely heavily on a trial-and-error approach. Some general strategy usually involves placing window centres in larger intervals and later additional windows can always be added in the regions where satisfactory sampling or distribution overlap has not been achieved. The unbiased probability distribution obtained from the *i*th window corresponds to:

$$P^i(\xi) = P^{Bi}(\xi)\exp\left(\beta\left(V^{Bi}(\xi) - f^i\right)\right) \tag{51}$$

To obtain the final estimate about PMF for the given coordinate, the overlapping distributions from different windows need to be combined and unbiased. The most popular is weighted histogram analysis method (WHAM), initially introduced by Ferrenberg and Swendsen,[36] and later extended by Kumar *et al.*[37] The best estimate of the probability distribution $P(\xi)$ is assumed to be a linear combination of the window specific probability $P_i(\xi)$ with weights $\pi_i(\xi)$:

$$P(\xi) = C\sum_i \pi_i(\xi)P^i(\xi) \tag{52}$$

These weights are determined by minimizing the expected statistical error on $P(\xi)$ and result with the final expression for the probability distribution:

$$P(\xi) = C\frac{\sum_k n_k P^{Bk}(\xi)}{\sum_j n_j \exp\left(-\beta\left(V^{Bj}(\xi) - f^j\right)\right)} \tag{53}$$

The constants f^i in the solution are obtained by solving their defining equation in a self-consistent manner. The uncertainties in the obtained free energy are usually computed by performing a bootstrap error analysis.

Recently, novel estimators have been developed that are considered superior to the WHAM because of their ability to additionally reduce statistical errors. One of these methods is called umbrella integration (UI).[38] This approach combines window sampling technique with thermodynamic integration,[39] which is the limiting case of a strong bias. The method avoids iterations necessary to calculate f^i by calculating the unbiased derivative of free energy from the biased probability distribution of a window i:

$$\frac{\partial F_i}{\partial \xi} = \frac{\partial F_i^B}{\partial \xi} - \frac{\partial V_i^B}{\partial \xi} = -\frac{1}{\beta}\frac{\partial \ln P_i^B}{\partial \xi} - \frac{\partial V_i^B}{\partial \xi} \tag{54}$$

To combine the different windows, the reaction coordinate is divided into bins that span the whole range of ξ and are independent of the windows. For each bin, centred at ξ_{bin} bin, the windows are combined by a weighted average:

$$\frac{\partial F(\xi)}{\partial \xi}|_{\xi_{bin}} = \sum_i^{windows} p_i\left(\xi_{bin}\right)\left(\frac{\partial F_i(\xi)}{\partial \xi}\right)_{\xi_{bin}} \tag{55}$$

$$p_i(\xi) = \frac{N_i P_i^B(\xi)}{\sum_i N_i P_i^B(\xi)} \tag{56}$$

The p_i is the normalized weight for N_i number of steps sampled for window i. Umbrella integration does not require window distributions to overlap, unlike WHAM. This is because

the weights are not calculated from the numerical distribution, which is prone to statistical error, but from the normal distribution for $P_i^B(\xi)$.

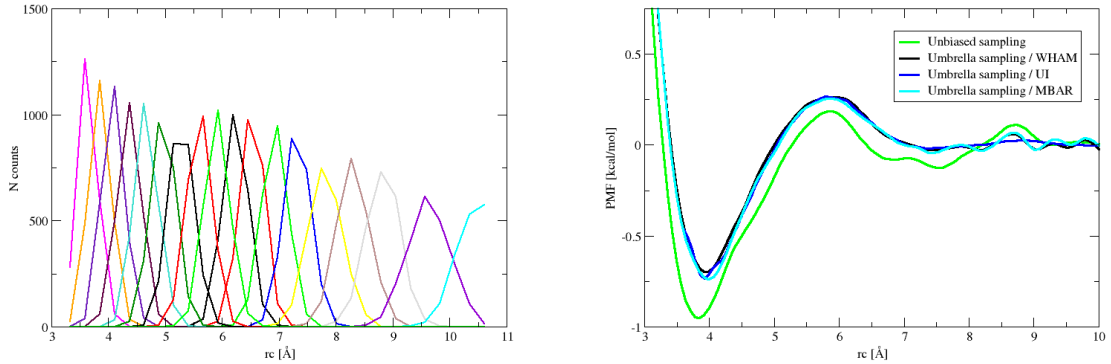


Figure 1.4 (a) Biased probability distributions extracted from the windows constructed to sample distances between two methanes. (b) PMF between two methane molecules in a box of waters obtained by using different estimators (WHAM, UI, MBAR) and same datasets. The curves are compared to the PMF resulting from the unbiased simulation.

Another alternative is given with the most recent method for estimating the PMF from the multiple equilibrium states based on Bennett acceptance ratio.[40] This multistate Bennett acceptance ratio (MBAR) estimator does not require the sampled energy range to be discretized to produce histograms, eliminating bias due to energy binning. It is stated that in the large sample limit, MBAR has the lowest variance of any known estimator for this type of data and provides the uncertainties. In this approach, the dimensionless free energy is estimated for configurations that correspond to Boltzmann distribution $q_i(\xi) \equiv \exp[-u_i(\xi)]$ by solving self-consistently the following expression:

$$\widehat{f}_i = -\ln \sum_{j=1}^{K} \sum_{n=1}^{N_j} \frac{\exp[-u_i(\xi_{jn})]}{\sum_{k=1}^{K} N_k \exp[\widehat{f}_k - u_k(\xi_{jn})]} \tag{57}$$

The estimated free energies $\widehat{f}_i$are determined uniquely only up to an additive constant, so only differences $\Delta\widehat{f_{ij}} = \widehat{f}_j - \widehat{f}_i$ will be meaningful. The uncertainty in the estimated free energy difference can be computed from the covariance matrix which is an integral part of this calculation:

$$\delta^2 \Delta\hat{\hat{f}}_{ij} \equiv \mathrm{cov}\left(-\ln \hat{\hat{Z}}_j/\hat{Z}_i\,, -\ln \hat{Z}_j/\hat{Z}_i\right) \tag{58}$$

More details about these equations can be found in the original publication of Shirts and Chodera.[40]

Figure 1.4 shows results extracted from the umbrella sampling datasets by the estimators listed in the section – WHAM, UI and MBAR. As it can be seen, all these estimators provide roughly the same answer for the given dataset, although they use the available data in different ways. There is a slight mismatch between PMF obtained from the unbiased simulation by direct counting and the PMF originating from umbrella, but the main features are captured in both cases. This implies that system configurations were properly sampled for a given coordinate, but even for a system as simple as two methanes in water, this is not a trivial task.

1.3.4.2 Steered Molecular Dynamics

Steered molecular dynamics is a non-equilibrium method that can be used to compute equilibrium free energies. Namely, the system is driven from the initial to the final state by applying the external force along the chosen coordinate. In the AMBER software package, it is done by restraining a system to sample configurations in region around the centre (ξ_o) of a harmonic potential $V_r(t)$, which is moved in a time-dependent fashion along the coordinate ξ (Figure 1.5):

$$V_r(t) = \frac{1}{2}k\left(\xi - \xi_0(t)\right)^2 \tag{59}$$

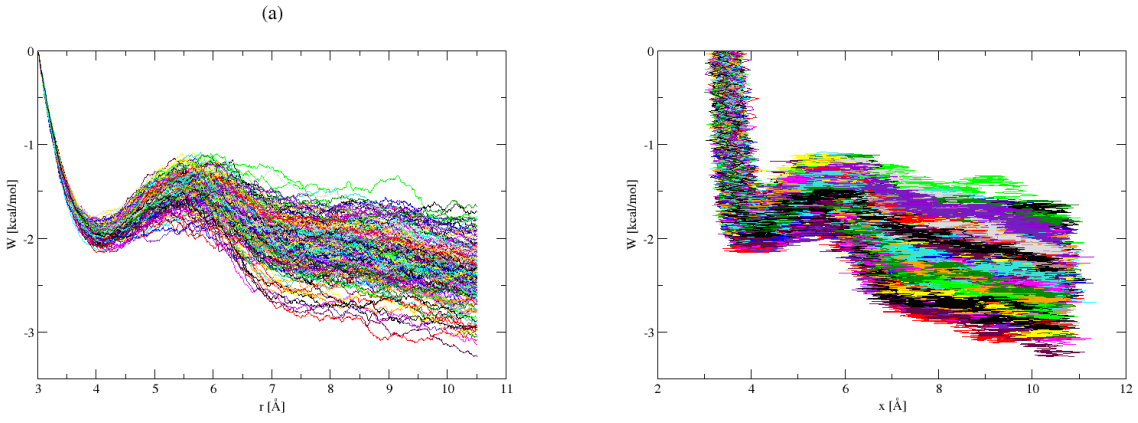


Figure 1.5 The work invested to drive two methane molecules apart in a box of waters along the chosen reaction coordinate, which has been defined as the distance between the solute molecules (panel left). However, the actual distances sampled during the steered molecular dynamics simulations are distributed around the defined centres along the pathway (panel right).

It can be thought of as a time-dependent umbrella sampling (left panel in Figure 1.5). The estimator of the free energy difference between the initial and final state (ΔF) is provided with the Jarzynski identity:[41]

$$\langle e^{-\beta W} \rangle = e^{-\beta \Delta F} \tag{60}$$

The Jarzynski identity relates the work (W) invested to bring the system from the initial (ξ_i) to the final state (ξ_f), averaged over multiple trajectories connecting these two states, to the change in the free energy associated with that process. The work is a functional of the trajectory and it will be different for each realization of the $\xi(t)$. If the transition from ξ_i to ξ_f is done in the infinite time interval, the system is able to remain close to equilibrium conditions and work is equal to the corresponding free energy:

$$\lim_{t\to\infty} W_t = \Delta F \tag{61}$$

For these quasi-static processes, the distribution of the measured work values can be described with a delta function with the exact value in ΔF. However, the experiments are done in a finite time, which results with higher average values of the work done for a given process ($\langle W \rangle > \Delta F$), distributed with a finite variance. There is also a possibility that a certain individual realization requires amount of work less than ΔF.

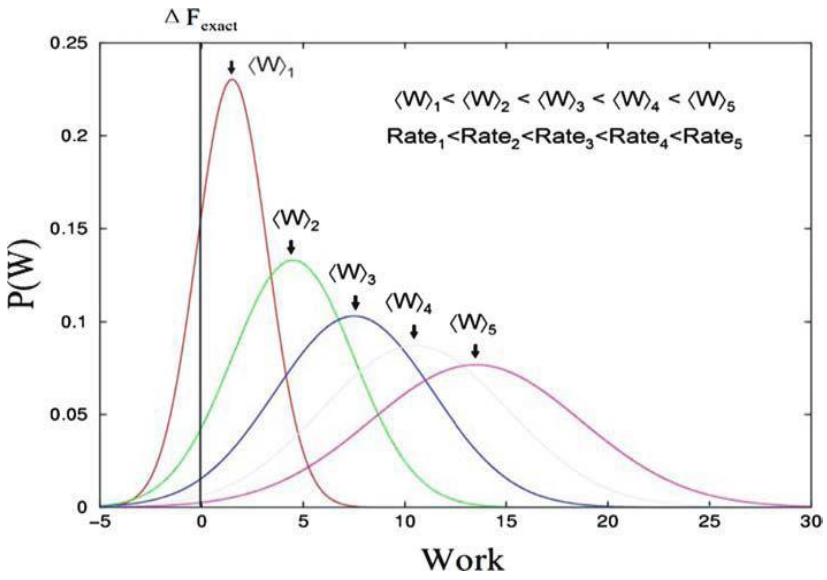


Figure 1.6 The probability distributions of the amount of work required to drive the system from the initial to the final state depend on the switching rate used in the experiment.[42,43]

The distribution of the work values obtained from numerous trajectories depends on the transition rate – faster switching shifts the average work toward higher values and increases the variance (Figure 1.6). In other words, there is an increase in the dissipated work ($W_d =$

$\langle W \rangle - \Delta F$), that is associated with the increase of entropy during an irreversible process. The Jarzynski identity requires a weighted average of these work probability distributions, and the weight factors have the exponential form. The trajectories corresponding to low amounts of total work have the greatest impact on this non-linear average. These trajectories are rare, and with increasing the switching rate, their number is additionally decreased, resulting with convergence problems.

The Jarzynski identity is a special case of a more general Crooks fluctuation theorem for stochastic microscopically reversible dynamics:[44]

$$\frac{\mathrm{P_F}(+\omega)}{\mathrm{P_R}(-\omega)} = e^{+\omega} \tag{62}$$

Here ω is the entropy production of the driven system measured over sometime interval, $\mathrm{P_F}(+\omega)$ is the probability distribution of this entropy production, and $\mathrm{P_R}(-\omega)$ is the probability distribution of the entropy production when the system is driven in a time-reversed manner. If the entropy production is defined as $\omega_F = -\beta\Delta F + \beta W = \beta W_d$, the fluctuation theorem can be expressed in terms of the amount of work performed on a system that starts in the equilibrium:

$$\frac{\mathrm{P_F}(+\beta W)}{\mathrm{P_R}(-\beta W)} = e^{-\beta\Delta F + \beta W} = e^{+\beta W_d} \tag{63}$$

Namely, it is possible to extend the Jarzynski identity to a more general class of equalities between the work and the free energy change. Crooks' path ensemble average relates the forward average of an arbitrary functional $\mathcal{F} = \mathcal{F}(\Gamma)$ of the phase space trajectory $\Gamma = \{p(t), r(t)\}$ to its work-weighted average in the reverse process:

$$\langle \mathcal{F} e^{-\beta W_\mathrm{d}} \rangle_F = \langle \hat{\mathcal{F}} \rangle_R \tag{64}$$

If $f(W)$ is any finite function of the work then the path ensemble average is given with:

$$e^{-\beta\Delta F} = \frac{\langle f(+W) \rangle_F}{\langle f(-W) e^{-\beta W} \rangle_R} \tag{65}$$

where $\langle \cdots \rangle_F$ denotes average over forward trajectories, and $\langle \cdots \rangle_R$ for time-reversed trajectories. This is important in the context of improving the accuracy and convergence of free energy calculations based on the Jarzynski nonequilibrium work relation, or Crooks theorem in the more general case. The most accurate ΔF from a statistical perspective is obtained with Bennett acceptance ratio method.[45] Although it has been derived for the instantaneous switching, Bennett's method can be adjusted for finite switching times. The minimal statistical error is achieved when following expression is used:

$$e^{-\beta\Delta F} = \frac{\langle(1+\exp\{+\beta W+C\})^{-1}\rangle_F}{\langle(1+\exp\{+\beta W-C\})^{-1}\rangle_R}\exp(-C) \tag{66}$$

If we assume that we have collected n_F measurements of the work from the forward process and n_F from the reverse process, then the optimal choice of the constant C is $-\beta\Delta F + \ln n_{\mathrm{F}}/n_{\mathrm{R}}$ and the equation is iteratively solved. The accuracy of the calculated free energy difference is improved by including the values of work obtained with reverse processes in addition to those obtained from the forward trajectories. Namely, it has been observed that the forward trajectories with the highest influence on the weighted average strongly resemble to typical trajectories generated in time-reverse manner.[46] In this way, the estimators of the free energy difference or potential of mean force are optimized.

The estimator for the potential of mean force from driven non-equilibrium processes was firstly developed by Hummer and Szabo[47,48] for unidirectional experiments, followed by derivation of the estimators that use both forward and reverse trajectories by Minh and Adib.[49,50] Hummer and Szabo's method divides data obtained from the repeated pulling experiments into time slices (t), which correspond to the intervals of the reaction coordinate similar to "windows" from umbrella sampling. Each trajectory contributes to the appropriate time slice, in which all the data is combined and binned into histograms. The final PMF ($\Delta G_0(z)$) is estimated by using weighted histogram analysis method (WHAM):

$$e^{-\beta\Delta G_0(z)} = \frac{\sum_t \langle\delta(z-z_t)e^{-\beta W_0^t}\rangle_F e^{\beta\Delta F_t)}}{\sum_t e^{-\beta[V(z,t)-\Delta F_t])}} \tag{67}$$

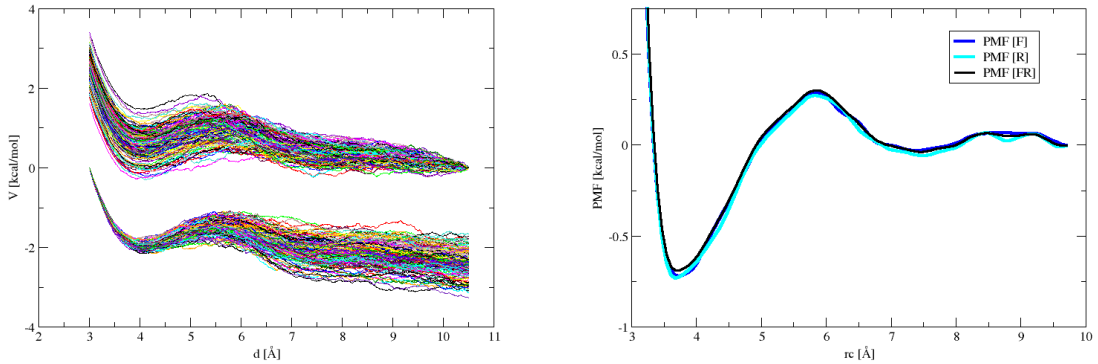


Figure 1.7 Trajectories generated by SMD for separating two methanes in water in a forward and reverse direction along the chosen coordinate (left panel). From each dataset the potential of mean force was obtained by using Hummer and Szabo's estimator developed for unidirectional experiments, while Minh and Adib's estimator was used to combine both forward and reverse trajectories into a single curve.

Minh and Adib's method is an extension of Hummer and Szabo's approach to bidirectional pullings. In this approach, the obtained trajectories are again divided into time slices, but now data from both forward and reverse processes are used to improve convergence. The trajectories from the reverse process can be included in the forward path ensemble when their density is reweighted by $e^{\beta(W-\Delta F)}$, according to Crooks path ensemble average. Reweightening of the reverse trajectories require knowledge of ΔF (and ΔF_t for each time slice), which is estimated by BAR method. Both datasets are then used to create histograms from time slices and after applying WHAM, the final expression for PMF estimate is given below:

$$e^{-\beta\Delta G_0(z)} = \frac{\sum_t \left[\langle \frac{n_F \delta(z-z_t) e^{-\beta W_0^t}}{n_F + n_R e^{-\beta(W-\Delta F)}} \rangle_F + \langle \frac{n_R \delta(z-z_{\tau-t}) e^{-\beta W_{\tau-t}^t}}{n_F + n_R e^{\beta(W+\Delta F)}} \rangle_R \right] e^{\beta \Delta F_t)}}{\sum_t e^{-\beta[V(z,t)-\Delta F_t])}} \tag{68}$$

Figure 1.8 shows all the curves representing PMF between two methane molecules in water calculated with free energy methods listed in this section. Although steered MD is widely regarded as a least accurate method for free energy estimation, it performs quite well in the case study involving two methanes in water. Again, simplicity of the system allows production of enough trajectories in forward and reverse direction to unravel the underlying potential of mean force in a reasonable time.

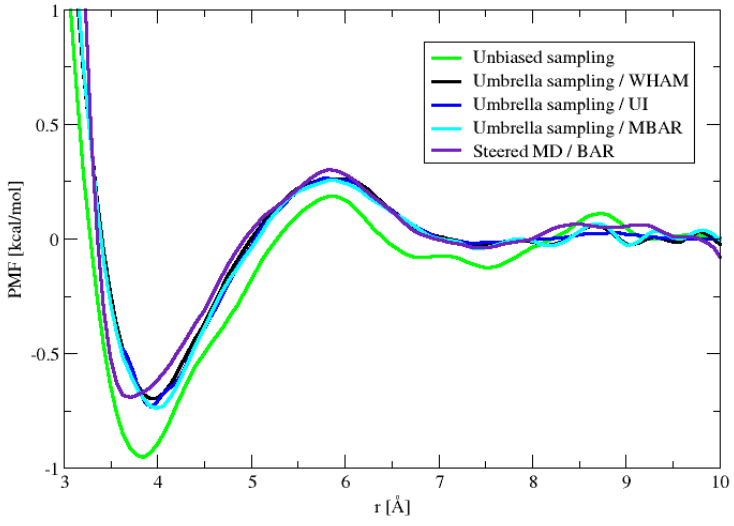


Figure 1.8 Comparison of PMF curves computed using different methods: direct counting from the unbiased simulation (green); umbrella sampling approach combined with 3 different estimators – WHAM (black), UI (blue), MBAR (cyan); and finally, PMF obtained from steered MD simulations using the estimator developed by Minh and Adib for bidirectional pulling experiments (indigo).

The main conclusion would be that a proper choice of the generalized coordinate and thorough sampling will lead to the correct answer independent of the method used to estimate PMF. However, when complex systems are explored in a similar fashion, this can become a complicated and cumbersome task.

1.3.4.3 Implicit Solvation

In the implicit solvation model, as its name suggests, the solute is not surrounded with individual solvent molecules; instead, it is immersed in a uniform dielectric medium. This simplified approach aims to capture the mean influence of the solvent molecules on the solute, that is, a statistical mechanical formulation of this concept relies on the potential of mean force (PMF) exerted on a solute.[51] Implicit models significantly reduce the computational costs compared to the explicit solvent treatment. Namely, the latter approach would require averaging over multiple solvent configurations to address the solvent effects exerted on the solute, and this is especially cumbersome task for large systems, such as solvated biomacromolecules. There are other difficulties in using explicit solvent in the molecular dynamics simulation that affect free energy calculations, such as truncation of the long range electrostatic or summing them over an infinite periodic array using Ewald techniques.[52] Therefore, it is of practical importance to develop models that incorporate the solvent effect in an implicit manner. The implicit solvation model avoids statistical errors associated with the averages obtained from the simulations with large number of explicit solvent molecules. It allows prediction of the electrostatic properties for these systems, which makes it a convenient method for calculation of the free energies of solvation or binding and other related properties, such as pK_a shifts, electrostatic potential etc. One of the fundamental approaches to calculate electrostatic interactions within the implicit solvation framework involves the solution of the Poisson-Boltzmann equation:[53]

$$-\nabla \cdot \big(\epsilon(\mathbf{r})\nabla u(\mathbf{r})\big) + \bar{\kappa}^2(\mathbf{r})\sinh\big(u(\mathbf{r})\big) = \left(\frac{4\pi e^2}{k_B T}\right)\sum_{i=1}^{N_m} z_i \delta(\mathbf{r}-\mathbf{r}_i) \qquad (69)$$

where $u(\mathbf{r})$ denotes the dimensionless electrostatic potential at a field position $\mathbf{r}$, and $\epsilon(\mathbf{r})$ is the permittivity that takes the values of the appropriate dielectric constants in the different regions of the model. For example, the value ϵ_m in the molecular region, and a second value ϵ_w in both the solution region and an ion-exclusion layer surrounding the molecule. The modified Debye-Hückel parameter $\bar{\kappa}(\mathbf{r})$ is proportional to the ionic strength of the solution and it is

dielectric independent. The molecule is represented by N_m point charges at positions $q_i = z_i e$ at positions $\mathbf{r}_i$. The constants e, k_B, T represent the electron charge, Boltzmann constant, and the absolute temperature, respectively. This is the nonlinear Poisson-Boltzmann equation, and its solution is usually approximated by solving the linearized form:

$$-\nabla \cdot \big(\epsilon(\mathbf{r})\nabla u(\mathbf{r})\big) + \bar{\kappa}^2(\mathbf{r})u(\mathbf{r}) = \left(\frac{4\pi e^2}{k_B T}\right)\sum_{i=1}^{N_m} z_i \delta(\mathbf{r} - \mathbf{r}_i) \qquad (70)$$

Analytical solutions to Poisson-Boltzmann equations are available only for a few simple cases. This has resulted in the development of various numerical methods to solve this problem. The algorithms to solve PB equation are implemented in most molecular dynamics software packages or they exist as the independent software. In this thesis, most of the PB calculations were performed with Adaptive Poisson-Boltzmann Solver (APBS).[54]

Determining the free energies within the implicit solvation framework usually requires usage of the free energy cycles. The free energy cycle constructed to calculate free energy of solvation and used by APBS is presented in Figure 1.9:[55]

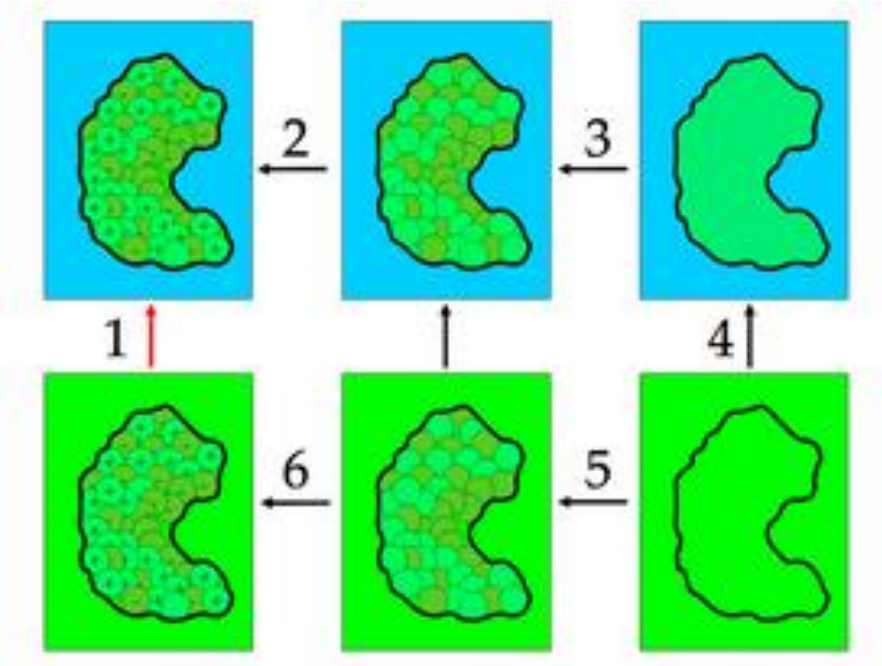


Figure 1.9 The free energy cycle used by APBS (http://www.poissonboltzmann.org/apbs/) to calculate the solvation energy (step 1). Step 2 indicates charging of the solute in solution (e.g., inhomogeneous dielectric, ions present). Steps 3 and 5 are associated with the attractive solute-solvent dispersive interactions (e.g., an integral of Weeks-Chandler-Andersen interactions over the solvent-accessible volume). Step 4 stands for the introduction of repulsive solute-solvent interaction (e.g., cavity formation). Finally, step 6 represents the charging of the solute in a vacuum or homogeneous dielectric environment in the absence of mobile ions.

The total free energy (ΔG_{tot}) can be decomposed into polar and nonpolar contributions, which are calculated separately. The polar contribution corresponds to the free energy difference of charging the solute in two different dielectric media, e.g. vacuum and the solvent:

$$\Delta G_p = \Delta G_2 - \Delta G_6$$

This step requires certain caution regarding the so called self-interaction, which the additional energy term resulting from the charge distribution interacting with itself. These self-interaction energies are typically very large and extremely sensitive to the problem discretization, such as the grid parameters used to numerically estimate electrostatic potential. Therefore, it is recommended to use the identical parameters to avoid the errors by cancelling these artefacts.

The nonpolar part consists of the energy necessary to create a cavity in the solution and the energy associated with dispersive interactions between the solute and solvent:

$$\Delta G_{np} = (\Delta G_3 - \Delta G_5) + \Delta G_4$$

A number of methods were developed to estimate nonpolar contribution to the total solvation energy, most of them based on the solvent accessible surface area (SASA).[56,57] These methods provide a crude approximation for prediction of the nonpolar free energies of solvation,[58] often resulting with a severe qualitative offset between numbers obtained with different methods.[59] However, in many cases, the electrostatic contribution to the free energy of solvation is dominant and the nonpolar part is often neglected for practical purposes. This approximation is even more justified for the free energy cycles where these terms nearly cancel out, because the nonpolar contributions for the same species in different solvent are often quite similar, as they depend mainly on the molecular geometry.

1.4 REFERENCES

1 Levine, I. N. *Quantum Chemistry* **1991**, 4th Edition, Prentice-Hall, New Jersey.

2 Leach, A. R. *Molecular Modelling; Principles and Applications* **1996**, Addison Wesley Longman Ltd., Harlow.

3 Foresman, J. B.; Frisch, Æ. *Exploring Chemistry with Electronic Structure Methods* **1996**, 2nd Edition, Gaussian Inc., Pittsburgh.

4 Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory* **1982**, Macmillan Publishing, New York.

5 Cramer, C. J. *Essentials of Computational Chemistry – Theories and Models* **2004**, Second Edition , J.Wiley & Sons Inc., New York.

6 Hartree, D. R. *Proc. Cam. Phil. Soc*. **1928**. 24, 246.

7 Fock, V. Z. *Phys*. **1930**, 61, 126.

8 Roothan, C. C. *Rev. Mod. Phys*. **1951**, 23, 69.

9 Hall, G. C. Proc. Roy. Soc. (London) **1951**, A205, 541.

10 Pople, J. A.; Nesbet, R. K. *J. Phys. Chem*. **1954**, 22.

11 Kohn, W.; Hohenberg, P. *Phys. Rev. B* **1964**, 136, 864.

12 Kohn, W.; Sham, L. *Phys. Rev. A* **1965**, 140, 1133.

13 Becke, A. D. *Phys. Rev. B* **1988**, 38, 3098.

14 Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, 37, 785.

15 Kohn, H.; Becke, A. D.; Parr, R. G. *J. Phys. Chem*. **1996**, 100, 12974.

16 Becke, A. D. *J. Chem. Phys*. **1993**, 98, 1372.

17 Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett*. **1996**, 77, 3865; **1997**, 78, 1396.

18 Burke, K. J. *Chem. Phys*. **2012**, 136, 150901.

19 Zhao, Y.; Truhlar, D. G. *Chem. Phys. Lett*. **2011**, 502(1-3), 1.

20 Grimme, S. *J. Comput. Chem*. **2006**, 27, 1787.

21 Burke, K.; Werschnik, J.; Gross, E. K. U. *J. Chem. Phys*. **2005**, 123, 062206.

22 Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys*. **1999**, 110, 4703.

23 Henry, D.J., Sullivan, M.B., Radom, L. *J. Chem. Phys*. **2003**, 118, 4849.

24 Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. (Theochem)* **1999**, 462, 1999.

20 Vreven, T., Morokuma, K., Farkas, Ö., Schlegel, H. B., Frisch, J. M. *J. Comput. Chem*. **2003**, 24, 760.

21 Vreven, T., Byun, K. S., Komaromi, I., Dapprich, S., Montgomery, J. A., Jr., Morokuma, K., Frisch, J.M. *J. Chem. Theory Comput.* **2006**, 2, 815.

27 Vreven, T.;Frisch, M. J.; Kudin, K. N.; Schlegel, H. B.; Morokuma, K. *Mol. Phys*. **2006**, 104, 701.

28 Chipot, C.; Shell, M. S.; Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology - Chapter 1* (edited by C. Chipot, A. Pohorille), Springer-Verlag, Berlin Heidelberg, **2007**

29 Chandler, D. *Introduction to Statistical Mechanics*, Oxford University Press, New York, **1987**.

30 Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem*. **2010**, 31, 1569.

31 Kirkwood, J. G. *J. Chem. Phys*. **1935**, 3, 300.

32 Darve, E. *Free Energy Calculations: Theory and Applications in Chemistry and Biology – Chapter 4*, (edited by C. Chipot, A. Pohorille), Springer-Verlag, Berlin Heidelberg, **2007**.

33 http://www.iams.sinica.edu.tw/lab/jlli/thesis_andy/thesis.html

34 Laio, A. *Advanced sampling techniques for numerical simulations: lecture notes*, **2008.**

35 Trzesniak, D.; Kunz, A.-P. E.; van Gunsteren, W. F. *ChemPhysChem* **2007**, 8, 162.

36 Ferrenberg, A. M.; Swendsen, R.H. *Phys. Rev. Lett*. **1989**, 63, 1195.

37 Kumar, S.; Bouzida, D.; Swendsen, D. H.; Kollman, P. A.; Rosenberg J. M. *J. Comp. Chem*. **1992**, 13, 1011.

38 Kästner, J.; Thiel, W. *J. Chem Phys*, **2005**, 123, 144104.

39 Kirkwood, J. G. *J. Chem. Phys*. **1935**, 3, 300.

40 Shirt, M. R.; Chodera, J. D. *J. Chem. Phys*. **2008**, 129, 124105.

41 Jarzynski, C. *Phys. Rev. Lett*. **1997**, 78, 2690.

42 Xiong, H.; Crespo, A.; Marti, M.; Estrin, D.; Roitberg, A. E. *Theor. Chem. Acc*. **2006**, 116, 338.

43 Collin, D.; Ritort , F.; Jarzynski, C.; Smith, S. B.; Tinoco Jr., I.; Bustamante, C. *Nature* **2005**, 437, 231.

44 Crooks, G. E. *Phys. Rev. E* **1999**, 60, 2721.

45 Bennett, C. H. *J. Comput. Phys*. **1976**, 22, 245.

46 Jarzynski, C. *Phys. Rev. E* **2006**, 73, 046105.

47 Hummer, G.; Szabo, A. *Proc. Natl. Am. Sci. USA* **2001**, 98, 3658.

48 Minh, D. D. L.; Chodera, J. D. *J. Chem. Phys*. **2011**, 134, 024111.

49 Minh, D. D. L.; Adib, A. B. *Phys. Rev. Lett*. **2008**, 100, 180602.

50 Minh, D. D. L.; Chodera, J. D. *J. Chem. Phys*. **2009**, 131, 134110.

51 Roux, B.; Simonson, T. *Biophys. Chem*. **1999**, 78, 1.

52 Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, Oxford Science Publications, Clarendon Press, Oxford, **1989**.

53 Holst, M. J. *The Poisson-Boltzmann Equation: Analysis and Multilevel Numerical Solution* (doctoral thesis), University of Illinois, **1994**.

54 Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. USA* **2001**, 98, 10037.

55 http://www.poissonboltzmann.org/apbs/examples/solvation-energies

56 Weiser, J.; Shenkin, P. S.; Still, W. C. *J. Comput. Chem*. **1999**, 20, 217.

57 Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. USA* **2006**, 103. 22 8331.

58 Chen, J.; Brooks III, C. L. *Phys. Chem. Chem. Phys*. **2008**, 10, 471.

59 Genheden, S.; Kongsted, J.; Söderhjelm, P.; Ryde, U. *J. Chem. Theory Comput*. **2010**, 6 (11), 3558.

2. RADICAL ENZYMES

2.1 INTRODUCTION

For many years, free radicals have been regarded as rare species in enzymatic reactions, mostly involving metalloenzymes and the semiquinone/quinine cofactors as the main radical species in the catalysis.[1] The second half of 20th century marks a significant increase in the number of discovered enzymes that either contain radicals or are able to stabilize radical intermediates.[2] Namely, radicals are high-energy species and are usually short-lived, which makes them difficult to observe and characterize. However, the advances in modern experimental techniques and spectroscopy have made it possible to detect and study these systems in more detail. Another obstacle in studying radical enzymes is their sensitivity to the presence of oxygen, a biradical species that can easily quench the radical formed in the protein environment. This may be the reason why majority of the discovered radical enzymes originates from anaerobic microorganisms, where they catalyze many key steps in their metabolism.[3] Apart from oxygen sensitivity, the radical reactivity could also lead to a number of intramolecular side reactions involving neighbouring residues or by dimerizing. These unwanted pathways are seemingly prevented by the protein architecture. The existence of highly reactive radical species opens the door to new reaction pathways in enzyme catalysis when the low energy processes are not available, but it also requires a controlled environment to avoid possible damage of the cellular structures. In general, radical chemistry is more demanding than classical acid-base catalysis because of the controlled generation, storage, and decomposition that is required for species with unpaired electrons.

The usual classification of the radical enzymes is based on how the radical is generated. The majority of enzymes create their own organic radicals, but there is also a significant number of those that rely on activating enzymes (activases), whose duty is hydrogen abstraction from an amino acid in the catalytic unit. This may result, for example, with the stable glycyl, cysteinyl, tryptophanyl or tyrosinyl radicals. Mechanisms of catalytic free radical generation include photoactivation of cofactors, interaction of molecular oxygen with non-heme binuclear iron, with heme or Cu-tyrosine centres. A very important mechanism involves the homolytic cleavage of the C–Co bond of the cofactor adenosylcobalamin (vitamin B_{12}) into cob(II)alamin and the 5'-deoxyadenosyl radical (DOA•). The same 5'-deoxyadenosyl radical is formed when S-

adenosylmethionine (SAM) undergoes fission of the S–CH2 bond after reduction of methionine by [4Fe-4S] cluster. An additional classification based on amino acid sequence comparisons is applicable to the SAM-dependent enzymes, giving rise to the now quite large SAM superfamily of radical enzymes.[4] The SAM superfamily and coenzyme B_{12}-dependent enzymes are oxygen independent, although for the latter ones there is no apparent common evolutionary origin.[5] More about adenosyl radical based chemistry can be found in a recent review written by Marsh *et al*.[6] The SAM superfamily is of special interest for this thesis and will be discussed in more detail in Section 2.2.

A very good illustration for radical initiation possibilities in the catalysis is given with the example of ribonucleotide reductase (RNR).[7] These enzymes catalyze the reduction of ribonucleotides to 2'-deoxyribonucleotides, which are basic elements necessary for DNA synthesis. RNRs are divided into three classes that correspond to three different mechanisms of generating the thiyl radical in the active site.[8] Class I RNR is present both in humans and *E. Coli* under aerobic conditions and it contains binuclear iron centre that activates oxygen and forms a tyrosyl radical. This was the first stable radical derived from an amino acid to be discovered in an enzyme.[9] In the presence of the substrate, the tyrosyl radical induces the generation of the thiyl radical over a distance of 40 Å by aligning intermediary tyrosine residues.[10] It is assumed that this long distance is a measure of protection of the thiyl radical against oxygen. Class II RNR uses adenosylcobalamin as radical generator,[11] while class III uses S-adenosylmethionine for the same purpose.[12]

In addition to the above mentioned mechanisms, another interesting radical formation in the context of this thesis takes place via photoactivation in the light-dependent DNA photolyases. Radical species in these systems are generated by excitation of the cofactor flavin adenine dinucleotide in the fully reduced state ($FADH^-$), followed by the electron transfer to the DNA lesion. Formation of the anionic radical lesion is a key step in the DNA repair mechanism catalyzed by these enzymes.[13] After the successful repair, the electron is transferred back to FAD.

Since the focus of this thesis is set on pyruvate formate-lyase, a member of SAM superfamily, and (6-4) photolyase, the following sections will be devoted to some general features of the families to which these enzymes belong.

2.2 SAM SUPERFAMILY

In 2001, a bioinformatic study was made by Sofia *et al.*,[4] which identified a superfamily of metalloenzymes that catalyze diverse reactions involved in various biological pathways, including biosynthesis of a large number of cofactors and antibiotics, the biosynthesis and repair of DNA, and general bacterial metabolism.[14] The initial study predicted around 600 enzymes belonging to this superfamily, but the number has arisen to some 3000 by this date. Many of them are still not identified and structure is solved for a limited number of SAM enzymes. However, discovery of this superfamily has revived the interest for radical enzymology and this research field has received increased attention from the scientific community in the recent years. This interest is additionally confirmed by dedicating entire issue of *Biochimica et Biophysica Acta* to radical SAM enzymes and radical enzymology (issue 1824, Nov. 2012).

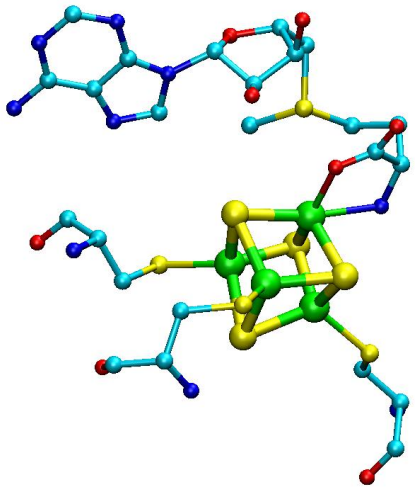


Figure 2.1 Typical coordination of [4Fe-4S] by SAM and three cysteines.

These enzymes are specific in their utilization of the S-adenosylmethionine (SAM) to generate 5'-deoxyadenosyl radical, which subsequently removes hydrogen atom from a target molecule.[15] A common feature used to identify the radical SAM superfamily is a sequence motif containing three cysteine residues, CxxxCxxC, which coordinate three iron atoms of a [4Fe–4S] cluster.[16] The presence of [4Fe–4S] cluster is essential for catalytic activity for SAM enzymes.[17] Namely, it acts as a binding site for SAM, whose methionine moiety coordinate the fourth (unique) iron in the cluster with its amino and carboxylate group (Fig. 2.1).[18, 19] The other important function of this cluster is reduction of the SAM sulfonium ion, which leads to the cleavage of the C-S bond and results with methionine and 5'-deoxyadenosyl (DOA•) radical.[20] The initial oxidation state of the [4Fe–4S] cluster is +1 and changes to +2 after the electron transfer to SAM. There are several modes by which SAM enzymes control the inherent reactivity of the reduced iron–sulfur cluster and the SAM sulfonium, and the resulting radical chemistry. They include reduction potential modifications and structural changes upon substrate modification to avoid SAM cleavage uncoupled from the catalysis and protect the radical environment.[21,22] An alternative SAM cleavage pathway was recently suggested for diphthamide biosynthetic

enzyme Dph2 and glycerol dehydratase activating enzyme (GDH-AE) which yields 5'-methylthioadenosine and a 3-amino-3-carboxypropyl radical (ACP•),but further investigation is required to support this interesting possibility. [23,24]

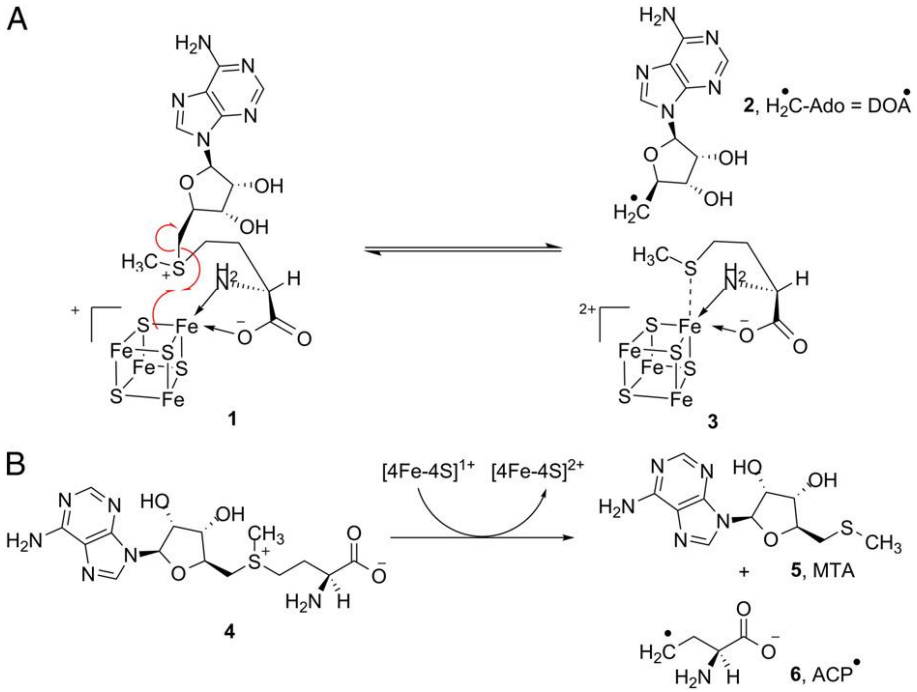


Figure 2.2. SAM cleavage pathways. The conserved $[4Fe–4S]^{+1}$ cluster provides the electron for reductive SAM cleavage resulting in methionine and a DOA• (A) or less common 5'-methylthioadenosine and an ACP• (B).[24]

Radical SAM enzymes are divided into three subclasses, based on the further fate of DOA• radical.[25] The first class share some similarities with B_{12}-dependent enzymes and involves catalytic reactions, in which SAM is reversibly cleaved to provide 5'-deoxyadenosyl radical and regenerated in each catalytic cycle. The representatives of this class are lysine 2,3-aminomutase (LAM) and spore photoproduct lyase (SPL). The second subclass uses DOA• radical as a substrate to abstract hydrogen atom from a glycyl residue to activate the substrate enzyme, known as glycyl radical enzymes (GRE). The cleavage of SAM in this case is irreversible. This subclass is called glycyl radical enzyme activases (GRE-AE), including pyruvate formate-lyase activase (PFL-AE), glycerol dehydratase activase (GDH-AE) and many others. Finally, the last subgroup also results with irreversible SAM cleavage and formed DOA• radical is used solely as a co-substrate in various reactions, in which the first step is hydrogen abstraction. The substrate derived radical can then undergo a range of reactions, such as additions, eliminations and different rearrangements.

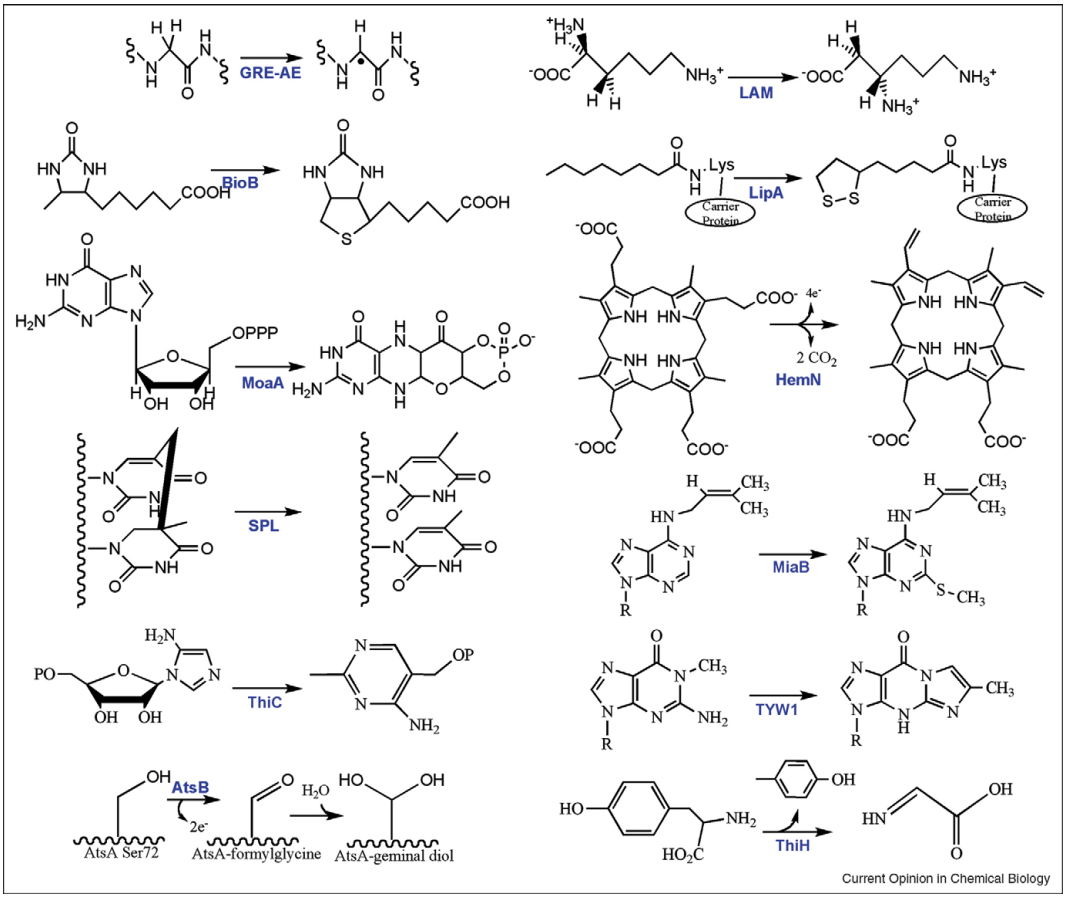


Figure 2.3. Representative reactions catalyzed by the radical SAM enzymes. Abbreviations used: GRE-AE, glycyl radical activating enzymes such as pyruvate formate-lyase activating enzyme; LAM, lysine 2,3-aminomutase; BioB, biotin synthase; LipA, lipoyl synthase; MoaA, molybdopterin cofactor biosynthesis enzyme; HemN, oxygen-independent coproporphyrinogen oxidase; SPL, spore photoproduct lyase; MiaB, tRNA methylthiolation enzyme; ThiC and ThiH, enzymes involved in thiamine biosynthesis, TYW1, tRNA modification enzyme; AtsB, formylglycine-generating enzyme.[21]

Although initially SAM was referred to as "a poor man's adenosylcobalamin",[26] a vast diversity of chemical reactions initiated by SAM in comparison to B_{12} has changed that perspective, changing the metaphor to "a rich man's adenosylcobalamin".[27] Available information about radical SAM enzymes indicates that they are of ancient origin, and were among the earliest biological catalysts to function by radical mechanisms. Many functionalities of SAM enzymes still remains unknown, although significant progress has been made in recent years based on the increasing amount of experimental and structural data.[28]

2.2.1 GLYCYL RADICAL ENZYMES

The mechanism of pyruvate formate-lyase (PFL) is a major topic of this thesis, and this enzyme belongs to the family of glycyl radical enzymes. Glycyl radical enzymes (GRE) act as substrates to the cognate activases (GRE-AE), specific enzymes containing [4Fe-4S] cluster necessary to generate radical from SAM, which in turn abstracts hydrogen from strictly conserved glycine residue in GREs.[29] In this way, the radical is stored in the form of the stable glycyl radical.[30] PFL was the first enzyme discovered to have radical located on the backbone rather than on a protein side chain.[31,32] During catalysis, the radical is transferred from glycyl to the substrate via a proximal cysteine residue in the active site. The glyclyl radical enzymes identified so far catalyze different reactions in microbes under anaerobic conditions.[33]

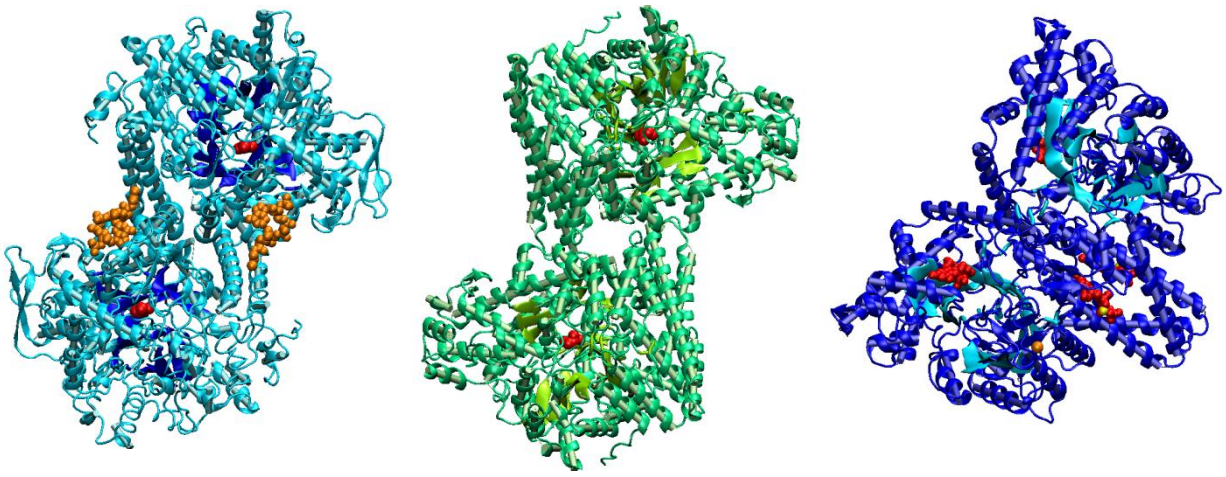


Figure 2.4 Representatives of glycyl radical enzymes with the highlighted common β-barrel motif: (a) pyruvate formate-lyase in complex with substrates pyruvate (red) and coenzyme A (orange); (b) glycerol dehydratase in complex with glycerol (red); (c) anaerobic ribonucleotide reductase in complex with deoxyguanosine triphosphate (red), Mn^{2+} (yellow) and Zn^{2+} (orange).

The majority of GREs usually appear as homodimers with subunit size of 80-100 kDa, but they can comprise additional subunits and activases.[34] Although they differ significantly in sequence, most GREs are structurally homologous, according to the available X-ray structures of the GREs (Figure 2.4). A common motif is a 10-stranded β-barrel core surrounded by α-helices in each monomer, with two finger-like loops protruding into the centre of the β-barrel.[35] One of the loops contains the radical storing glycine in a conserved hydrophobic stretch (RSVXG) close to the C-terminus of the protein,[36] while the other carries the conserved cysteine that directly participates in the catalysis.

The glycyl radical is stabilized by delocalization of the unpaired electron to the neighbouring peptide bonds of the protein backbone by means of the so-called captodative effect.[37] However, glycyl radical is very sensitive to the presence of oxygen and the enzymes undergo irreversible cleavage of the polypeptide chain at the radical site.[38] In the case of PFL, a deactivation mechanism that reversibly removes radical from the enzyme has been developed to avoid irreversible inactivation when cell shifts from anaerobic to aerobic regime.[39] Hence, glycyl radical enzymes require strictly anaerobic conditions.

The most studied representatives of GREs include pyruvate formate-lyase (PFL), anaerobic ribonucleotide reductase (ARNR), benzylsuccinate synthase (BSS) and glycerol dehydratase (GDH). PFL was the first GRE discovered and it catalyzes reversible conversion of pyruvate to acetyl-CoA and formate during anaerobic metabolism of microorganisms. It requires activation by its activase (PFL-AE), which belongs to SAM superfamily. PFL-AE uses SAM to generate 5'-deoxyadenosyl radical which in turn abstracts hydrogen from the conserved glycine residue in PFL:

$$\left[\text{-NH-CH}_2\text{-CO-}\right] + \text{SAM} + e^- \longrightarrow \left[\text{-NH-}\dot{\text{C}}\text{H-CO-}\right] + \text{5'-deoxyadenosyl} + \text{methionine}$$

Upon activation, the radical moves from glycine to the conserved cysteine. There is an additional cysteine residue in the active site to which radical is transferred and it attacks carbonyl group of pyruvate. This attack results with C-C bond cleavage and formation of the acetylated enzyme and formate, followed by transfer of the acetyl group from the enzyme to coenzyme A (CoA):

$$\text{PFL} + \text{pyruvate} \rightleftharpoons \text{acetyl-PFL} + \text{formate}$$
$$\text{acetyl-PFL} + \text{CoA} \rightleftharpoons \text{PFL} + \text{acetyl-CoA}$$

The existence of another cysteine residue in the active site is specific for PFL, as is the radical addition to substrate. Namely, GREs mostly rely on hydrogen abstraction as a mean of substrate activation.

Hydrogen abstraction is mechanistic step used in the conversion of ribonucleotides to deoxyribonucleotides, catalyzed by anaerobic ribonucleotide reductase (ARNR or RNR III).[40] Activated ARNR has a radical stored on glycine and then transferred to cysteine, which abstracts 3'-H atom from the ribonucleotide substrate. Elimination of 2'-OH group leads to formation of keto-radical, subsequently reduced with formate. The final product is deoxyribonucleotides, a DNA building block, presenting RNRs as possible links between the

RNA and DNA worlds.[41] ARNR is an example of GRE which has a dimeric activase as an integral part of the entire enzyme complex. [42]

Hydrogen abstraction by cysteine residue is also used in the reaction of radical addition of toluene to fumarate, catalyzed by benzylsuccinate synthase (BSS). Formation of benzylsuccinate is the first step in the fermentation of toluene by various sulphate and nitrate-reducing bacteria.[43]

Glycerol dehydratase (GDH) is glycyl radical enzyme that catalyzes the conversion of glycerol to 3-hydroxypropanal,[44] but it is interesting to mention in the context of radical enzyme chemistry that there is B_{12}-dependent enzyme catalyzing this same reaction.[45] Another interesting fact is related to its activase (GDH-AE), which seems to use unusual pathway for SAM cleavage yielding 5'-methylthioadenosine and a 3-amino-3-carboxypropyl radical. The latter abstracts hydrogen from the conserved glycine.

2.2.2 Spore Photoproduct Lyase

Bacterial spores are one of the most resistant and longest-lived cells, developed to survive in the extreme conditions. These extreme conditions include a strong UV irradiation, which can cause damage in DNA molecules, especially in the case of two neighbouring pyrimidine bases. In the B-form of DNA, two adjacent pyrimidines can dimerize upon exposure to UV light, leading to the formation of cyclobutane pyrimidine dimer (CPD) and (6-4) photolesion. However, DNA in the spores is transformed to the A-form[46] and dimerization of two pyrimidine bases results with a particular lesion characteristic for spores: 5-thyminyl-5,6-dihydrothymine or spore photoproduct (SP).[47] These lesions accumulate in the dormant spores and the SP content can go up to 8% of the total thymine in genomic DNA. Such a high content of damaged DNA is fatal for germinated bacteria. Hence, it is of vital importance to repair the DNA before germination. [48] A general DNA repair pathway found in prokaryotes and eukaryotes, although at different level of complexity, is nucleotide excision repair (NER). However, there is alternative pathway specific for bacteria that rely on the in-situ repair done by spore photoproduct lyase (SPL).[49] This enzyme repairs the SP lesion exclusively, while the CPD and (6-4) lesion are repaired by DNA photolyases (see Section 2.3).

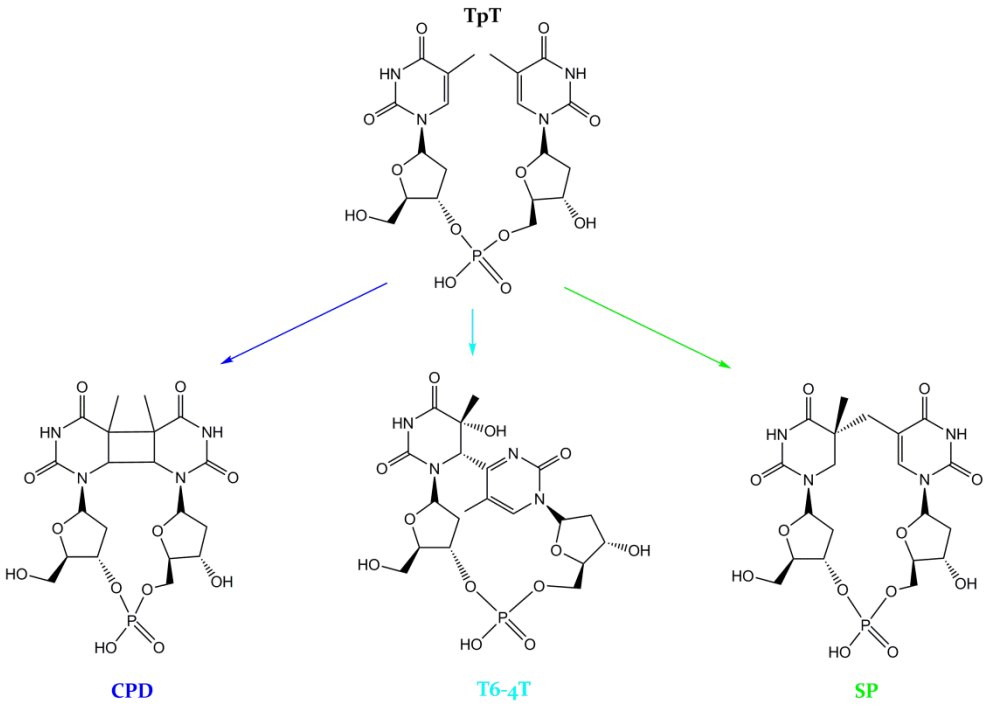


Figure 2.5 Two adjacent pyrimidine bases can dimerize upon exposure to UV radiation resulting with the formation of DNA lesions: cyclobutane pyrimidine dimer (CPD), pyrimidine-pyrimidone dimer (6-4 lesion), and spore photoproduct (SP).

SPL belongs to SAM superfamily and contains the characteristic sequence motif CxxxCxxC required for coordination of the [4Fe-4S] cluster.[50] This cluster again is necessary for reductive cleavage of SAM, which binds to the unique iron atom of the cluster in a bidentate fashion. SAM cleavage results with a formation of the reactive 5'-deoxyadenosyl radical (5'-dAdo•) that initiates the repair process. The experiments have shown that SPL uses SAM as a cofactor and it is regenerated in every catalytic cycle, similar to the behaviour observed for lysine 2,3-aminomutase.[51] These two enzymes belong to the first class of SAM superfamily enzymes.

The SPL and its mechanism has been a mystery for a long time, but it has been extensively studied in the last decade and even the crystal structure was solved recently.[52] The SPL appears as a monomer consisting of 340 amino acids. It exhibits the common fold of radical SAM enzymes and forms a partial $(\alpha/\beta)_6$ triose phosphate isomerase (TIM) barrel, which enables SPL to accept large substrates. At the top of TIM barrel is located a binding site for the single [4Fe-4S] cluster, buried inside of the enzyme. Binding site for the DNA containing the SP lesion contains a region rich in lysine and arginine residues positioned in way that enables interaction with negatively charged phosphates. The lesion is flipped out from the double helix to reach the active site.

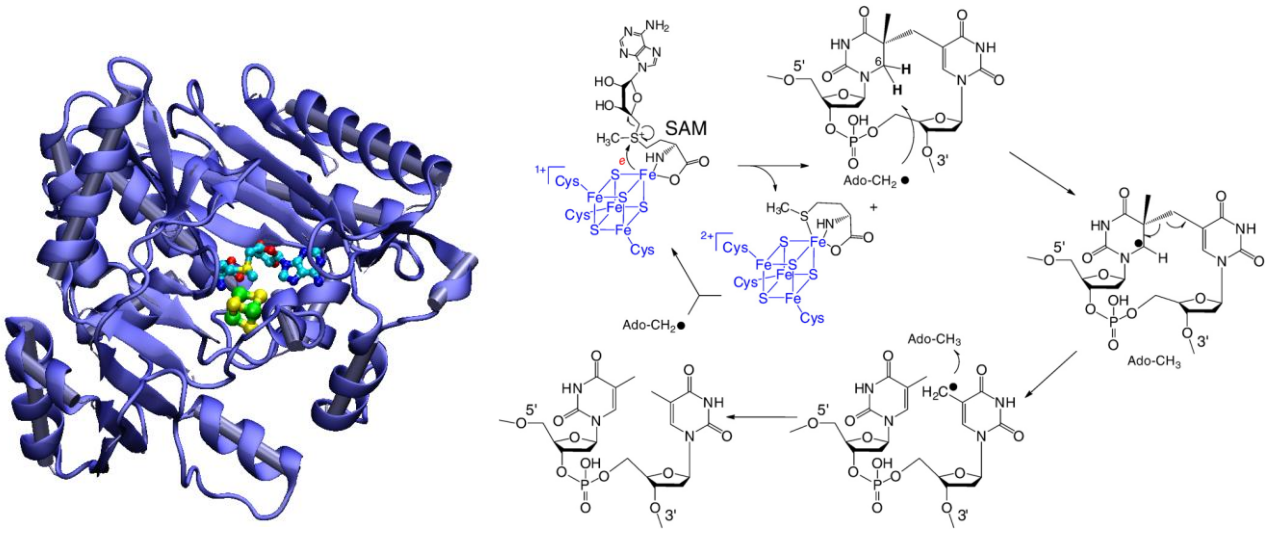


Figure 2.6 A crystal structure of spore photoproduct lyase with [4Fe-4S] centre and SAM (PDB entry: 4FHC). The currently accepted mechanism is shown on the right, but recent studies suggest that the hydrogen in the final step is donated by a protein residue, possibly Cys141, instead of SAM.[52]

Despite the increase of the biochemical and structural data, the exact mechanism of DNA repair by SPL is not yet fully explained. The latest studies indicate that upon reductive cleavage of SAM, the resulting DOA• radical abstracts hydrogen from the substrate to yield a radical on C6 atom.[53] The methylene bridge in this radical species then undergoes a homolytic cleavage to produce a thymine methyl radical. The initial assumption was that thymine allyl radical abstracts hydrogen from 5'-dAdo, but experiments do not support this theory. As an alternative, it was suggested that hydrogen is taken from a protein residue, supposedly from the conserved Cys141 (residue numbering according to SPL extracted from *B. subtilis*).[54] This assumption was further verified in the mutation experiments in which Cys141 was replaced with alanine and this mutant is characterized with a significant drop in the activity compared to the wild type.[55] However, certain aspects of SPL mechanism are still a puzzle, especially the SAM regeneration step, and clarification of the mechanistic details of this interesting enzyme is ongoing work.

2.3 DNA PHOTOLYASES

Exposure to UV radiation (280-320 nm) can cause damage in DNA molecules by inducing dimerization of two adjacent pyrimidine bases. The most common DNA lesions are cyclobutane pyrimidine dimer (CPD) and pyrimidine-pyrimidone photoproduct, called (6-4) lesion. CPD lesion is formed by [2+2] cycloaddition of the C5=C6 double bonds of the neighbouring pyrimidines, while (6-4) lesion is product of Paterno-Büchi reaction between C5=C6 and C4=O/N bonds. The (6-4) lesion can also form Dewar isomers. Successful repair of these lesions is essential for cell survival because of their high mutagenic and carcinogenic potential. The enzymes capable of repairing DNA and restoring the initial monomers by photoreactivation are known as DNA photolyases. These enzymes share significant similarity in sequence (15-70%) and some common mechanistic features; however, each enzyme repairs only one lesion. They are usually referred to as CPD photolyase and (6-4) photolyase, which repair CPD and (6-4) lesion, respectively. These light-dependent enzymes are absent in many species, including placental mammals. DNA photolyases exhibit great sequence homology to another interesting protein found in plants, animals, and some bacteria, known as cryptochromes. These enzymes show no photolyase activity, but instead they regulate some of the blue-light responses in plants such as growth and development and synchronize circadian rhythms with the daily light-dark cycles in animals.[56]

DNA photolyases are monomeric proteins consisting of 450-550 amino acids and two noncovalently bound chromophores, one of which is always flavin adenine dinucleotide (FAD). FAD is a mandatory cofactor for successful catalysis, as it both affects specific binding of DNA and because it directly participates in the repair. [57-59] It has to be in the fully reduced anionic state ($FADH^-$) to be catalytically active.[60,61] The presence of the other cofactor is not obligatory for catalysis and it serves as a light harvester. Usually this cofactor is methylentetrahydrofolate (MTHF), or less often 8-hydroxy-7,8-dimethyl-5-deazariboflavin (8-HDF). The second cofactor increases the repair rate 10-100-fold under conditions of limiting light due to a higher extinction coefficient and an absorption maximum at longer wavelengths in comparison to FAD. FAD is bound to the photolyase in an unusual U-shaped form with the isoalloxazine and adenine rings in close proximity. Its binding site involves 14 amino acids, most of which are conserved in the photolyase/cryptochrome family.[62] If the enzyme is found in an inactive state with FAD in the form of neutral semiquinone radical ($FADH^•$) or fully oxidized (FAD), photolyases can photoreduce the cofactor to $FADH^-$ state via reversible electron transfer (ET) from certain amino acid residues. In *E. coli* CPD photolyase, this process involves a chain of

tryptophan residues (Trp382, 359, 306),[63] while in (6-4) photolyase a tyrosine residue was identified as a final electron donor.[64]

Photolyase is a structure-specific DNA binding protein whose specificity is determined by the backbone structure of DNA at the binding site in contrast to the sequence-specific DNA binding proteins which rely on hydrogen-bond donors and acceptors in the grooves of the duplex.[65] The surface of the binding pocket of DNA is conveniently enriched with positively charged residues to strengthen interaction with the negative phosphate backbone of DNA.[66] DNA binding includes flipping of the dimerized bases into the active site to make stable enzyme-substrate complex.[67, 68]

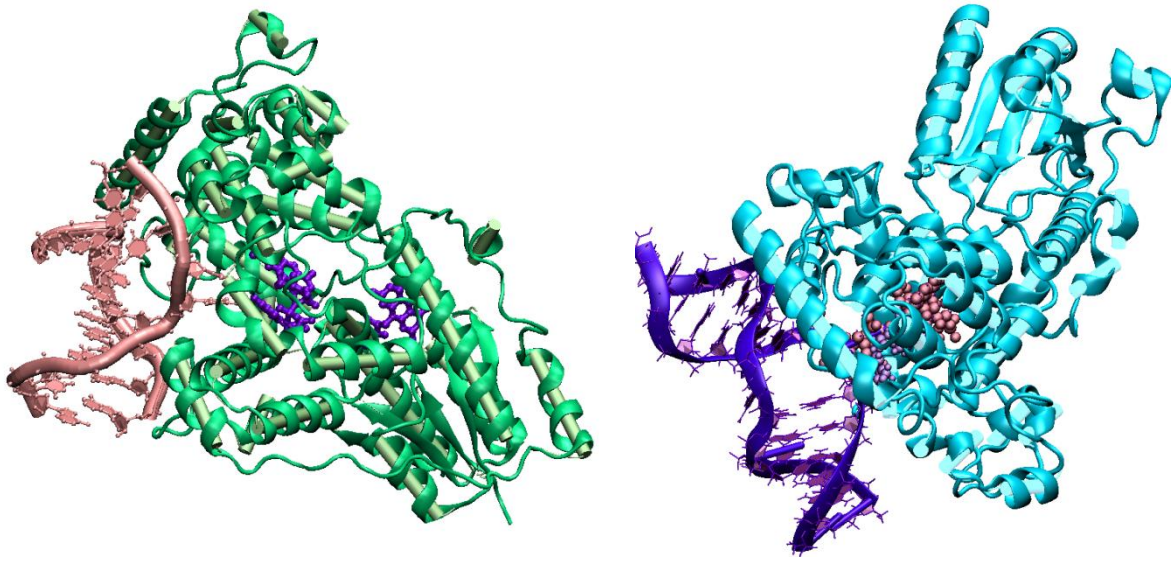
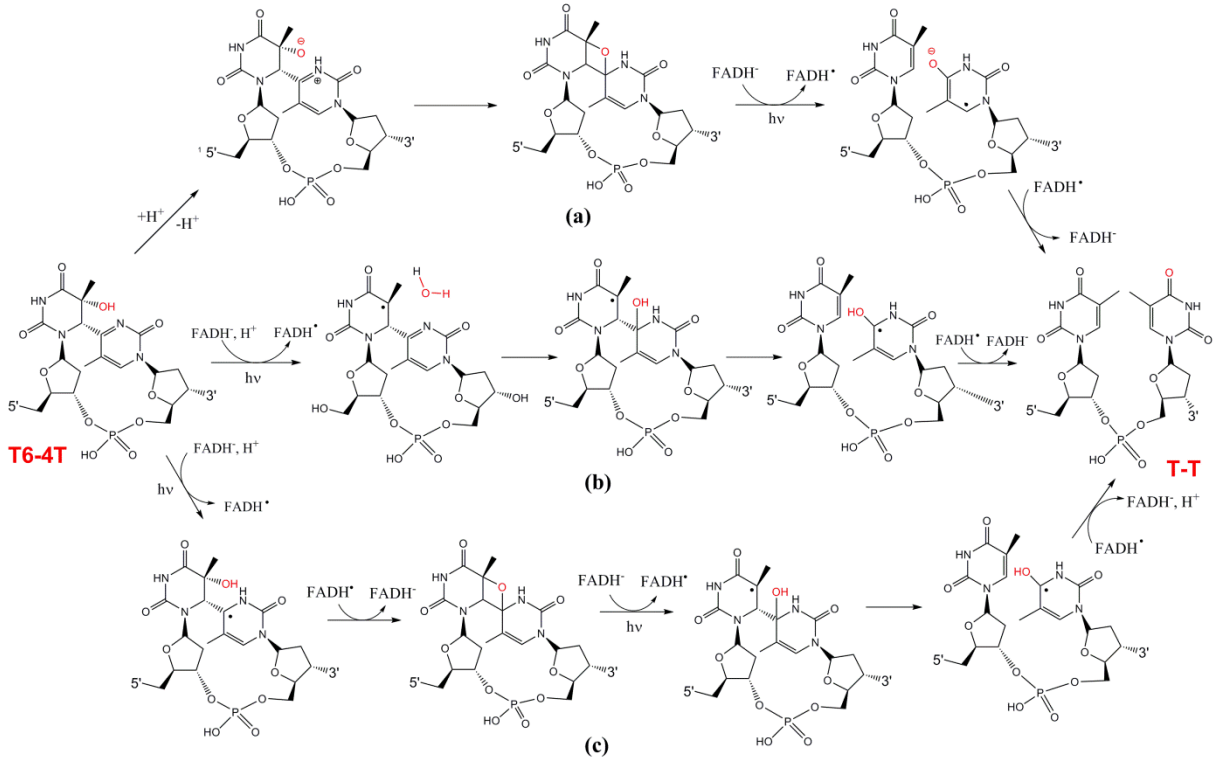


Figure 2.7 Crystal structures of CPD photolyase (left) and (6-4) photolyase (right) with the DNA lesions flipped into the active site and FAD cofactors.

Catalysis is initiated by the absorption of the near-UV/blue light either directly by the reduced $FADH^-$ or by the system photoantenna, followed by excitation energy transfer to the flavin via dipole-dipole interaction. The next step includes electron transfer from the excited $FADH^{*-}$ to the lesion, which turns into radical anion, while $FADH^\bullet$ remains in the form of stable neutral semiquinone radical. In the case of CPD lesion, a ketyl radical is formed at C4 next to a cyclobutane ring.[69] The ketyl radical acts as a nucleophile, which leads to bond scission in the cyclobutane and formation of a resonance-stabilized ketyl radical and a restored thymine base.[70] To fully restore the other monomer, the electron is back-transferred to $FADH^\bullet$ and CPD photolyase is again ready for new catalytic cycle. CPD photolyase has been studied in great detail and the currently accepted mechanism has been verified both experimentally[71] and theoretically.[72,73]

However, this is not the case with (6-4) photolyase, whose crystal structure was only recently solved for *D. melanogaster*. There are several suggested mechanisms, but none of them is yet widely accepted by the scientific community (Scheme 2-1). (6-4) photolyase is also capable of repairing Dewar isomers of (6-4) lesions.[73,74] One of the early and long-living proposals included the reversal of the formation of (6-4) lesion, which proceeds via oxetane (or azetidine) intermediate.[75] It was suggested that this intermediate is formed in the „dark" step, catalyzed by these two histidines which acted as a general acid-base pair. This intermediate would then undergo reaction similar to that of the CPD lesion; after the electron transfer from FADH•⁻ it would cleave back to the initial monomers.



Scheme 2-1 (6-4) photolyase repair mechanisms proposed by (a) Hitomi *et al*;[11] (b) Maul *et al.*;[16] (c) Sadeghian *et al.*[19] Protons are donated or accepted by the active site histidines, but the residues are omitted for clarity.

However, after the crystal structure was solved, that mechanism was discarded as unlikely. Several alternative nonoxetane mechanisms were suggested since, presented in Scheme 2-1 and most of them involve acid-base chemistry provided by the active-site histidines. In the mechanism proposed by Maul *et al.* (Scheme 2-1b), one of the histidines protonates the migrating hydroxyl group, making it a better leaving group after electron injection from FADH•⁻. The water molecule subsequently attacks the acylimine to form a radical intermediate, followed by rapid dissociation into the original pyrimidine bases. The catalytic cycle is closed by the electron transfer back to FADH• coupled with a loss of the proton. An

interesting alternative was provided in the computational study of the radical anionic T6-4T lesion carried out by Domratcheva *et al*., where a direct hydroxyde transfer via non-adiabatic pathway was suggested.[76] Another computational study performed by Sadeghian *et al*. (Scheme 2-1c) suggests that (6-4) catalysis is actually a two-photon process, where the energy of the first photon is used to form the oxetane intermediate, catalyted by protonated histidine, while the second one is required for cleavage of that cyclic intermediate.[77] However, none of these mechanisms is able to provide a model entirely consistent with the experimental observations and still leaving some open questions. One of those questions is the assignment of the correct protonation states to the active-site histidines, which are of crucial importance for the catalysis, and this is the problem tackled in this thesis.

2.4 References

1 Frey, P. A. Annu. *Rev. Biochem*. **2001**. 70, 121.

2 Buckel, W. *Angew. Chem. Int. Ed*. **2009**, 48, 6779 .

3 Buckel, W.; Golding, B. T. *Annu. Rev. Microbiol*. **2006**, 60, 27.

4 Sofia, H.J.; Chen, G.; Hetzler, B.G.; Reyes-Spindola, J.F.; Miller, N.E. *Nucleic Acids Res*. **2001**, 29, 1097.

5 Banerjee, R.; Ragsdale, S. W. *Annu. Rev. Biochem*. **2003**, 72, 209.

6 Marsh, E. N. G.; Patterson, D. P.; Li, L. *ChemBioChem* **2010**, 11, 604.

7 Stubbe, J. *Proc. Natl. Acad. Sci. USA* **1998**, 95, 2723.

8 Licht, S.; Gerfen,G. J.; Stubbe, J. *Science* **1996**, 271, 477.

9 Atkin, C. L.; Thelander, L.; Reichard, P.; Lang, G. *J. Biol. Chem*. **1973**, 248, 7464.

10 Seyedsayamdost, M. R.; Xie, J.; Chan, C. T.; Schultz, P. G.; Stubbe, J. *J. Am. Chem. Soc*. **2007**, 129, 15060.

11 Hogenkamp, H. P. *Pharmacol. Ther*. **1983**, 23, 393.

12 Eliasson, R.; Fontecave, M.; Jornvall, H.; Krook, M.; Pontis, E.; Reichard, P. *Proc. Natl. Acad. Sci. USA* **1990**, 87, 3314.

13 Sancar, A. *Chem. Rev*. **2003**, 103 (6), 2203.

14 Frey, P. A.; Hegeman, A. D.; Ruzicka, F. J. *Crit. Rev. Biochem. Mol. Biol*. **2008**, 43, 63.

15 Wang, S. C.; Frey, P. A. *Trends Biochem. Sci*. **2007**, 32, 101.

16 Layer, G.; Heinz, D. W.; Jahn, D.; Schuber, W.-D. *Curr. Opin. Chem. Biol*. **2004**, 8, 468.

17 Hiscox, M. J.; Driesner, R. C.; Roach, P. L. *Biochem. Biophys. Acta* **2012**, 1824 (11), 1165.

18 Krebs, C.; Broderick, W. E.; Henshaw, T. F.; Broderick, J. B.; Huynh, B. H. *J. Am. Chem. Soc*. **2002**, 124, 912.

19 Walsby, C. J.; Ortillo, D.; Yang, J.; Nnyepi, M. R.; Broderick, W. E.; Hoffman, B. M.; Broderick, J. B. *Inorg. Chem*. **2005**, 44, 727.

20 Cosper, N. J.; Booker, S. J.; Ruzicka, F.; Frey, P. A.; Scott, R. A. *Biochemistry* **2000**, 29, 15668.

21 Duschene, K. S.; Veneziano, S. E.; Silver, S. C.; Broderick, J. B. *Curr. Opin. Chem. Biol*. **2009**, 13, 74.

22 Wang, S. C.; Frey, P. A. *Biochemistry* **2007**, 46, 12889.

23 Zhang, Y.; Zhu, X.; Torelli, A. T.; Lee, M.; Dzikovski, B.; Koralewski, R.M.; Wang, E.; Freed, J.; Krebs, C.; Ealick, S. E.; Lin, H. *Nature* **2010**, 465, 891.

24 Demick, J. M.; Lanzilotta, W. N. *Biochemistry* **2011**, 50, 440.

25 Booker, S. J. *Curr. Opin. Chem. Biol*. **2009**, 13(1), 58.

26 Frey, P. A. *FASEB J*. **1993**, 7, 662.

27 Frey, P. A.; Magnusson, O. T. *Chem. Rev*. **2003**, 103, 2129.

28 Dowling, D.P.; Vey, J.L.; Croft, A.K.; Drennan, C. L. *Biochim. Biophys. Acta*. **2012**, 1824(11), 1178.

29 Eklund, H.; Fontecave, M. *Structure*. **1999**, 7(11), R257.

30 Hioe, J.; Savasci, G.; Brand, H.; Zipse, H. *Chem. Eur. J.* **2011**, 17(13), 3781.

31 Knappe, J., Neugebauer, F.A., Blaschkowski, H.P., Gänzler, M. *Proc. Natl. Acad. Sci. USA* **1984**, 81, 1332.

32 Wagner, A.F.V., Frey, M., Neugebauer, F.A, Schäfer, W., Knappe, J. *Proc. Natl. Acad. Sci. USA* **1992**, 89, 996.

33 Selmer, T.; Pierik, A. J.; Heider, J. *Biol. Chem.* **2005**, 386, 981.

34 Lehtiö, L.; Goldman, A. *Protein Eng. Des. Sel.* **2004**, 17 (6), 545.

35 Leppänen, V. M.; Merckel, M. C.; Ollis, D. L.; Wong, K. K.; Kozarich, J. W.; Goldman, A. *Structure* **1999**, 7, 733.

36 Sawers, G. *FEMS Microbio. Rev*. **1999**, 22, 543.

37 Himo, F. *Chem. Phys. Lett*. **2000**, 328, 270.

38 Wagner, A. F. V.; Frey, M.; Neugebauer, F. A; Schäfer, W.; Knappe, J. *Proc. Natl. Acad. Sci. USA* **1992**, 89, 996.

39 Nnyepi, M. R.; Peng, Y.; Broderick, J. B. *Arch. Biochem. Biophys*. **2007**, 459, 1.

40 Sun, X.; Ollagnier, S.; Schmidt, P. P.; Atta, M.; Mulliez, E.; Lepape, L.; Eliasson, R.; Graslund, A.; Fontecave, M.; Reichard, P.; Sjoberg, B. M. *J. Biol. Chem*. **1996**, 271, 6827.

41 Stubbe, J. *Curr. Opin. Struc. Biol*. **2000**, 10(6), 731.

42 Ollagnier, S., Mulliez, E., Schmidt, P.P., Eliasson, R., Gaillard, J., Deronzier, C., Bergman, T., Gräslund, A., Reichard, P. and Fontecave, M. *J. Biol. Chem*. **1997**, 272, 24216.

43 Leuthner, B., Leutwein, C., Schulz, H., Hörth, P., Haehnel, W., Schilz, E., Schägger, H., Heider, J. *Mol. Microbiol*. **1998**, 28, 615.

44 O'Brien, J. R.; Raynaud, C.; Croux, C.; Girbal, L.; Soucaille, P.; Lanzilotta, W. N. *Biochemistry* **2004**, 43, 4635.

45 Toraya, T. *Cell. Mol. Life Sci*. **2000**, 57(1), 106.

46 Mohr, S. C.; Sokolov, N. V.; He, C. M.; Setlow, P. *Proc. Nat. Acad. Sci. USA* **1991**, 88, 77.

47 Nicholson, W. L.; Setlow, B.; Setlow, P. *Proc. Nat. Acad. Sci USA*. **1991**, 88, 8288.

48 Setlow, P. *Annu. Rev. Microbiol*. **1995**, 49, 29.

49 Munankata, N.; Rupert, C. S. *J. Bacteriol*. **1972**, 111, 192.

50 Rebeil, R.; Sun, Y.; Chooback, L.; Pedraza-Reyes, M.; Kinsland, C.; Begley, T. P.; Nicholson, W. L. *J. Bacteriol*. **1998**, 180(18), 4879.

51 Cheek, J.; Broderick, J. *J. Am. Chem. Soc*. **2002**, 124, 2860.

52 Benjdia, A.; Heil, K.; Barends, T.R.; Carell, T.; Schlichting, I. *Nucleic Acids Res*. **2012**, 40(18), 9308.

53 Yang, L.; Lin, G.; Liu, D.; Dria, K. J.; Telser, J.; Li, L. *J. Am. Chem. Soc*. **2011**, 133, 10434.

54 Li, L. *Biochim. Biophys. Acta* **2012**, 824(11), 1264.

55 Yang, L; Lin, G.; Nelson, R. S.; Jian, Y.; Telser, J.; Li, L. *Biochemistry* **2012**, 51(36), 7173.

56 Sancar, A. *Chem. Rev*. **2003**, 103(6), 2203.

57 Sancar, G. B.; Smith, F. W.; Reid, R.; Payne, G.; Levy, M.; Sancar, A. *J. Biol. Chem*. **1987**, 262, 478.

58 Jorns, M. S.; Baldwin, E. T.; Sancar, G. B.; Sancar, A. *J. Biol. Chem*. **1987**, 262, 486.

59 Payne, G.; Sancar, A. *Biochemistry* **1990**, 29, 7715.

60 Payne, G.; Heelis, P. F.; Rohrs, B. R.; Sancar, A. *Biochemistry* **1987**, 26, 7121.

61 Cichon, M. K.; Arnold, S.; Carell, T. *Angew. Chem. Int. Ed*. **2002**, 41, 767.

62 Komori, H.; Masui, R.; Kuramitsu, S.; Yokoyama, S.; Shibata, T.; Inoue, Y.; Miki, K. *Proc. Natl. Acad. Sci. USA*. **2001**, 98, 13560.

63 Park, H.-W.; Kim, S.-T.; Sancar, A.; Deisenhofer, J. *Science* 2005, 268, 1866.

64 Weber, S.; Kay, C. W. M.; Mögling, H.; Möbius, K.; Hitomi, K.; Todo, T. *Proc. Natl. Acad. Sci. USA* **2002**, 99 (3), 1319.

65 Sancar, A. *Biochemistry* **1994**, 33 (1), 2.

66 Deisenhofer, J. *Mutat. Res*. **2000**, 460, 143.

67 Mees, A.; Klar, T.; Gnau, P.; Hennecke, U.; Eker, A. P.; Carell, T.; Essen, L. O. *Science* **2004**, 306, 1789.

68 Maul, M. J.; Barends, T. R. M.; Glas, F. A.; Cryle, M. J.; Domratcheva, T.; Schneider, S.; Schlichting, I.; Carell, T. *Angew. Chem. Int. Ed*. **2008**, 47, 10076.

69 Seyedsayamdost, M. R.; Xie, J.; Chan, C. T.; Schultz, P. G.; Stubbe, J. *J. Am. Chem. Soc*. **2007**, 129, 15060.

70 Jiang, W.; Yun, D.; Saleh, L.; Bollinger Jr., J. M.; Krebs, C. *Biochemistry* **2008**, 47, 13736.

71 Chang, C.-W.; Guo, L.; Kao, Y.-T.; Li, J.; Tan, C.; Li, T.; Saxena, C.; Liu, Z.; Wang, L.; Sancar, A.; Zhong, D. *Proc. Natl. Acad. Sci. USA* **2010**, 107, 2914.

72 Harrison, C. B.; O'Neill, L. L.; Wiest, O. *J. Phys. Chem*. **2005**, 109,7001.

73 Glas, A. F.; Kaya, E.; Schneider, S.; Heil, K.; Fazio, D.; Maul, M. J.; Carell, T. *J. Am. Chem. Soc*. **2010** 132 (10), 3254.

73 Masson, F.; Laino, T.; Röthlisberger, U.; Hutter, J. *ChemPhysChem* **2009**, 10, 400.

74 Glas, A. F.; Schneider, S.; Maul, M. J; Hennecke, U.; Carell, T. *Chem. Eur. J*. **2009**, 15(40),10387.

75 Hitomi, K.; Nakamura, H.; Kim, S.-T.; Mizukoshi, T.; Ishikawa, T.; Iwai, S.; Todo, T. *J. Biol. Chem*. **2001**, 13, 10103.

76 Domratcheva, T.; Schlichting, I. *J. Am. Chem. Soc*. **2009**, 131, 17793.

77 Sadeghian, K.; Bocola, M.; Merz, T.; Schütz, M. *J. Am. Chem. Soc*. **2010**, 132, 16285.

3. Pyruvate Formate-Lyase

3.1 Introduction

Pyruvate formate-lyase (PFL) is a key enzyme in the anaerobic glucose metabolism of *E. coli* and other microorganisms, where it catalyzes reversible cleavage of pyruvate to formate and acetyl-coenzyme A in a so called *ping-pong* reaction:[1]

$$\text{PFL} + \text{pyruvate} \rightleftharpoons \text{acetyl-PFL} + \text{formate}$$

$$\text{acetyl-PFL} + \text{CoA} \rightleftharpoons \text{PFL} + \text{acetyl-CoA}$$

The generation of acetyl-CoA and CO_2 from pyruvate under aerobic conditions is catalyzed by pyruvate dehydrogenase complex with the participation of thiamine diphosphate as the common coenzyme. Another system that can be used to generate acetyl CoA is pyruvate: ferredoxin/flavodoxin oxidoreductase. PFL is expressed under both aerobic and anaerobic conditions in facultative anaerobes, such as *E.coli*, but its expression increases significantly in a low oxygen environment.[2] Namely, PFL is very sensitive to oxygen because of the radical chemistry used in its catalysis. PFL is the first discovered glycyl radical enzymes (see Chapter 2), a class of enzymes that store radical on Cα atom of glycine in the polypeptide chain.[3] The glycyl radical in proteins exhibit great stability due to delocalization of the unpaired electron to the neighbouring peptide bonds of the protein backbone. The glycyl radical in *E. coli* PFL can survive for several days at 273 K and a few hours at 303 K in strictly anaerobic environment. The expression of PFL and its activating enzyme is tightly coupled to the level of oxygen and the subtleties of this fascinating control mechanism are still under investigation.[4]

3.1.1 Structure and Activation of PFL

PFL is a homodimer comprised of two identical subunits, each having 759 residues in their protein chain (85 kDa).[5] The crystal structure of the inactive form of PFL is available,[6] but also as a complex with its substrate pyruvate and the inhibitor oxamate in the presence of the coenzyme A.[7] Substrate binding is achieved through the formation of the salt bridge between

the pyruvate's carboxylate group and Arg435 and with hydrogen bonding of pyruvate carbonyl with Arg 176. In addition to these electrostatic interactions, this almost planar substrate finds itself in a "sandwich" between to hydrophobic residues; Trp333 and Phe432 (Figure 3.1c).[7]

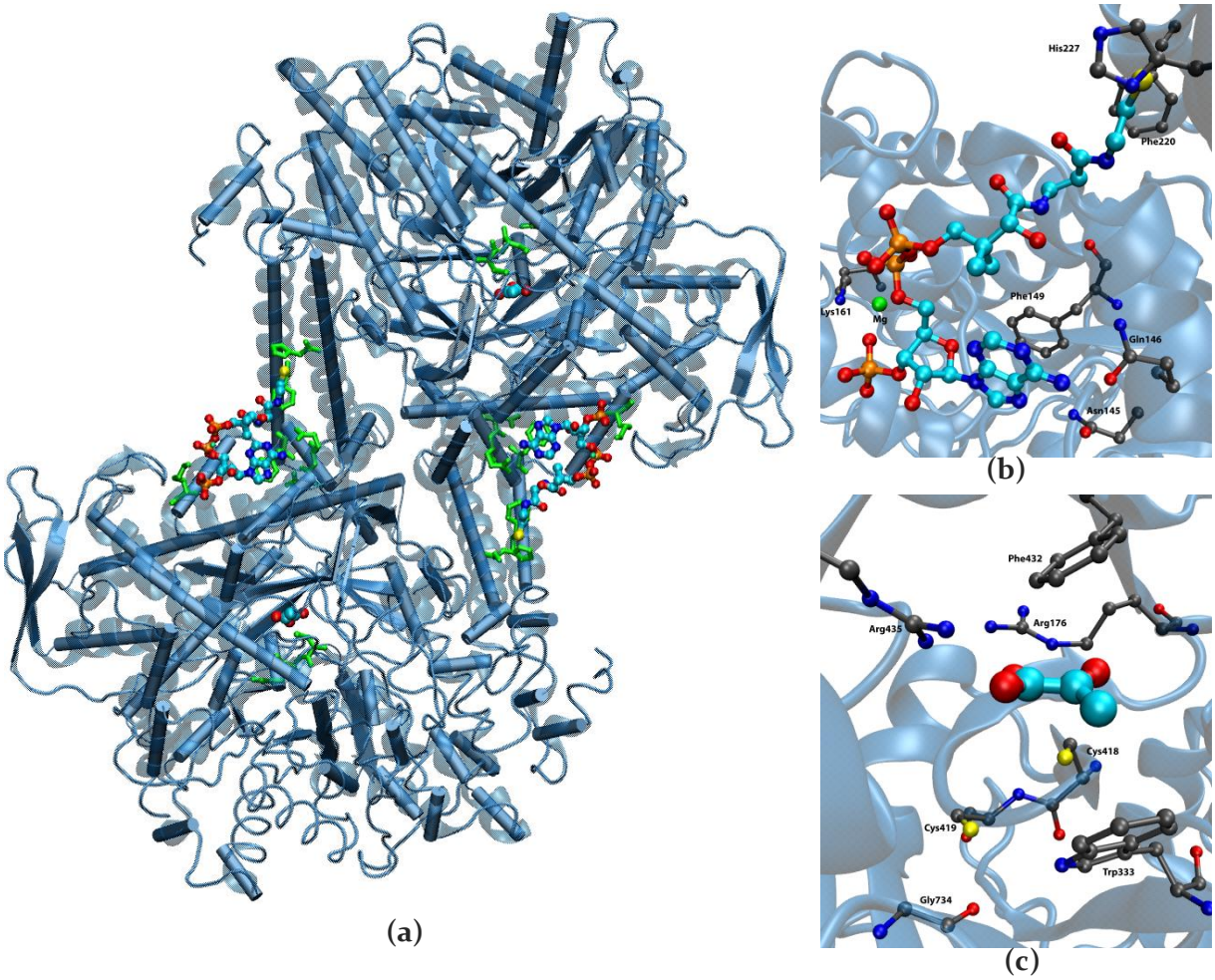


Figure 3.1 The crystal structure of PFL (a) with a closer view of the binding sites of coenzyme A (b) and pyruvate (c).

The binding site of the other substrate, coenzyme A, has been located close to the interface between two subunits in a dimer, and one CoA binds to the surface of each subunit (Figure 3.1b). The binding site comprises of a short α-helix that includes strictly conserved Asn145, Gln146 and Phe149. The latter one makes stacking interactions with imidazole ring of adenine moiety, while Asn145 and Gln 146 further strengthen binding through the formation of hydrogen bonds with the adenine amino group. Additional interaction between CoA and PFL is provided by a salt bridge between the 3' and 5' phosphates and Lys161. According to the crystallographic data, there is a density peak near the phosphate, which was interpreted as Mg^{2+}. Bound CoA adopts unusual *syn* conformation in respect to the N-glycosidic bond, although the *anti* is the preferred conformation of the free CoA in solution. In this *syn*

conformation, the thiol group on pantothenate chain is located at the predominantly hydrophobic pocket formed by the side chains of residues Phe200, His227, Leu197, and Ala224 of the opposing monomer. This form of binding places CoA 30 Å away from both active sites. According to biochemical data, binding of oxamate has no effect on CoA affinity for PFL. It also seems that CoA is not required during the first half-reaction of pyruvate cleavage, but if it is present, it has a spectator role. This implies that certain conformational changes are necessary to allow CoA to reach the active site and undergo acetylation. It has been suggested that ribose and pantothenate moiety might rotate around N-glycosidic bond and change from *syn* to *anti* conformation, which would enable CoA to reach the active site. This transition is expected to be energetically favourable, but how this change occurs is unknown. According to this assumption, this changed binding site would allow binding of the free CoA from the solution in the anti form upon the completion of the first half reaction. The exact mechanism of CoA acetylation remains obscure.

The active site of PFL contains glycyl radical (Gly734), which serves as a radical storage, and two neighbouring cysteines involved in catalysis, Cys418 and Cys419. The experiments based on mutagenesis and inhibitor testing showed that both cysteines play a crucial role in the enzyme catalysis. However, their mutation has no effect on the enzyme activation, i.e. glycyl radical is formed independent of the cysteine residues.[8] From the results of electronic paramagnetic resonance (EPR) experiments it was noticed that glycyl radical exchanges α-hydrogen atoms with solvent faster than expected and that mutation of Cys419 affects the equilibrium of this process, but mutation of Cys418 has no such effect.[9] This information suggests that Cys419 catalyzes this rapid exchange, which implied close spatial proximity of Gly734 and Cys419. This conclusion was later verified when crystal structure was solved.[6] Additional experimental result supporting this conclusion came with studying of the inactivation of PFL in the presence of oxygen, when sulfinyl (RSO·) i peroxyl (ROO·) protein radicals are formed upon contact with oxygen, located on Cys419 i Gly734, respectively.[10] An interesting observation related to the cleavage of polypeptide chain caused by the oxygen has been made after discovery of YfiD protein (14 kDa). Namely, YfiD protein associates with cleaved PFL chains to restore its activity by introducing a new glycyl radical centre, instead of the one lost in the oxygen inactivation process. Basically, it serves as a spare part for PFL glycyl radical domain in cells that have experienced oxidative stress.[11]

To become fully active, PFL undergoes posttranslational modification in which glycyl radical is introduced into system. This process is catalyzed by PFL activating enzyme (PFL-AE), a monomeric protein (28 kDa) which contains [4Fe-4S] centre and belongs to SAM superfamily

(see Chapter 2).[12] Namely, this reaction requires S-adenosylmethionine (SAM or AdoMet) and reduced flavodoxin as co-substrates:[13]

$$\left[\text{-NH-CH}_2\text{-CO-}\right] + \text{SAM} + e^- \longrightarrow \left[\text{-NH-}\dot{\text{C}}\text{H-CO-}\right] + \text{5'-deoxyadenosyl} + \text{Met}$$

The side products of this reaction are 5'-deoxyadenosine and methionine. General mechanism involves formation of transient 5'-deoxyadenosyl radical, followed by hydrogen abstraction from Gly734 in *E.coli* PFL. The [4Fe-4S] cluster needs to be reduced prior to SAM cleavage. The activation of PFL requires an allosteric effector molecule, pyruvate or oxamate, although it is unclear how these compounds affect activation process. Apparently, oxamate has no affinity for the catalytic site of the activated PFL or to site which modulates its activity. Only one active site per dimer is activated.[14]

Recent years were marked with significant advances toward general better understanding of SAM radical enzymes, including PFL-AE.[15] It was the first activase for which the crystal structure was solved,[16] providing a better insight in this powerful radical machinery. Considering the fact that Gly734 is buried 8 Å under protein surface, it was a long standing question how 5'-deoxyadenosyl is able to selectively abstract hydrogen from that atom. It has been suggested that PFL undergoes significant conformational changes for this reaction to take place. Recent study shows that PFL has two possible conformation states in the equilibrium;[17] a closed state with buried Gly734 found in the crystal structures, which is believed to be the resting state of the enzyme. This is the state of the catalytically active PFL, where Gly734 radical is close to the active site and protected from the solvent. The second conformation of PFL is an open state, in which the loop carrying Gly734 is solvent exposed and available to interact with PFL-AE. It seems that these two states are in equilibrium, in which the closed state is favoured. However,

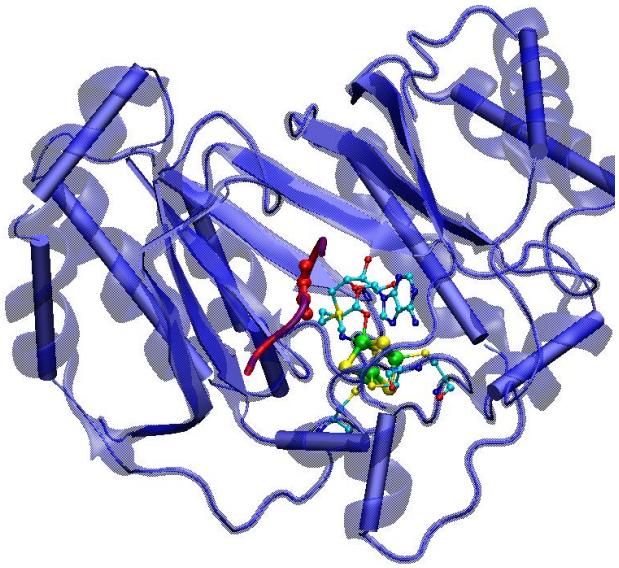


Figure 3.2 The crystal structure representing PFL-AE with SAM bound to [4Fe-4S] cluster. In addition, a peptide made of 5 residues is bound to the enzyme mimicking the environment of Gly734 in PFL (red), which is a natural substrate.

presence of PFL-AE shifts that equilibrium toward open state.

When it comes to deactivation of PFL, a long-standing hypothesis was that a special protein AdhE was responsible for quenching glycyl radical in PFL when cell switches back to aerobic metabolism[18]. In this way, PFL remains intact and can be potentially become re-activated again under anaerobic conditions. However, later study has shown that AdhE is not PFL deactivating enzyme, while small molecules such as mercaptoethanol and dithiothreitol proved as efficient deactivators of PFL under given reaction conditions.[19] Larger thiols, such as cysteine or glutathione, show no detectable deactivation. The results suggest that thiols must access the active site glycyl radical directly in order to efficiently promote inactivation, since the smallest thiols provide the most efficient inactivation. Non-thiolic reductants, such as dithionite and ascorbic acid, also inactivate PFL, with dithionite showing a rate similar to that of DTT and ascorbate being signifciantly slower.

3.1.2 Suggested Mechanisms of Catalysis

Upon activation, PFL catalysis proceeds in two steps and it is fully reversible (K_{eq}=750). The turnover number in the forward direction is k_{cat}=770 s^{-1}, while the reverse reaction is less efficient with k_{cat}=260 s^{-1}. In the first step pyruvate is cleaved into formate and acetyl group (K_{eq}=50), which remains bound to the enzyme until transfer to coenzyme A in the second half-reaction (K_{eq}=15):[1]

$$\text{PFL} + \text{pyruvate} \rightleftharpoons \text{acetyl-PFL} + \text{formate}$$

$$\text{acetyl-PFL} + \text{CoA} \rightleftharpoons \text{PFL} + \text{acetyl-CoA}$$

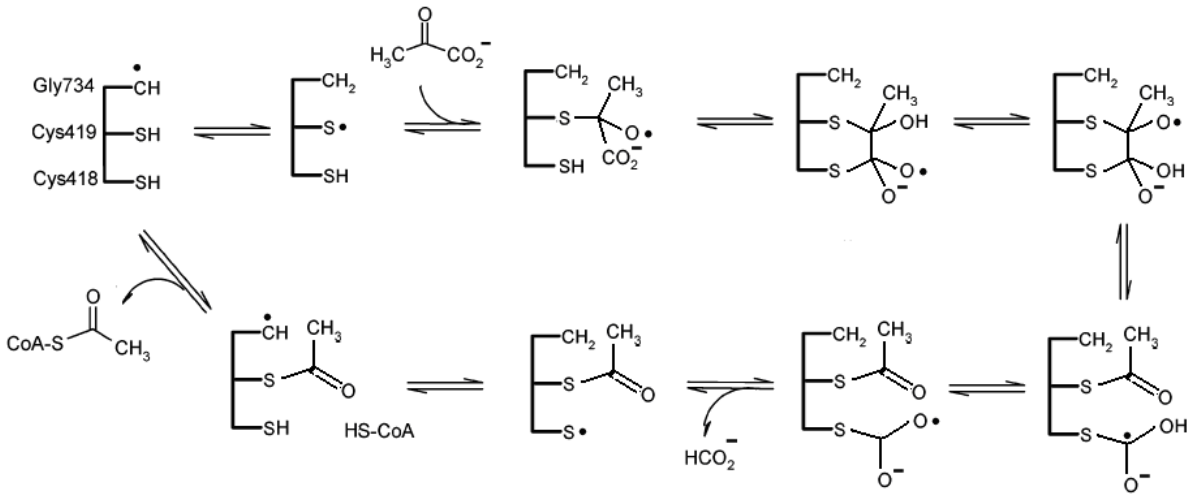
The presence of CoA is not mandatory for the first half-reaction to take place, but it can be bound to the enzyme in a spectator mode. This type of catalysis is known as *ping-pong* mechanism. This mechanism involves more than one substrate and covalent or non-covalent modification of the enzyme (acetyl-PFL) as result of the reaction with the first substrate (pyruvate), followed by the reaction with the second substrate (CoA) and restoration of the initial state of the enzyme (PFL).

In the first step, it is assumed that a cysteine residue attacks carbonyl carbon after radical transfer from Gly734, which results with C-C bond cleavage. It has been established that besides Gly734, two cysteines are crucial for catalysis – Cys418 and Cys419. Radical is

transferred from Gly734 to Cys419 due to their spatial proximity, but it was a long-standing puzzle which cysteine is the site of acetylation. Now, it has been widely accepted that Cys418 is the carrier of the acetyl group. Evidence supporting this hypothesis comes from the mutagenesis experiments, which reveal that exchange of acetyl group is hindered when Cys418 is mutated to serine, while the same mutation of Cys419 has no effect on the mentioned exchange.[8] This indicates that Cys418 is the primary reactant in the thioester exchange with CoA. Another important observation is related to the hydrogen exchange rate on glycyl radical, which remains unchanged upon enzyme acylation, implying that Cys419 should be available to catalyze this process.[9] Experiments involving inactivation of PFL by hypophosphite, a close analogue of formate, have shown that the analogue reacts with E-acetyl radical intermediate to produce 1-hydroxyethyl phosphonate, covalently bound to Cys418.[20] Finally, according to the crystallographic data, the distance between Sγ atom of Cys418 and carbonyl carbon (C2) is 2.6 Å. The angle of 103° with the carbonyl group of pyruvate (O-C2-S), which is close to the optimal angle of 109° for the radical attack on sp2 hybridized carbonyl carbon.[6] Based on these findings, Cys418 has been identified as the preferred site of acetylation.

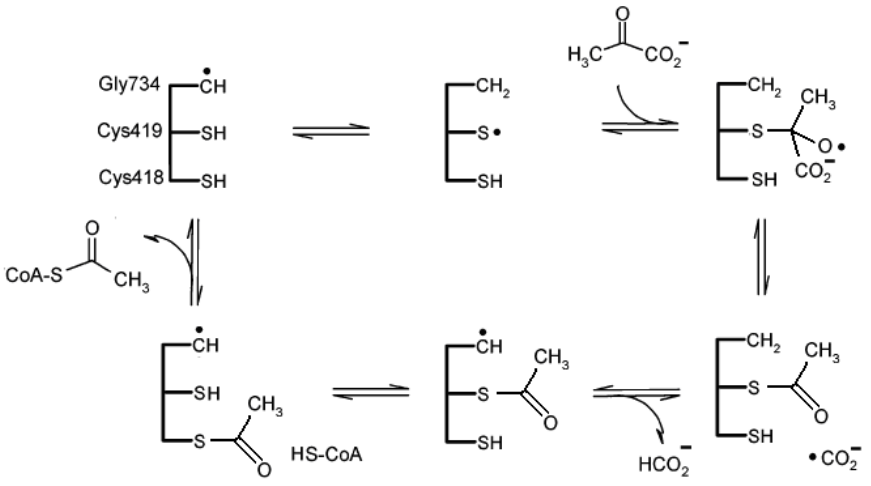
For the second half-reaction was initially thought to proceed via non-radical mechanism, but it has been established that the radical content has a dramatic influence on the acetyl exchange rates. Namely, rates of acetylation and deacetylation by acetyl-CoA/CoA via non-radical pathway are about 10^5 slower than its radical counterpart.[8] There are several hypotheses regarding the mechanism of acetyl transfer to CoA, but none of them has been yet corroborated, leaving some open questions. The most important one involves the conformational changes that PFL should undergo in order to allow CoA to reach the active site and take over the acetyl group. What exactly triggers this change and what kind of structural modifications are required remains unclear.

Of course, over the years several different mechanisms of PFL were suggested, adapting to the new information gathered during investigation of this interesting enzyme. The first mechanism was proposed by Knappe and co-workers (Scheme 3-1).[8] In this mechanism, Cys419 attacks pyruvate, forming thiohemiketal and the radical is transferred to Cys418. The Cys418 thiyl radical forms an adduct with the thiohemiketal carboxyl group, followed by a hydrogen shift to yield the alkoxy-radical intermediate that undergoes the homolytic C-C bond cleavage. The products of this reaction are the acetylated Cys419 and formate-radical adduct on Cys418. The latter dissociates to form free formate, followed by the restoration of Gly734 radical. The final transfer of acetyl to CoA uses Cys418 as relay.



Scheme 3-1. Reaction mechanism of PFL suggested by Knappe *et al.* in 1993.[8]

An alternative mechanism was proposed by Kozarich *et al.*[9] and the first step of that mechanism involves hydrogen transfer from Cys419 to Gly734, followed by addition of thiyl radical onto substrate to form tetrahedral oxy-radical intermediate. This intermediate subsequently dissociates to acetylated Cys419 and formyl radical. The latter abstracts hydrogen from Gly734, regenerating the glycyl radical. The acetyl moiety shifts from Cys419 to Cys418 via reversible transesterification, and finally to CoA (Scheme 3-2).

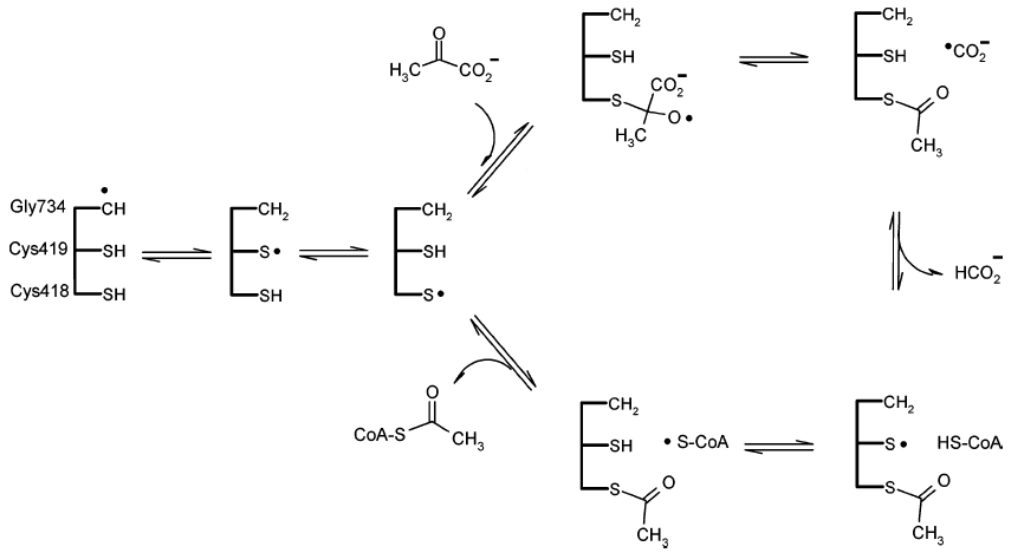


Scheme 3-2. The mechanism proposed by Kozarich *et al.* in 1995. [9]

In 1998. Himo and Eriksson made a theoretical study of this mechanism using small model systems to describe the important reactants, which confirmed the plausibility of the homolytic radical reaction pathway.[21] In their study, cysteine residues were represented with methylthiol

and pyruvate was in the neutral form. The important modification of the suggested mechanism was introduced by suggesting that acetyl transfer from the active site cysteine to CoA proceeds via radical mechanism. The study showed that the thiyl radical attack on thioacetate is energetically more favourable than its thiol or thiolate counterparts, indicating that the previously considered heterolytic acetyl transfer is less likely possibility. The lack of general acid-base catalyst in the active site supports the hypothesis of homolytic acetyl transfer.

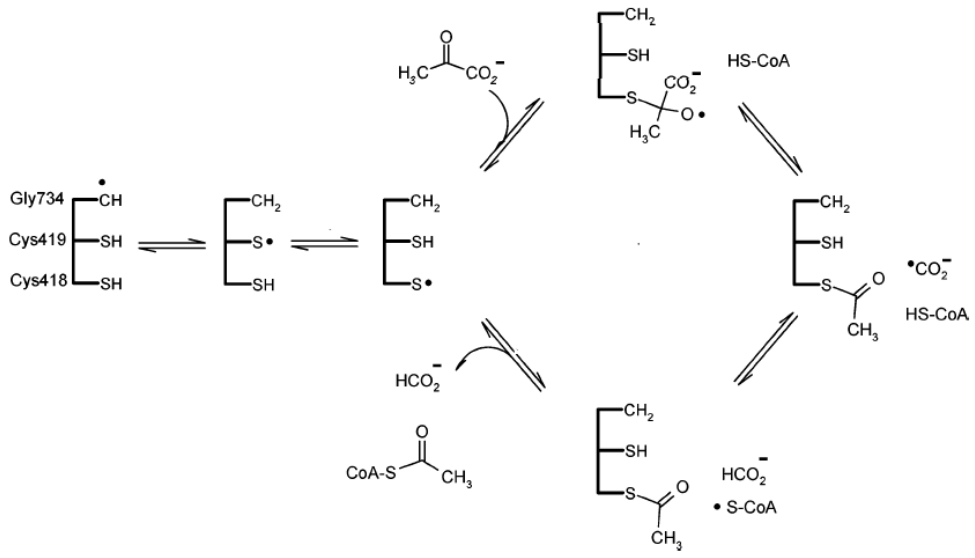
Kozarich *et al*. did not consider the possibility of the direct attack of Cys418 on pyruvate in their mechanism, which was assumed to be buried deeply in the protein interior. However, the structural data obtained from crystallographic experiments shed a new light on the roles played by two cysteines in the active site. It was shown that direct addition of Cys418 to pyruvate is entirely consistent with the structural parameters, as described above. Based on this fact and other structural and biochemical experiments, including theoretical studies, a new mechanism of PFL catalysis was proposed by Knappe *et al*. in 1999.[6] This mechanism starts with substrate binding in the active site, triggering radical generation at Gly734 by the activase. The radical is then transferred to Cys418, using Cys419 as a relay. The Cys418 thiyl attacks carbonyl carbon of substrate to yield tetrahedron oxy-radical intermediate that dissociates to thioacetate and formyl. The formyl radical regenerates thiyl radical on Cys419 by hydrogen abstraction. The second half-reaction was suggested to proceed via hydrogen transfer between CoA and Cys419, resulting with thiyl radical on CoA. After transacetylation, radical on Cys418 is restored and ready to enter new catalytic cycle (Scheme 3-3).



Scheme 3-3. The mechanism proposed by Knappe *et al*. in 1999.

Another theoretical study of the mechanism described above was made by Lucas *et al*. in 2003.[22] In this study, small model systems representing the active site were again used to describe the PFL catalysis, but the neutral species used in previous study were replaced with negatively charged substrate and products. The protein surrounding was described with polarizable continuum model. The putative tetrahedron intermediate was not found on the reaction pathway, but a quasi-planar transition state and concerted mechanism was suggested with formyl radical as leaving group. Another interesting observation was made regarding hydrogen abstraction from Cys419 by formyl radical, which proceeds seemingly without a barrier. Hydrogen abstraction from Cys410 is also stereochemically more favourable than the same reaction involving Gly734. The overall reaction was characterized as exothermic (-5.8 kcal mol^{-1}).

In 2004., Himo and Guo revisited the PFL mechanism in a computational study using large models of the active site up to 75 atoms that were based on the available crystal structure containing pyruvate.[23] In addition to negatively charged pyruvate, this model includes two positively charged arginine side chains (Arg176 and Arg435) that play important role in the substrate binding (salt bridge formation and hydrogen bonding), and catalytic triad (Gly734, Cys418, Cys419). Moreover, Asp661 forming a salt bridge with Arg176 is also included and represented with acetate. In this so-called frame model, certain atoms are kept frozen to their X-ray position, usually those where truncation was made. A new perspective to the investigated mechanism was provided by introducing an alternative step in which formyl radical abstracts hydrogen directly from CoA, rendering acetyl transfer from Cys418 to CoA energetically feasible (Scheme 3-4).



Scheme 3-4. Reaction mechanism for PFL modified by Himo and Guo in 2004.

3.1.3 Inhibition of PFL by Substrate Analogues

One of the commonly used PFL inhibitors is oxamate, an isosteric and chemically inert analogue of natural substrate pyruvate. It binds to the active site in a mode very similar to that of pyruvate, except the oxamate experiences slight rotation (6°) in comparison to pyruvate and it is 0.4 Å further away from Cys418.[7] It acts as a competitive inhibitor with rather weak binding affinity K_I > 20 mmol dm^{-3}, while value of K_M ~ 2 mmol dm^{-3} was measured for pyruvate at 30^0C. Activation of PFL requires presence of either oxamate or pyruvate due to their effector activity, which is the main reason for frequent usage of oxamate in experiments involving PFL. However, oxamate undergoes no further reaction once it occupies the active site and one of the aims of this thesis is to provide an explanation for such behaviour.

Another pyruvate analogue used in the experiments is methacrylic acid, for which K_I=0.42 mmol dm^{-3} was determined. In contrast to oxamate, PFL undergoes a suicide reaction with methacrylate that results with 2-carboxy-propyl substituted Cys418 residue.[24] This is the additional argument supporting the hypothesis about Cys418 as a site of acetylation during the catalysis.

Formate is the substrate for PFL in the reverse reaction and its analogue hypophosphite was also used in a set of biochemical experiments. Hypophosphite reacts with the acetylated enzyme to form 1 –hydroxyethylphosphonate with a thioester linkage to the Cys-418, which is characterized as a dead-end product.[20,25]

3.1.4 REFERENCES

1 Knappe, J.; Blaschkowski, H.P.; Gröbner, P.; Schmitt, T. *Eur. J. Biochem*. **1974**, 50, 253.

2 Sawers, G.; Summpann, B. *J. Bacteriol*. **1992**, 174, 3474.

3 Wagner, A. F. V.; Frey, M.; Neugebauer, F. A.; Schäfer, W.; Knappe, J. *Proc. Natl. Acad. Sci. USA* **1992**, 89, 996.

4 Yang, J.; Naik, S.G.; Ortillo, D. O.; García-Serres, R.; Li, M.; Broderick, W. E. Huynh, B. H.; Broderick, J. B. *Biochemistry*, **2009**, 48 (39), 9234.

5 Conradt, H.; Hohmann-Berger, M.; Hohmann, H. P.; Blaschkowski, H. P.; Knappe, J. *Arch. Biochem. Biophys*. **1984**, 228(1), 133.

6 Becker, A.; Fritz-Wolf, K.; Kabsch, W.; Knappe, J.; Schultz, S.;Wagner, A. F. V. *Nat. Struct. Bio*. **1999**, 6, 969.

7 Becker, A.; Kabsch, W. *J. Biol. Chem*. **2002**, 277 , 40036.

8 Knappe, J.; Elbert, S.; Frey, M.; Wagner, A. F. V. *Biochem. Soc. Trans*. **1993**, 21, 731.

9 Parast, C.V.; Wong, K. K.; Lewisch, S. A.; Kozarich, J. W. *Biochemistry* **1995**, 34, 2392.

10 Reddy, S.G.; Wong, K. K.; Parast, C.V.; Peisach, J.; Magliozzo, R.; Kozarich, J.W. *Biochemistry* **1998**, 57, 558.

11 Wagner, A. F.; Schultz, S.; Bomke, J.; Pils, T.; Lehmann, W. D.; Knappe, J. *Biochem. Biophys. Res. Comm*. **2001**, 285 (2), 456.

12 Külzer, R.; Pils, T.; Lappl, R.; Hüttermann, J.; Knappe, J. *J. Biol.Chem*. **1998**, 273, 4897.

13 Frey, M.; Rothe, M.;Wagner, A.F.V.; Knappe, J. *J. Biol.Chem*. **1994**, 269, 12432.

14 Unkrig, V.; Neugebauer, F. A.; Knappe, J. *Eur. J. Biochem*. **1989**, 154, 723.

15 Vey, J. L.; Drennan, C. L. *Chem. Rev*. **2011**, 111, 2487.

16 Vey, J. L.; Yang, J.; Li, M.; Broderick, W. E.; Broderick, J. B.; Drennan, C. L. Proc. *Nat. Acad. Sci. USA* **2008**, 105 (42), 16137.

17 Peng, Y.; Veneziano, S. E.; Gillispie, G. D.; Broderick. J. B. *J. Biol. Chem*. **2010**, 285 (35), 27224.

18 Kessler, D.; Herth, W.; Knappe, J. *J. Biol. Chem*. **1992**, 267(25), 18073.

19 Nnyepi, M. R.; Peng, Y.; Broderick, J. B. *Arch. Biochem. Biophys*. **2007**, 459, 1.

20 Plaga, T. W.; Frank, K.; Knappe. J. *Eur. J. Biochem*. **1988**, 178, 445.

21 Himo, F.; Eriksson, L. A. *J. Am. Chem. Soc*. **1998**, 120, 11449.

22 Lucas, M. F.; Fernandes, P. A.; Eriksson, L. A.; Ramos, M. J. *J. Phys. Chem. B* **2003**, 107, 5751.

23 Guo, J.-D.; Himo, F. *J. Phys. Chem. B* **2004**, 108, 15347.

24 Plaga, W.; Vielhaber, G.; Wallach, J.; Knappe, J. *FEBS Letters* **2000**, 466, 45.

25 Unkrig, V.; Neugebauer, F. A.; Knappe, J. *Eur. J. Biochem*. **1989**, 154, 723.

3.2 A Small-Model Approach in Modelling PFL Catalysis: First Half-Reaction

Computational chemistry has proven itself to be a valuable tool for studying chemical reactions and their mechanisms. Particularly important in recent years has been the application of these tools to problems of biological importance, such as enzymatic catalysis. Such systems are, however, inherently large and their treatment necessarily requires some kind of approximation. One solution has involved treating the entire system semi-empirically[1] using, for example, the divide-and-conquer approach to facilitate the computation.[2] Alternatively, one may divide the system into different regions, each treated with a different theoretical approach. Examples of this type include the QM/MM methodology[3-9] or, more generally, a multi-layered ONIOM approach.[10,11] Finally, the system can be simply truncated to a size amenable to the target theoretical treatment, while attempting to keep the salient features of the biological system under study.[12,13] Despite its apparent shortcomings, this latter approach has enjoyed many successful applications,[14-16] revealing fundamental mechanistic aspects while, in particular, allowing the retention of high accuracy in the calculations.[17]

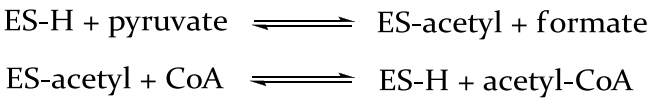
A fundamental issue associated with the construction of a small model system is how to assign the protonation state of titratable groups, and hence determine the charge of the model itself. Even though this problem also exists for layered and even semi-empirical treatments, it is more prominent in the field of small models. Perhaps the most common approach for small models is to use neutral systems in as far as it is possible. The argument underlying this choice is that the interior of proteins is predominately associated with low polarity and therefore significant charge separation is not to be expected.[18,19] This logic has been extended to the relatively widespread treatment of the protein environment as a homogeneous medium with a low dielectric constant, usually $\varepsilon \approx 4$.[13,20]Nevertheless, many important conclusions have arisen from the use of charged models,[15,17] even for the systems which had appeared to be exemplary for the neutral model approach.[21,22]

Charged species are naturally able to undergo strong interactions with other charged and polar residues in the protein. A common example is a charged carboxylate group, which is often found participating in hydrogen bonds (ionic or salt bridges) with positively charged amino-acid side chains, such as those of arginine and lysine. Even though these interactions are associated with charge separation, they do serve to partially neutralize the substrate's charge, thus serving as a potential alternative justification for the use of neutral models.

It is the aim of this paper to investigate, by means of a case study, the mechanistic effect of choosing different protonation states for a substrate carboxylate group that is involved in a salt bridge with an arginine residue. By comparing the different treatments with results that explicitly include the salt bridge we hope to derive a sound recommendation for the best choice of protonation state, under these circumstances, in small-model treatments.

We have chosen the substrate mechanism of Pyruvate Formate-Lyase (PFL) to serve as our case study for this investigation. The reason underlying this choice is that it is a relatively well-studied system, whose mechanism has already been computationally investigated using a number of different small-model approaches.[23-25]

PFL is a key enzyme of anaerobic glucose metabolism in *E.coli* and other microorganisms, catalyzing the CoA-dependant reversible cleavage of pyruvate into acetyl-CoA and formate:[26]



Scheme 3-5 Substrate transformation catalyzed by PFL.

In its active form PFL is a glycyl radical enzyme in which an unpaired electron is located on the Cα atom of Gly734.[27] This radical is generated through the action of PFL-activase, a specific activating enzyme. Besides Gly734,[28] it has been established that two neighboring cysteines, namely Cys418 and Cys419, are required for the catalysis.[29] In addition, two arginine residues, Arg176 and Arg435, serve to bind and position the substrate, by forming salt bridges with the carboxylate group of pyruvate (Figure 3.3).[30]

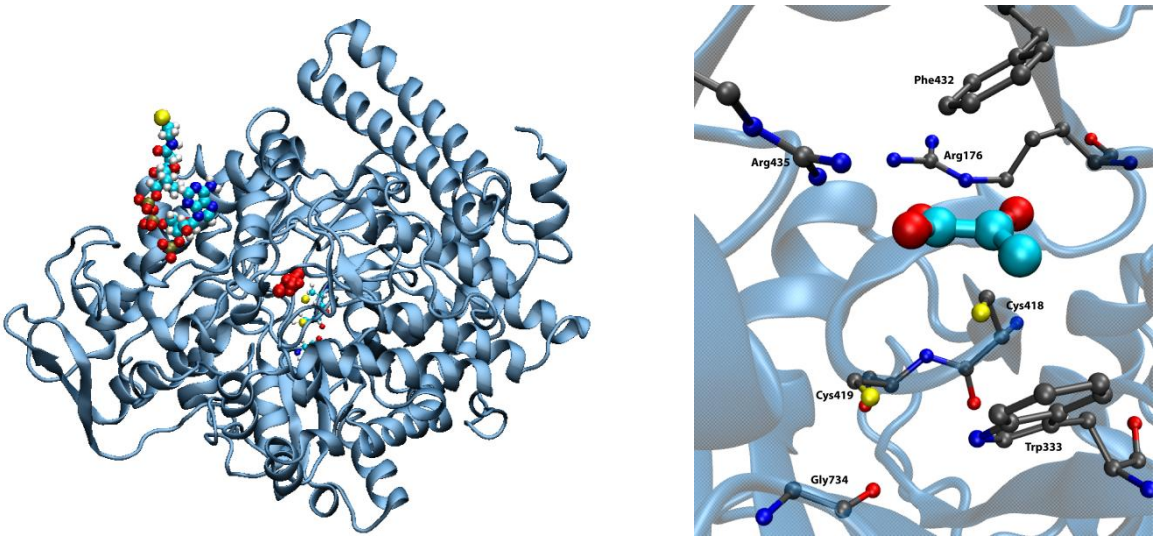
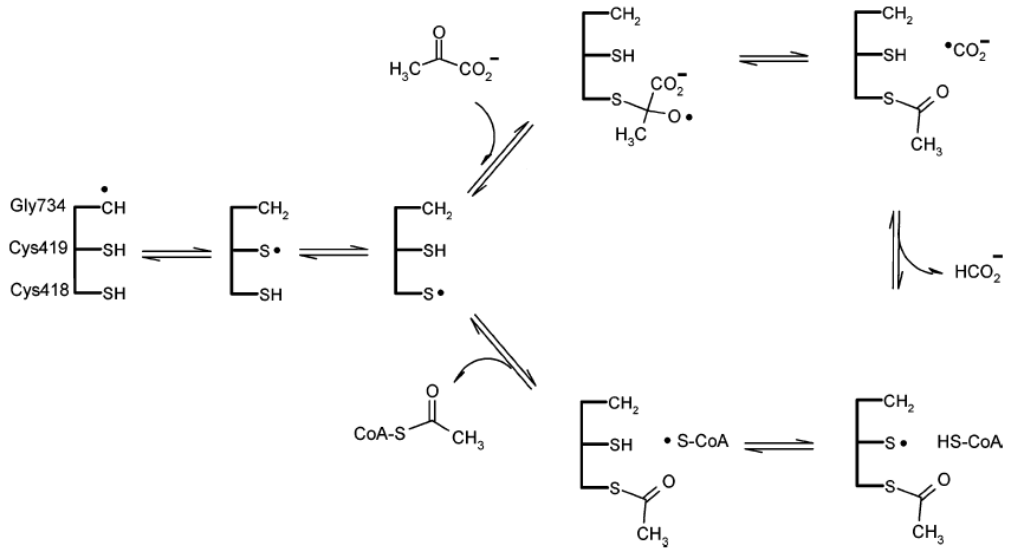


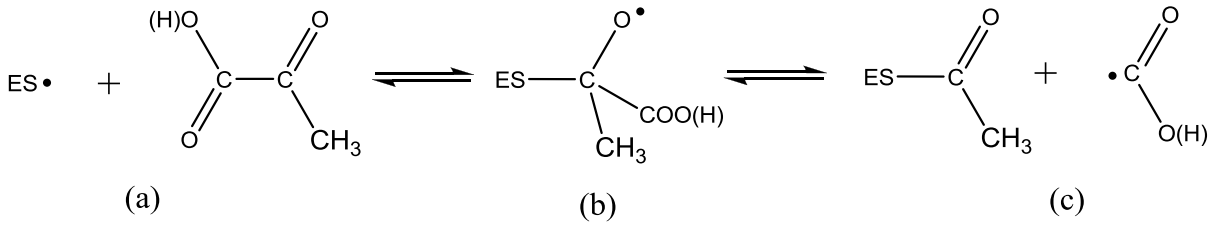
Figure 3.3 (a) Crystallographic structure of the PFL monomer in complex with pyruvate and CoA; (b) structure of the PFL's active site.[30]

The currently accepted mechanism of PFL, shown in Scheme 3-6, was proposed by Knappe *et al*. in 1999.[31] According to this mechanism, hydrogen is shuttled from Cys418 through Cys419 onto the glycyl radical. The resulting cysteinyl radical at position 418 subsequently attacks the carbonyl carbon of pyruvate forming a putative tetrahedral intermediate. Fragmentation of this intermediate into formyl radical and acetylated Cys418 is followed by the formation of formate and eventual actetylation of CoA. The exact details of the latter stages of the mechanism have not yet been established. They are, however, beyond the scope of the present contribution.



Scheme 3-6 Currently accepted PFL mechanism proposed by Knappe *et al.* in 1999.[31]

Instead we focus here on the transformation of the substrate itself (the first half reaction), which involves the radical addition and fragmentation steps, as shown in Scheme 3-7.



Scheme 3-7 Fragmentation of pyruvate into acetylated enzyme and formyl radical, catalyzed by PFL.

Prior to the suggestion of the mechanism shown in Scheme 3-6,[31] Eriksson and Himo completed the first small-model computational study of the PFL mechansim.[23] Therein, the methyl thiyl radical was used as a model for the protein-bound cysteinyl radical, a choice justified on the basis of the relevant S-H bond strengths. The substrate was represented by neutral pyruvic acid (protonated pyruvate). Their calculations showed, for the first time, that a

homolytic addition/elimination mechanism (as shown in Scheme 3-6 and Scheme 3-7) was energetically feasible. Specifically, they located a single tetrahedral intermediate (b, in Scheme 3-7) connected to states (a) and (c) in Scheme 3-7 by two separate transition structures. This study, using the B3LYP/6-311+G(2d,2p), was instrumental in construction of what is now the generally accepted mechanism.

Following the solution of the X-ray structure of PFL,[30,31] Lucas *et al*. published a new small-model computational study of the substrate mechanism using the B3LYP/6-311++(3df,3pd) methodology.[24] On this occasion, however, the substrate was represented by the anionic pyruvate (deprotonated pyruvic acid), a choice justified by the crystal structure as well as the known pK_a of pyruvic acid. The anionic substrate model predicts a concerted mechanism for the addition-elimination process, proceeding directly from state (a) to state (c) with no evidence for an intermediate corresponding to state (b).

In a more recent study, Himo and Guo employed an extended (but still truncated) model, consisting of an anionic representation of the substrate (pyruvate) and (charged) models for those amino acids considered crucial for catalysis, including Arg176 and Arg435 (Figure 3.3b).[25] During the calculations, selected atoms of the protein residues used for building the model were kept fixed at their X-ray positions (the so-called *frame-model*). In this study, which employed the B3LYP/6-311+G(2d,2p) level of theory, two intermediates were located along the reaction coordinate, as were three corresponding transition structures. Both intermediates (lying between states (a) and (c) in Scheme 3-7) were found to lie in shallow minima on the potential energy surface.

These three different small-model studies clearly show how the choice of model strongly affects even the qualitative appearance of the potential energy surface. It is in this context that we wish to directly compare, on an equal footing, the results of a neutral (carboxylic acid) model, an anionic (carboxylate) model, and an extended model incorporating a single salt bridge between the side chain of an arginine residue and the substrate carboxylate group. We wish to use these comparisons to recommend the most appropriate model for treating arginine-bound carboxylates in the context of truncated small models. In a related matter, by performing high-level molecular orbital calculations on these same systems, we wish to determine the reliability of the DFT approach on model systems of this type.

3.2.1 COMPUTATIONAL DETAILS

All geometry optimizations and frequency calculations employed the B3LYP/6-31+G(d) approach. Improved relative energies were obtained with the B3LYP/G3MP2Large level of theory and by using the composite G3(MP2)-RAD method. This method was developed by Radom and co-workers[32] (based on G3(MP2)[33]) to give more reliable predictions of the energies of radical species. The relative energy of two species using the G3(MP2)-RAD procedure is given by:

$$\Delta E = \Delta E[\text{RHF} - \text{UCCSD(T)/6} - 31\text{G(d)}] + \Delta E[\text{ROMP2/G3MP} \quad \text{2Large}] - \Delta E[\text{ROMP2/6} - 31\text{G(d)}] + \Delta E(\text{ZPVE})$$

All energies presented were calculated at 0 K and include an unscaled zero-point vibrational energy (ZPVE) correction obtained from the calculated Hessians. All computations were performed using methods implemented in the Gaussian03 program package,[34] except for the RHF-UCCSD(T) calculations,[35] which were obtained with Molpro (Version 2006.1).[36]

The connectivities of the potential energy surfaces presented were verified by following the intrinsic reaction coordinate (IRC) from each transition structure. In the case of the reactant and product complexes, the final geometries were obtained by optimizing the end-points of the relevant IRCs. These complexes were not, however, subjected to exhaustive conformational searches. Our rationale for this approach is simply that the relevance of the complexes to the enyzmatic reaction is already somewhat questionable. Such relevance would only be further diminished by locating the lowest-energy bi- or ter-molecular complex in each case.

3.2.2 Results and Discussion

3.2.2.1 Neutral Model

Our neutral model, shown in Figure 3.4, is virtually identical to the previously discussed model of Eriksson and Himo.[23] The only significant difference is that we elected to use the *syn* conformation of the carboxylic group as opposed to the lower energy *anti* orientation used previously.[23]

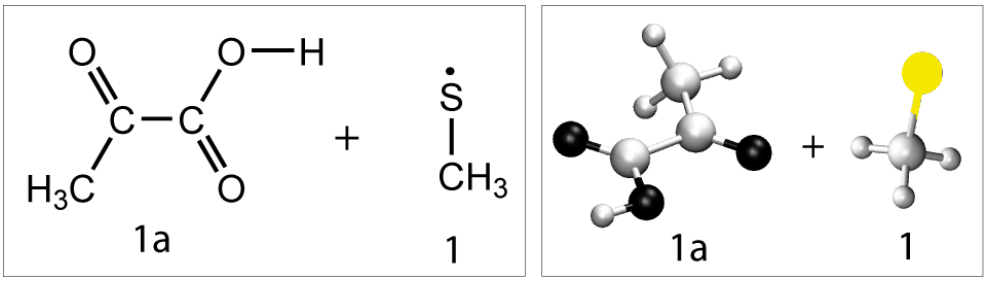


Figure 3.4 The neutral model consists of pyruvic acid (**1a**) and methylthiyl radical (**1**) instead of cysteine.

Our reasoning for using the less stable orientation (**1a**) is that we are attempting to mimic an arginine-bound carboxylate with our neutral model. The X-ray crystal structure, as well as a simple geometric argument, places the arginine residue on the side of the molecule opposite to the acetyl group. In addition, we deem the formation of an intramolecular hydrogen bond in the substrate to be unlikely at the active site of the enzyme.

The reaction profile for the addition of the methyl thiyl radical to pyruvic acid (in the conformation corresponding to **1a**) is shown in Figure 3.5. The first obvious difference between this figure and the results of Eriksson and Himo[23] is the appearance of two shallow minima, corresponding to intermediates, as opposed to just one (such as state (b) in Scheme 3-7). Closer inspection of these two intermediates reveals that they are similar in nature to those found with the somewhat larger model employed by Himo and Guo,[25] with the first (**3a**) having a slightly extended C-S bond while the second (**4a**) is characterized by a longer than normal C-C bond.

In agreement with the initial neutral model study,[23] we find that the rate limiting step corresponds to C-C bond cleavage, with **TS:4a→5a** being the highest stationary point on the potential surface. The mechanism resembles ascending a ladder, starting with the formation of the S-C bond (**TS:2a→3a**), followed by the transition between the two intermediates

(**TS:3a→4a**) and finally the cleavage of the C-C bond (**TS:4a→5a**). Both of the intermediates lie in shallow minima on the PES, but the second intermediate (**4a**) lies approximately 20 kJ/mol higher than the first one (**3a**). After the C-C bond is broken, the reaction proceeds downhill towards the separated products (**6+6a**), passing through a minimum corresponding to a gas-phase product complex (**5a**).

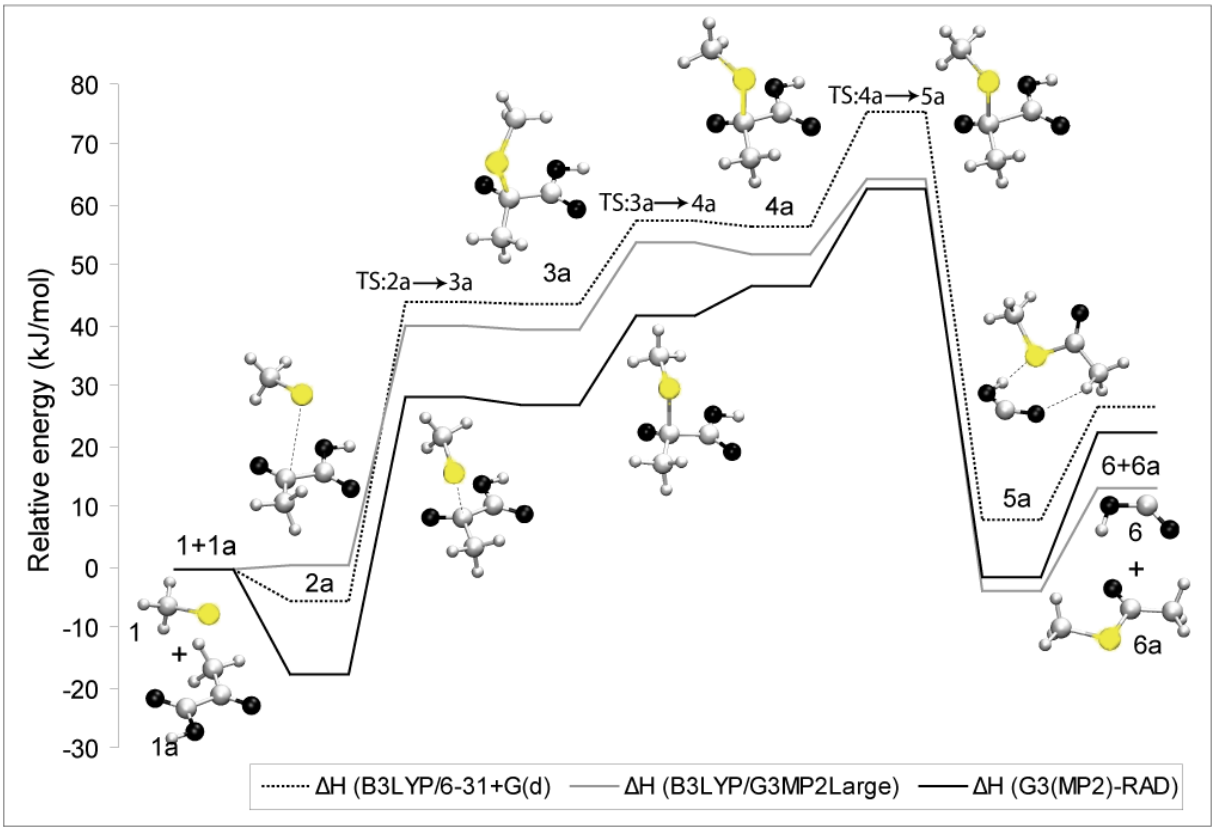


Figure 3.5 Neutral model. Addition of methylthiyl radical to carbonyl the carbonyl group of pyruvic acid at oK (geometries presented were optimized at B3LYP/6-31+G(d)).

It is interesting that a minor alteration in the carboxylic proton orientation is sufficient to introduce a second intermediate on the pathway. Indeed in some ways, this difference causes the pathway in Figure 3.5 to resemble more closely that in reference 25 (with two intermediates) than reference 23 (with a single intermediate).

Several previous studies have found that hybrid DFT methods such as B3LYP tend to underestimate reaction barriers.[37] Interestingly, we find exactly the contrary result here. That is, in comparison to G3(MP2)-RAD, B3LYP/6-31+G(d) tends to overestimate the energetics of the reaction in Figure 3.5 by between approximately 10 and 20 kJ mol^{-1}. B3LYP/G3MP2Large performs generally better but its discrepancies also tend towards overestimation of the barriers with respect to the more reliable treatment. There is, however, a relatively large discrepancy

between G3(MP2)-RAD and the large basis set DFT treatment concerning the overall endothermicity of the reaction. Another interesting observation can also be made in case of the reactant complex (**2a**). Both DFT treatments predict the complex to be very weakly bound (less than approximately 5 kJ mol^{-1}), whereas the G3-based methodology predicts a binding energy of almost 20 kJ mol^{-1}. This would tend to suggest that the binding energy of this complex is dominated by dispersion interactions, which are poorly handled by DFT.

3.2.2.2 Anionic Model

The presence of charge in a small model system can have a significant influence on the resulting mechanism of the modelled chemical reaction. Nevertheless, charge is sometimes a fact of life and it is important to know its impact. This is particularly true in the case of PFL because, as pointed out by Lucas *et al.*[24] , the pK_a of pyruvic acid is 2.5 and so it should be expected to be found as a charged species under physiological conditions. Our anionic model, shown in Figure 3.6, is identical to the simplest one employed by Lucas *et al.*[24]

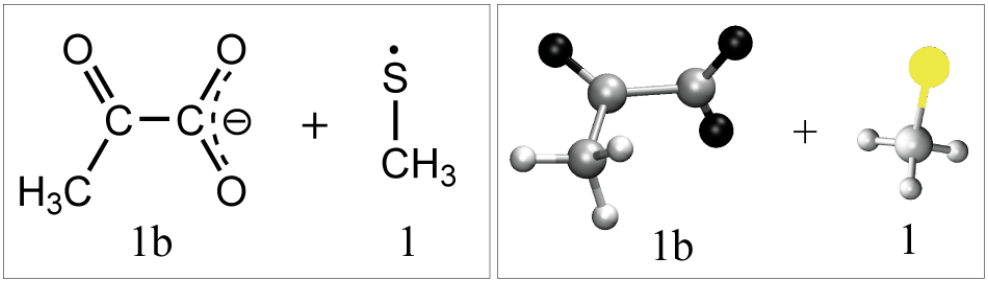


Figure 3.6 The anionic model consists of a negatively charged pyruvate (**1b**) and thiyl radical (**1**) instead of cysteine.

The results for the addition of the methylthiyl radical to the negatively charged pyruvate are shown in Figure 3.6. To keep the comparisons with the neutral model on an equal footing, the reaction profile is shown relative to the isolated products. Naturally, one consequence of this is the significantly stronger complexation energies of the reactant and product complexes due to charge-dipole type interactions.

In agreement with Lucas *et al.*,[24] we find the mechanism to be concerted, with no intermediate to be found between the putative states (a) and (c) shown in Scheme 3-7. Also in agreement with Lucas *et al.*, we find the transformation of **2b** to **3b** to be endothermic using B3LYP/6-31+G(d). However, with both B3LYP/G3MP2Large and G3(MP2)-RAD, this transformation is actually exothermic, making the structure **3b** the lowest stationary point on

the potential energy surface. As discussed by Lucas *et al.*[24] the C-C bond in this complex is not completely broken. In this sense it is a loose analogue of the second intermediate (**4a**) in Figure 3.5 (see also the subsequent Section 3.2.2.3 and reference 25 for a discussion of this point). However, in stark contrast to intermediate **4a**, there is no benefit associated with further dissociation of intermediate **3b**. This implies that the anionic mechanism would have no particular reason to progress beyond this point.

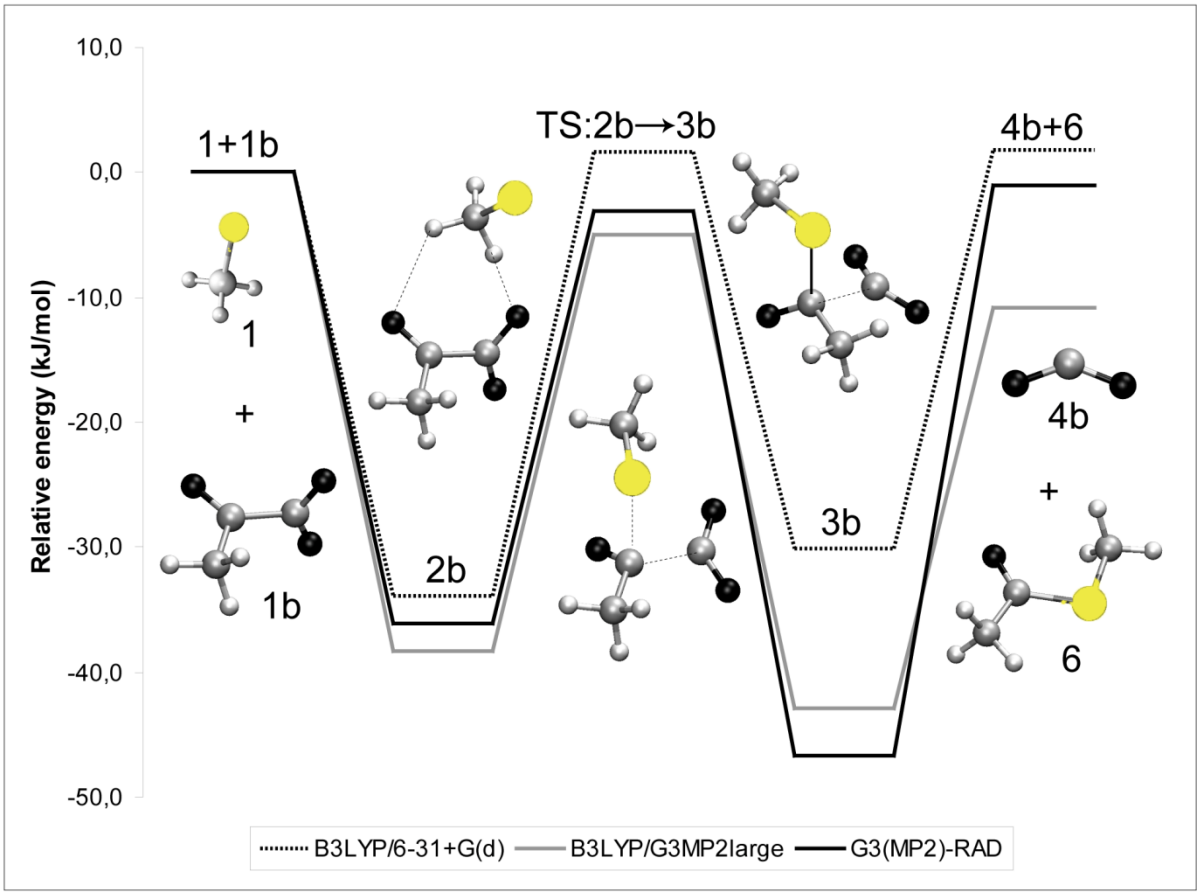


Figure 3.7 Anionic model: Addition of methylthiyl radical to carbonyl group of pyruvate at O K (geometries presented were optimized at B3LYP/6-31+G(d)).

The above result is actually important in the context of a potential QM/MM study of the PFL substrate transformation. More specifically, it could prove convenient to select the atoms shown in Figure 3.6 and Figure 3.7 as the QM region of the QM/MM Hamiltonian. If electrostatic embedding were employed such that the charges of the MM region were able to polarize the QM wave function, this should present no real problem. However, if a mechanical embedding approach were chosen, even if only for geometry optimizations,[11] one would expect the reaction profile to strongly resemble to the one shown in Figure 3.7, because the QM calculations in that case would not differ significantly from those presented here.

In a manner similar to that found in the neutral systems, the small basis set DFT results tend to overestimate the energies of the transition structure and intermediates when compared to the G3(MP2)-RAD results, particularly seriously so for the intermediate **3b**. The large basis set DFT calculations give much more satisfactory agreement with the higher level treatment, except for the overall exothermicity of the reaction.

3.2.2.3 Extended Model

As stated earlier, the principle aim of the work presented here is to compare, on an equal footing, the results from previously studied neutral and anionic systems to those obtained from one in which an anionic carboxylate is bound by an arginine residue. In the interest of continuing to obtain accurate results (G3(MP2)-RAD), we have chosen to truncate our model and represent the arginine with a methylguanidinium ion. Indeed, this simplification has already been successfully implemented in previous studies on salt bridges.[38-41]

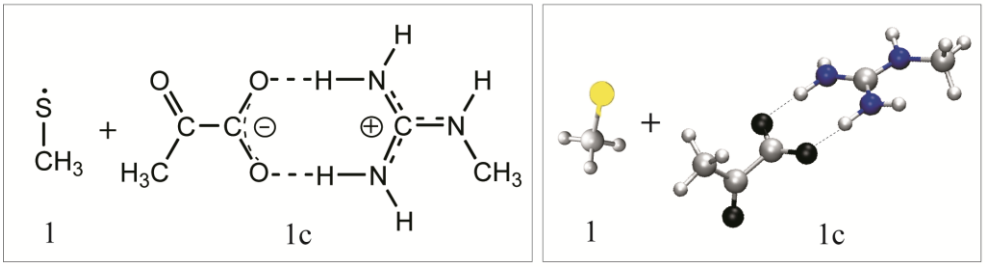


Figure 3.8 The extended model consists of pyruvate in a complex with protonated methylguanidinum and methylthiyl radical.

Three different tautomeric states can be envisioned for the resulting complex (**1c**), which is shown in Figure 3.8. Two such states are neutral while the third (pictured in Figure 3.8) is zwitterionic. The zwitterionic tautomer is calculated to be lower in energy than either of the neutral alternatives by all three theoretical methodologies used herein. In contrast, previous computational studies of carboxylate-methylguanidinium interactions, whose carboxylate group was provided by acetic acid, showed the neutral hydrogen-bonded complex to be slightly more stable in the gas phase.[38-40] This discrepancy is due to the fact that, the gas phase proton affinity of pyruvate (1395 kJ mol^{-1})[42] is around 55 kJ mol^{-1} lower than that of the acetate (1450 kJ mol^{-1}).[43,44]The other example where zwitterionic form is favored is for the complex of 2,5-dihydroxybenzoic acid and methylguanidinium.[41] It is important to note that the charge

separation leading to the zwitterionic state takes place already in the gas phase. Even a weakly polar environment would be expected to enhance this preference.

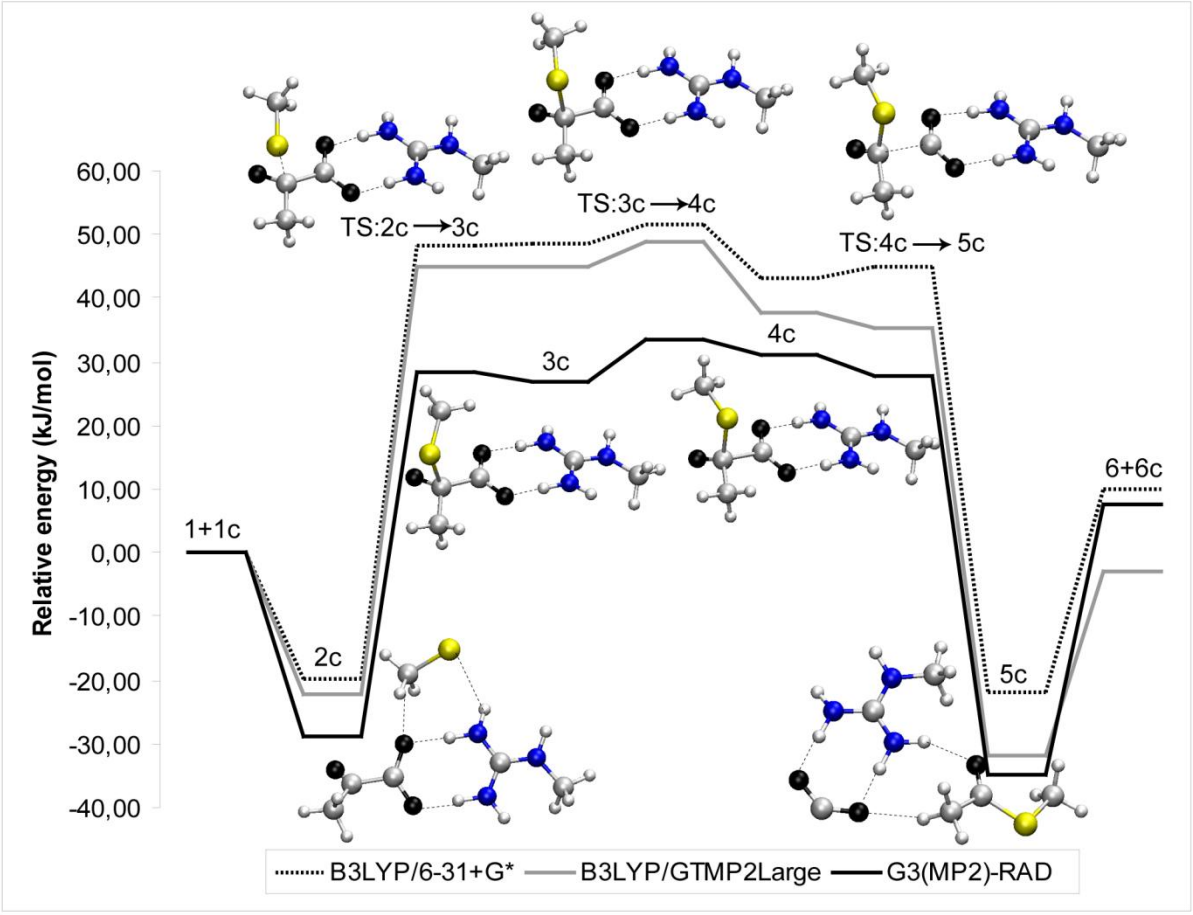


Figure 3.9 Exteneded model. Addition of methylthiyl radical to carbonyl group of pyruvate complexed with methlyguanidinium at oK (geometries presented were optimized at B3LYP/6-31+G(d) level of theory).

The potential energy surface obtained for the addition of methylthiyl radical to the complex between pyruvate and the methylguanidinium is shown in Figure 3.9. Following the complex formation, the reaction takes place in three steps. Firstly, the S-C bond is formed (**TS:2c→3c**). Subsequently, a transition between the two intermediates (**TS:3c→4c**) takes place and finally the C-C bond is cleaved (**TS:4c→5c**). In this respect, the potential energy surface in Figure 3.9 much more closely resembles that calculated in the context of the neutral model (Figure 3.5) than that obtained from the anionic model (Figure 3.7). Interestingly, the transformation shown in Figure 3.9 also coincides surprisingly well with that obtained by Guo and Himo with their significantly larger model.[25] In particular the C-C bond cleavage from intermediate **4c** is predicted to be very facile so that it is the transition structure between the two intermediates

(**TS:3c→4c**) that becomes the highest energy stationary point on the potential energy surface (in the present study this stationary point, like all others, was fully optimized). The facile C-C bond cleavage is likely due to the stabilization of the leaving formylate radical by the arginine model. This is also reflected in the fact that the product complex (**5c**) is predicted to be lower in energy than the reactant complex (**3c**) as well as in the reduced overall endothermicity with respect to the neutral model.

As alluded to earlier, the intermediate **4c** of the extended model corresponds closely to the product **3b** of the anionic model, with an extended C-C bond and delocalized spin distribution (the analogy here is stronger than between **4a** and **3b**). However, the presence of the methylguanidinium implies that, in contrast to the anionic pathway, there is an energetic gain to be made from fully breaking the C-C bond.

In a manner similar to that observed in the neutral and anionic systems, the small basis set DFT treatment overestimates the energies of the intermediates and transition structures with respect to G3(MP2)-RAD. The large basis set DFT methodology offers some improvement but discrepancies between this and the higher level treatment are, nevertheless, found to be as large as 20 kJ mol^{-1}.

Finally, it is interesting to compare the overall energetic demands calculated in each of the three models, using the G3(MP2)-RAD results. If one takes the separated products as the reference value (as in Figure 3.5, Figure 3.7 and Figure 3.9), the highest energy barriers to be overcome are 65.0, -3.0, and 35.0 kJ mol^{-1}, in the neutral, anionic, and extended models respectively. If, as some authors prefer, one takes the reactant complexes as the reference values, the barriers for the neutral, anionic and extended models correspond to 83.0, 32.0, and 65.0, kJ mol^{-1}, respectively.

3.2.3 CONCLUSION

Small model computational studies of the PFL substrate transformation, both from the literature and presented herein, show that the appearance of the potential energy surface and hence the qualitative description of the mechanism are quite sensitive to the model choice. In particular, the question arises as to how best to treat the protonation state of the carboxylic acid substituent. Indeed, this question is by no means unique to the PFL system and frequently arises in small model computational studies on biological systems. In order to address this question, we chose to use the well-studied PFL example as a case in point. In particular, we wished to establish the most realistic approach by which to truncate a system that contains an arginine-bound carboxylate motif.

Our calculations on the relevant salt-bridged system for PFL show that the radical addition/elimination mechanism proceeds through two intermediates requiring steps best described as C-S bond formation, isomerization, and C-C bond cleavage. Interestingly, the reaction profile for this idealized carboxylate-guanidinum complex is very similar to that calculated by Guo and Himo in their larger model.[25] Calculations on the truncated neutral carboxylic acid system (with the *syn* orientation of the carboxylic proton) give a very similar picture. That is, the reaction proceeds through two intermediates associated with three transition structures, with a reaction profile that is qualitatively very similar to the salt-bridged system. In contrast, truncating the salt-bridge motif to a bare carboxylate anion gives an altogether different qualitative picture. That is, no intermediates are found on the reaction pathway and the C-S bond formation, the isomerization, and the C-C bond cleavage are predicted to occur simultaneously in a single concerted step. In addition, the anionic pathway would be predicted to stop at an intermediate structure with an extended, but not fully cleaved, C-C bond. Both the extended and neutral models disagree with this aspect and predict further energetic benefit is to be derived from completely dissociating this bond.

Based on the reasons outlined above, as well as the quantitative energetics, we would certainly recommend that the neutral carboxylic acid constitutes a much better truncation of an arginine-bound carboxylate than does a bare carboxylate. This conclusion, however, is not based on assuming reduced charge separation in the protein's interior. Rather it stems from a direct comparison with a charge separated state.

The question naturally arises as to whether a truncation is at all necessary. The answer to this question is connected to the performance of DFT on this and related systems. The comparisons provided for each of the systems show that the small basis set DFT results repeatedly deviate from the higher level calculation by 20 kJ mol^{-1} or more. The larger basis set DFT calculations certainly reduce this discrepancy and may even be classified as a good approximation to the G3(MP2)-RAD results. Nevertheless, there are certain quantities (e.g. the TS for C-S bond formation and the overall heat of reaction) where the B3LYP/G3MP2Large results also deviate significantly from the higher-level calculations. As it is not always obvious when these deviations will arise, there is an argument for the continued use of high-level calculations. With this argument, however, comes the need for some kind of truncation.

As many authors have noted, an attractive form of truncation is the use of a layered energy function such as a QM/MM treatment. In this respect, it would be of interest to know whether treating the methylguanidinium fragment molecular mechanically provides a satisfactory solution to the truncation problem. As pointed out, this would need to be done with electrostatic embedding to provide meaningful answers and we are indeed pursuing such solutions. However, in cases where a layered treatment proves inconvenient or impractical, we would recommend that a neutral carboxylic acid (with a *syn* orientation of the carboxylic proton) constitutes a much better truncation of an arginine-bound carboxylate motif than does a bare carboxylate anion.

3.2.4 REFERENCES

1 Ababou, A.; van der Wart, A.; Gogonea, V.; Merz K. M. Jr. *Biophys. Chem*. **2007**, 125, 221.

2 Dixon S. L.; Merz K.M Jr. *Chem. Phys*. **1996**, 104, 6643.

3 Warshel, A.; Levitt, M. J. *Mol. Biol*. **1976**, 103, 227.

4 Singh, U. C.; Kollman, P. A. *J. Comput. Chem*. **1986**, 7, 718.

5 Bash, P. A.; Field, M. J.; Karplus, M. *J. Am. Chem. Soc*. **1987**, 109, 8092.

6 Gao, J.; Xia, X. *Science* **1992**, 258, 631.

7 Bakowies, D.; Thiel, W. *J. Phys. Chem*. **1996**, 100, 10580.

8 Schoneboom, J. C.; Lin, H.; Reuter, N.; Thiel, W.; Cohen, S.; Ogliaro, F.; Shaik, S. *J. Am. Chem. Soc*. **2002**, 124, 8142.

9 Murphy, R. B.; Philipp, D. M.; Friesner, R. A. J. *Comput. Chem*. **2000**, 21, 1442.

10 Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comp*. 2006, 2, 815.

11 Vreven, T.; Morokuma, K.; Farkas, O.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem*. **2003**, 24, 760-769.

12 Tantillo, D. J.; Chen, J.; Houk, K. N. *Curr. Opin. Chem. Biol*. **1998**, 2, 743–750.

13 Himo, F.; Siegbahn, P.E.M. *Chem. Rev*. **2003**, 103, 2421; Siegbahn, P. E. M.; Borowski, T. *Acc. Chem. Res*. **2006**, 39, 729.

14 Ban, F. Q.; Rankin, K. N.; Gauld, J. W.; Boyd, R. J. *Theor. Chem. Acc*. **2002**, 108, 1.

15 Mohr, M.; Zipse, H. *Chem. Eur. J*. **1999**, 5, 3046; Smith, D. M.; Buckel, W.; Zipse, H. *Angew. Chem. Int. Ed*. **2003**, 42, 1867-1870.

16 Borowski, T.; Broclawik, E.; Schofield, C. J.; Siegbahn, P. E. M.; *J. Comput. Chem*. **2006**, 740; Borowski, T.; de Marothy, S.; Broclawik, E.; Schofield, C. J.; Siegbahn P. E. M. *Biochemistry* **2007**, 46, 3682.

17 Smith, D. M.; Golding, B. T.; Radom, L. *J. Am. Chem. Soc.* **1999**, 121, 1383; Smith, D. M.; Golding, B. T.; Radom, L. *J. Am. Chem. Soc*. **2001**, 123, 1664.; Wetmore, S. D.; Smith, D. M.; Bennett, J. T.; Radom, L. *J. Am. Chem. Soc*. **2002**, 124, 14054; Sandala, G. M.; Smith, D. M.; Radom, L. *J. Am. Chem. Soc*. **2006**, 16004.

18 Siegbahn, P. E. M. *J. Phys. Chem*. **1996**, 100, 14672.; Siegbahn, P. E. M. *J. Am. Chem. Soc*. **1998**, 120, 8417.

19 Himo, F. *J. Phys. Chem. B* **2002**, 106, 7688.

20 Siegbahn, P. E. M. *Struc. Chem*. **1995**, 6, 271.

21 Pelmenschikov, V.; Cho, K. B.; Siegbahn, P. E. M. J. *Comput. Chem*. **2004**, 25, 311.

22 Fernandes, P. A.; Eriksson, L. A.; Ramos, M. J. *Theor. Chem. Acc*. **2002**, 108, 352.

23 Himo, F.; Eriksson, L.A. *J. Am. Chem. Soc*. **1998**, 120, 11449.

24 Lucas, M. F.; Fernandes, P. A.; Eriksson, L. A.; Ramos, M. J. *J. Phys. Chem. B* **2003**, 107, 5751.

25 Guo, J.-D.; Himo, F. *J. Phys. Chem. B* **2004**, 108, 15347.

26 Knappe, J.; Blaschkowski, H. P.; Gröbner, P.; Schmitt, T. *Eur. J. Biochem.* **1974**, 50, 253.

27 Knappe, J.; Neugebauer, F.A.; Blaschkowski, H.P.; Gänzler, M. *Proc. Natl. Acad. Sci. USA* **1984**, 81, 1332.

28 Himo, F. *Chem. Phys. Lett.* **2000**, 328, 270.

29 Knappe, J.; Elbert, S.; Frey., M.; Wagner, A. F. V. *Biochem. Soc. Trans*. **1993**, 21, 731.

30 Becker, A.; Kabsch, W. *J. Biol. Chem*. **2002**, 277, 400036.

31 Becker, A.; Fritz,-Wolf, K.; Kabsch, W.; Schultz, S.; Wagner, A. F. V.; Knapper *J. Nat. Struc. Biol*. **1999**, 6, 969.

32 Henry, D.J.; Sullivan, M.B.; Radom, L. *J. Chem. Phys*. **2003**, 118, 4849.

33 Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, 110, 4703.

34 Gaussian 03, Revision C.02, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A.; Gaussian, Inc., Wallingford CT, 2004.

35 Knowles, P. J ; Hampel, C.; Werner, H.-J. *J. Chem. Phys*. **1993**, 99, 5219; Erratum: *J. Chem. Phys.* **2000**, 112, 3106.

36 Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; T. Thorsteinsson, MOLPRO, version 2006.1, a package of ab initio programs,

37 Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Chem. Phys. Lett. 1997, 270, 419; Johnson, B. G.; Gonzales, C. A.; Gill, P. W. M.; Pople, J. A. *Chem. Phys. Lett*. **1994**, 221, 100.

38 Melo, A.; Ramos, M. J. *Int. J. Quant. Chem*. **1999**, 72, 157-176.

39 Melo, A.; Ramos, M. J.; Floriano, W.B.; Gomes, J. A. N. F.; Leao, J. F. R, Magalhaes, A. L., Maigret, B., Nascimento, M. C.; Reuter, N. *J. Mol. Struc. (Theochem)* **1999**, 463, 81.

40 Barill, X.; Aleman, C.; Orozco, M.; Luque, F. J. *Proteins* **1998**, 32, 67.

41 Kinsel, G. R.; Zhao,O.; Narayanasamy, J.; Yassin, F.; Dias, H. V. R.; Niesner, B.;Prater, K.; St. Marie, C.; Ly, L.; Marynick, D. S. *J. Phys. Chem. A* **2004**, 108, 3153.

42 Graul, S. T.; Schnute, M. E.; Squires, R. R. *Int. J. Mass. Spectrom. Ion. Proc*. **1990**, 96, 181.

43 Wenthold, P. G.; Squires, R. R. *J. Am. Chem. Soc*. **1994**, 116, 11890.

44 Cumming, J. B.; Kebarle, P. *Can. J. Chem* . **1978**, 56, 1.

3.3 A Compound QM/MM Procedure: Comparative Performance on the PFL System

Beginning with the seminal paper from Warshel and Levitt,[1] the concept of combining a quantum mechanical treatment with a molecular mechanical description (QM/MM) has continued to grow in importance.[2] The idea itself is conceptually simple but the details of the implementation are, however, complex, with many issues needing to be addressed.[3] While the QM/MM methodology was conceived to take the best aspects from both the QM and MM worlds, the practical complexity involved has led to the oft heard comment that one is actually left with the worst of both worlds, requiring sound knowledge and experience with both QM and MM techniques, and beyond, to ensure the generation of meaningful results.

The difficulties with hybrid QM/MM methods are numerous and include aspects like the boundary problem, which involves having to chose between using single link atoms,[4] double link atoms,[5] connection atoms,[6] pseudobonds,[7] quantum capping potentials,[8] strongly localized MOs,[9,10] extremely localized MOs,[11] or generalized hybrid orbitals.[12] One should also contend with issues such as which type of embedding to use,[13] how to cope with the MM charges close to the region boundary,[14] and whether to include a polarizable MM region.[15]

One issue that has received less attention in the QM/MM field is the adequacy of the QM description itself. Because of the sampling problem involved in the treatment of large systems, recent trends have moved towards what might normally be considered inaccurate QM methods. Examples include the semi-empirical,[16,17] empirical valence bond,[18] and self-consistent charge density functional tight-binding (SCC-DFTB)[19] approaches. Traditional density functional theory continues to prove a popular choice for the QM treatment in QM/MM models.[20,21] While all of these methods certainly have their place, it is well-known in the ab initio community that none of them are reliable enough able to supply chemical accuracy for reaction energetics. As recently outlined in an important contribution from the groups of Mullholand, Thiel, Shütz, and Werner[22] there is, in the current climate,[23] a clear need for higher accuracy QM methods in the QM/MM context.[24]

One of the most widespread and successful approaches for achieving chemical accuracy in the non-biological literature has been the Gaussian-*n* (G*n*) series of model chemistries, with *n* recently reaching a value of 4.[25] The predecessor G3 series of methods have proven to be quite successful[26] with the G3X model chemistry, in particular, displaying a MAD from experiment of 1.01 kcal mol^{-1} when applied to the G3/05 test set comprising 454 energies.[27] For comparison,

the MAD from experiment of the B3LYP methodology, for the same test set, was 4.14 kcal/mol.[27]

On the other hand, one of the most widespread and successful hybrid treatments has been the multilayered ONIOM method of Morokuma and co-workers.[28] Although often used to combine different QM layers,[29] the QM/MM variant has also proven a powerful tool.[30] In particular, the recent incorporation of electrostatic embedding[31] has rendered the formalism competitive with any of the alternatives. Importantly, due to its inherent flexibility, the ONIOM approach is well-suited for the incorporation of compound methods. Indeed, the combination of G*n* style methods in the ONIOM(QM:QM) context, such as IMOMO-G2MS,[32] ONIOM-G3B3,[33] and variants of ONIOM(G3:MP2) and ONIOM(G3:DFT)[34] have proven to be convenient methods for achieving chemical accuracy in medium-sized systems.

Indeed the ONIOM(G3:B3LYP) method has been found to be an acceptable method for achieving chemical accuracy in medium-sized systems.

We have therefore chosen to combine the G3 and ONIOM(QM:MM) methodologies as a transparent and convenient way of incorporating a chemical accuracy QM treatment into the QM/MM context. The specific variant presented here combines the G3(MP2)-RAD model chemistry[35] with the AMBER[36] molecular mechanics force field, and is thus denoted ONIOM(G3(MP2)-RAD:AMBER). Our choice of the G3(MP2)-RAD method stems from the fact that it has been designed specifically to perform well for radical-containing systems and our primary interest is in radical enzymes. For closed-shell systems, the method is very similar to the G3(MP2) and G3X(MP2) models.[37] Naturally, the ONIOM method is sufficiently flexible that employing any G*n* variant in an ONIOM(G*n*:AMBER) calculations is readily achievable.

As the method is designed to be applied to biological reaction mechanisms, we have decided to validate it in that context. Due to the sometimes ambiguous interpretation of experimental values for such applications, we have chosen to perform the initial validation on a small model system, where it is possible to obtain benchmarks with a full QM treatment. For that purpose we have chosen the substrate transformation Pyruvate Formate-Lyase.

Pyruvate formate-lyase (PFL) is a key enzyme of anaerobic glucose metabolism in *E.coli* and other microorganisms, catalyzing the CoA-dependant reversible cleavage of pyruvate into acetyl-CoA and formate:[38]

$$\text{ES-H + pyruvate} \rightleftharpoons \text{ES-acetyl + formate}$$

$$\text{ES-acetyl + CoA} \rightleftharpoons \text{ES-H + acetyl-CoA}$$

In its active form, PFL stores an unpaired electron spin on the Cα atom of glycine 734.[39,40] The radical is generated through the action of a specific activating enzyme (PFL-activase), after the active site of PFL has been occupied.[41]

Several X-Ray crystal structures of the catalytic subunit (non-radical form) of PFL have been determined. In addition to those obtained with the substrate pyruvate,[42] two structures have been solved with oxamate in the active site.[43]

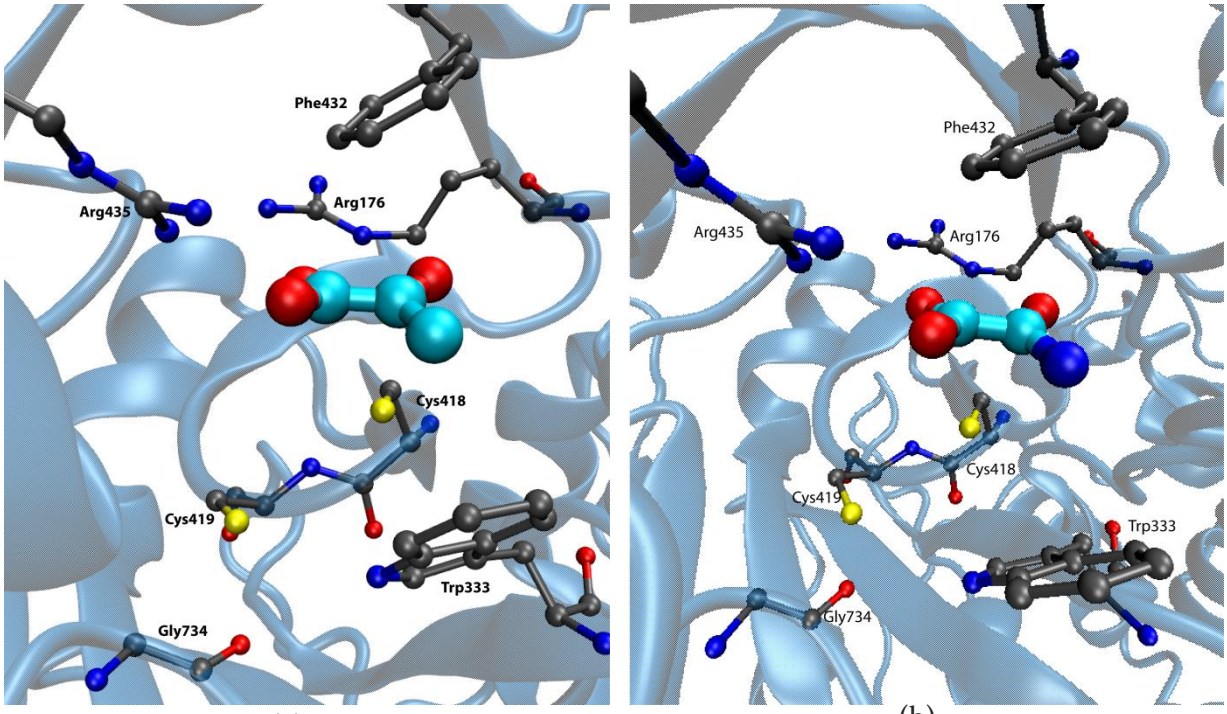
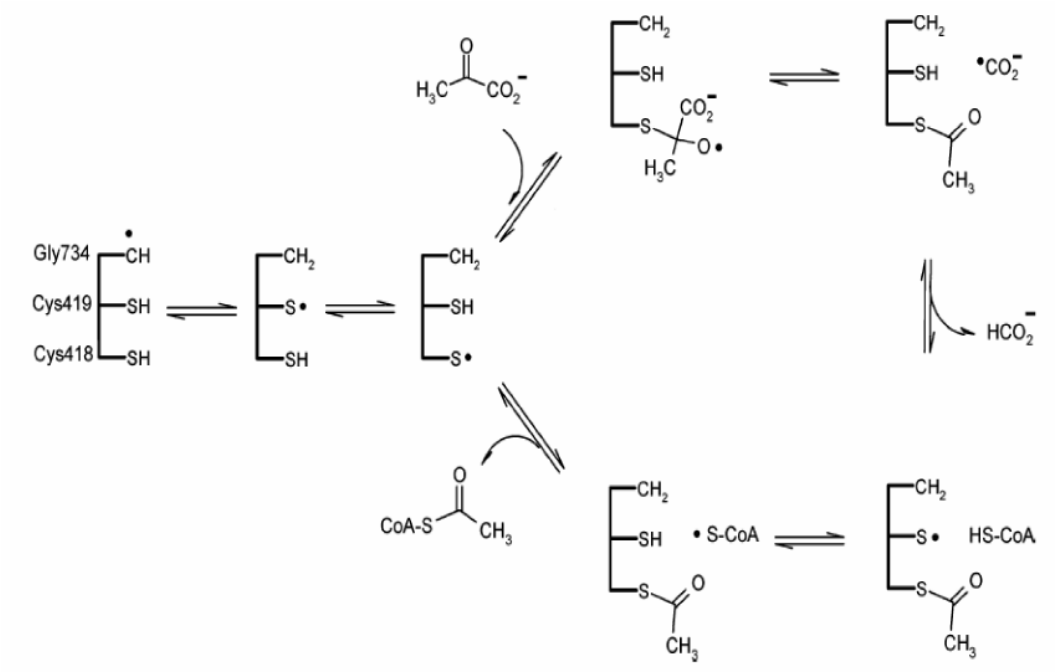


Figure 3.10 Crystallographic structure of the active site with bound pyruvate (a) and oxamate (b).

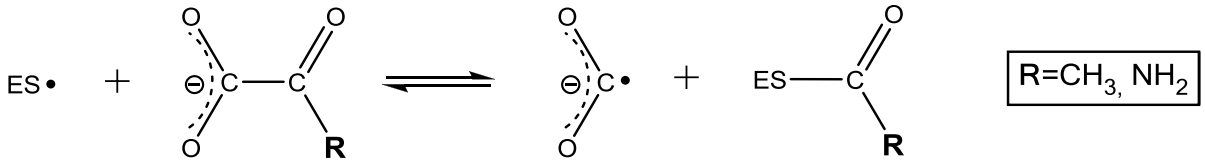
Oxamate is an isosteric and chemically inert analogue of the substrate pyruvate and binds to active site in a mode largely similar to that of pyruvate (Figure 3.10). Binding of oxamate triggers generation of the glycyl radical, but oxamate undergoes no further turnover. The apparent binding affinity of oxamate to the non-radical form of PFL around K_D=2mM is somewhat smaller than that measured for pyruvate of K_D=0.1 mM.[42] For the substrate turnover catalyzed by the radical form of the enzyme, a K_M value of 2 mM has been measured for pyruvate, while oxamate appears to be a competitive inhibitor with relatively weak binding affinity (K_I > 2mM).[38]

The currently accepted mechanism (Scheme 3-8), proposed by Knappe an co-workers in 1999,[42] was derived with the help of theoretical calculations[44] based on an earlier mechanism proposed by Kozarich *et al.*[45]



Scheme 3-8 Currently accepted substrate mechanism of PFL, proposed by Knappe *et al.*[42]

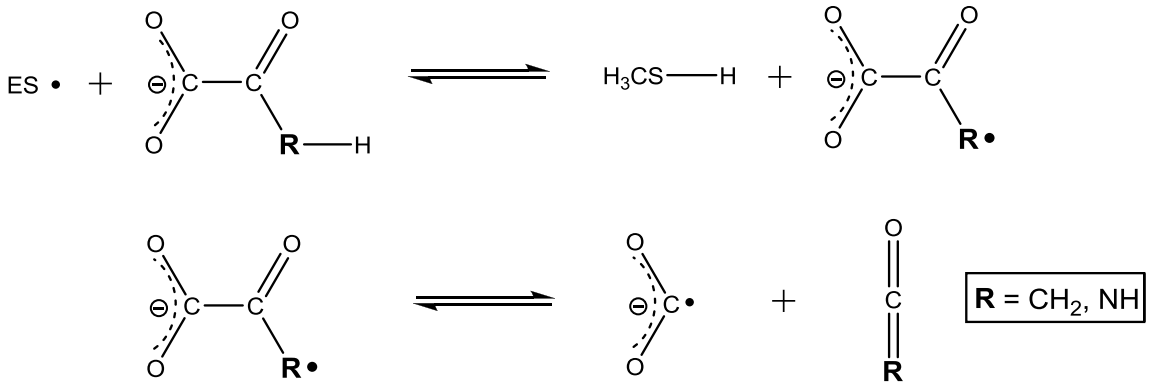
According to this proposal (Scheme 3-8), following activation, the radical centre is shuttled from Gly734, via Cys419, to Cys418. The subsequent addition of the ideally positioned (Figure 3.10) thiyl radical thus formed to C2 of the substrate results in the formation of a tetrahedral intermediate, which collapses into the formyl radical and acetylated Cys418. The later stages of the mechanism involve the quenching of the formyl radical and eventual acetyl transfer to CoA. For our current purposes, we are concerned only with the initial transformation of the substrate (shown in Scheme 3-9 with R=CH_3).



Scheme 3-9 Transformation of pyruvate/oxamate into acetylated enzyme and formyl radical, catalyzed by PFL.

In order to investigate the origin of the inhibitory effect of oxamate, as well as to have an additional system with which to calibrate the ONIOM(G3(MP2)-RAD:AMBER) method, we have chosen to determine the energetics of the consensus mechanism for this compound as well (R=NH_2, Scheme 3-9).

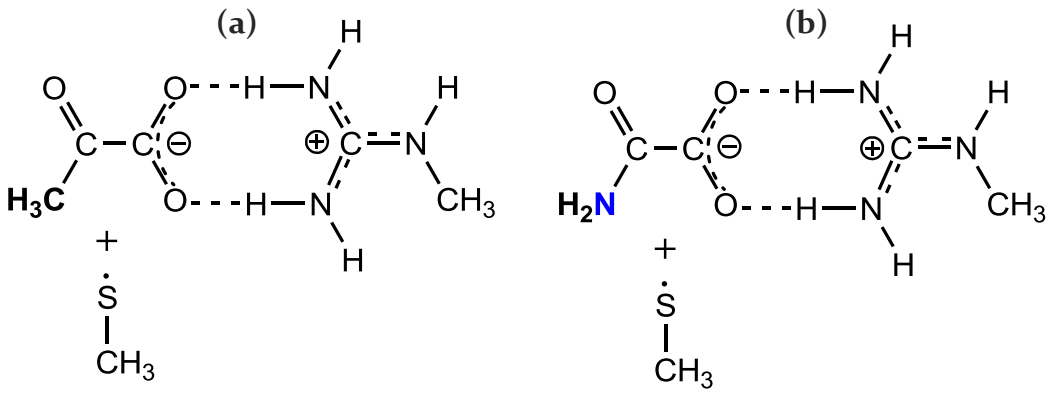
Using a similar motivation, we have also added the characterization of an alternative mechanism, for both the substrate and the inhibitor, to our validation set. The alternative mechanism involves initial H-atom abstraction from the substrate (or inhibitor). Although this mechanism is not known to play any role in the PFL catalyzed substrate reaction, it is of interest in the wider picture of SAM-dependent glycyl radical enzymes. Interestingly, PFL serves as a paradigmatic example of the glycyl radical enzymes,[46,47] and, indirectly, of the radical-SAM super-family as well.[48,49] However, the majority of the other known enzymes in these classes activate their substrates precisely by abstracting a hydrogen atom from a C-H bond. Examples include class III ribonucleotide-reductase (RNR),[50] benzylsuccinate-synthase (BSS)[51] and the coenzyme-B_{12}-independent glycerol dehydratase (GDH).[52] The uniqueness of the radical addition mechanism of PFL, combined with the considerable structural homology existing between PFL and the catalytically active subunits of both class III RNR[53] and GDH,[52] warrants the consideration of H-abstraction in the context of PFL. In particular, it is interesting to ask how facile the potential H-abstraction is and, furthermore, if such an activation process can actually facilitate the cleavage of the C-C bond of the substrate (or inhibitor, Scheme 3-10).



Scheme 3-10 Removal of hydrogen from the substrate by the cysteinyl radical and subsequent C-C bond cleavage.

Consideration of both mechanistic alternatives, for the substrate pyruvate and the inhibitor oxamate, implies the characterization of four different mechanisms. As mentioned earlier, we have evaluated these, in the first place, on small model systems where we are able to compare the ONIOM(G3(MP2)-RAD:AMBER) results directly with the pure QM G3(MP2)-RAD ones. The actual model system chosen is shown below in Scheme 3-11 and consists of the methylthiyl radical reacting with a complex between pyruvate (oxamate) and the methylguanidinium ion. The latter moiety is intended to model a nearby arginine residue.

Other small-model computational studies have also proven useful in understanding the mechanism of PFL,[44,54,55] an aspect which we have discussed in detail in a recent publication.[56] In that work, we used the model system shown in Scheme 3-11 to determine the best truncation of an arginine-bound carboxylate motif. As we present a treatment here in which the methylguanidinuim is represented as an MM fragment, a further utility of our results is to test the resulting QM/MM approach for treating this common enzymatic binding motif.



Scheme 3-11 The model consists of pyruvate (a) and oxamate (b) in complex with protonated methylguanidinum and thiyl radical instead of cysteine.

3.3.1 Computational Details

For the pure QM calculations, each stationary point was fully optimized and its Hessian evaluated at the B3LYP/6-31+G(d) level of theory. The vibrational frequencies obtained from the latter calculations were used to calculate an unscaled zero-point vibrational energy correction (ZPVE). In order to verify the connectivity of each potential energy surface, the intrinsic reaction coordinate (IRC) was followed (in both directions) from every transition structure described. The reactant and product complexes in each case were obtained by rigorous optimization of the end point from the relevant IRC. These complexes were not, however, subjected to exhaustive conformational searches to locate the lowest energy complex.[56]

Improved relative energies were obtained using the G3(MP2)-RAD methodology. This method, based on G3(MP2),[37] was developed to obtain more reliable predictions in energy for

radical species.[35] For reaction mechanisms such as those currently under investigation, the relevant quantity is the relative energy of two species:

$$\Delta E[G3(MP2)\text{-}RAD] = E[RHF\text{-}UCCSD(T)/6\text{-}31G(d)] + E[ROMP2/G3MP2Large] - E[ROMP2/6\text{-}31G(d)] + \Delta(ZPVE)$$

The QM/MM approach is almost exactly analogous to that described above. The difference is that the methylguanidinium fragment was treated molecular mechanically within the framework of the AMBER force field (see below for details). Geometries were thus obtained with the ONIOM(B3-LYP/6-31+G(d):AMBER) hybrid method. Due to the link-atom implementation in ONIOM, it is possible to calculate a full Hessian with 3N-6 vibrational degrees of freedom. This allows the normal calculation of vibrational frequencies, ZPVEs and IRCs (at least for small models), which we performed with the ONIOM(B3LYP/6-31+G(d):AMBER) hybrid for all relevant stationary points.

Improved QM/MM relative energies were obtained with the ONIOM(G3(MP2)-RAD:AMBER) combination. The relevant relative energy is defined as:

$$\Delta E[ONIOM(G3(MP2)\text{-}RAD:AMBER)] = E[ONIOM(RHF\text{-}UCCSD(T)/6\text{-}31G(d):AMBER)] + E[ONIOM(ROMP2/G3MP2Large:AMBER)] - E[ONIOM(ROMP2/6\text{-}31G(d):AMBER)] + \Delta(ZPVE)$$

All computations were performed using Gaussian03[57] except for the RHF-UCCSD(T) calculations,[58] which were carried out with Molpro (Version 2006.1).[59] As all ONIOM(QM:MM) calculations were performed with electrostatic embedding,[31] the ONIOM(RHF-UCCSD(T)/6-31G(d):AMBER) results included RHF-UCCSD(T) values polarized by the appropriate lattice of point charges.

The point charges for use in both the Molpro and Gaussian were obtained using RESP (restrained electrostatic potential) fitting of the HF/6-31G(d) ESP[60,61] of the methylguanidinium fragment. MM atom types for this fragment were assigned on the basis of the AMBER94 force field.[62] The Van der Waals (VdW) parameters for the QM atoms were assigned values specifically adjusted for the B3LYP-6-31+G(d):AMBER combination.[63] Our tests revealed, however, a negligible impact on the potential energy surfaces, when compared to the use of the standard AMBER94 VdW parameters.

All energies presented are calculated at temperature of 0 K. The spin densities were obtained by Mulliken population analysis at the level of theory used for geometry optimization.

3.3.2 Results and Discussion

3.3.2.1 Radical Addition to Pyruvate

The reaction profile for the addition of the methyl thiyl radical is shown in Figure 3.11. On the potential energy surface for this reaction two intermediates have been found and with three corresponding transition structures interconnecting them.[56]

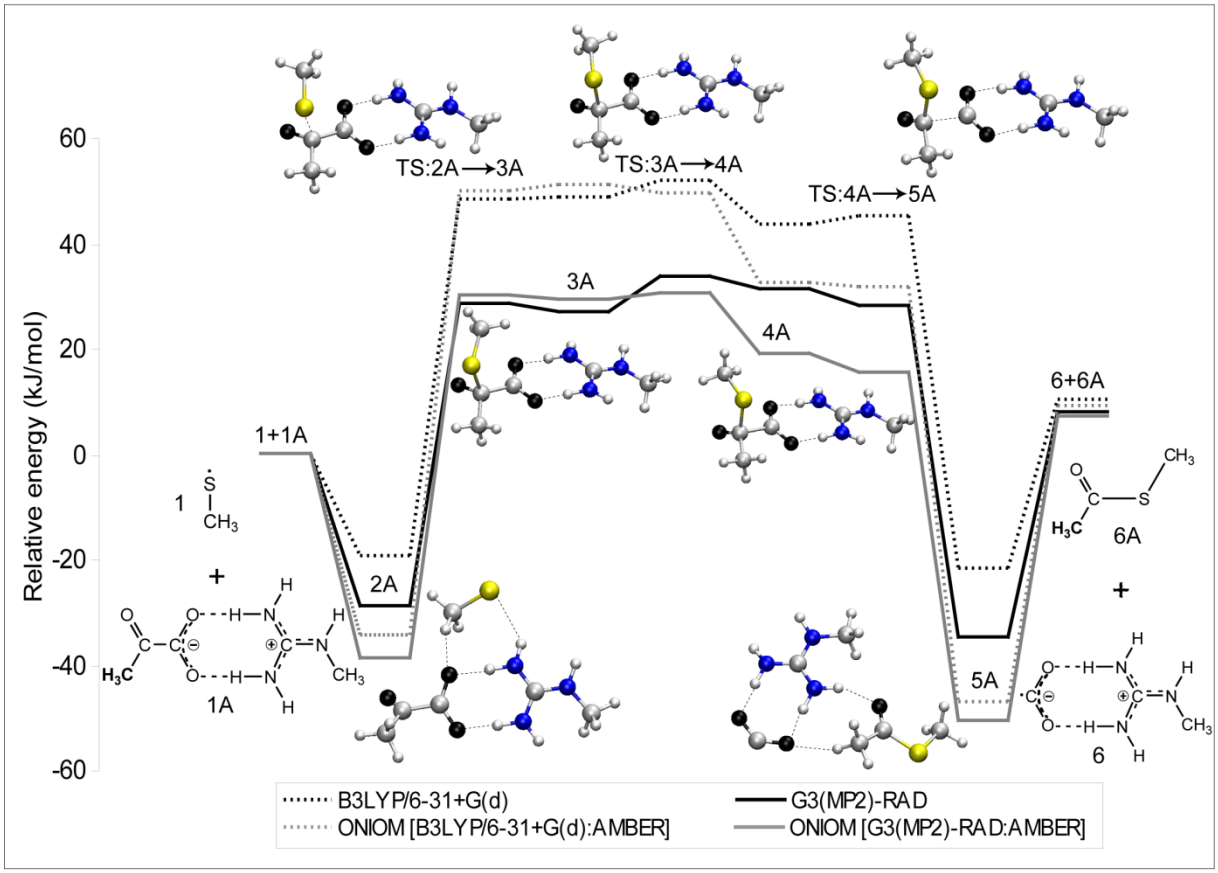


Figure 3.11 A comparison of QM and ONIOM results for addition of methylthiyl radical to carbonyl group of oxamate at oK (geometries presented were optimized at B3LYP/6-31+G(d) level of theory).

The highest energy barrier is associated with the formation of the S-C bond (TS:2A→3A), which results with formation of the first intermediate 3A. A transition from one shallow minimum (3A) to the other (4A) goes over low energy barrier (TS:3A→4A). Once the state 4A has been reached, C1-C2 bond is easily broken. The spin density over the course of the reaction changes as expected in that the radical character is transferred from sulphur to the carboxyl

group as the fragmentation proceeds. Minor delocalization of the spin density to methylguanidinium is present, especially in case of the formylate radical.

Comparison of the results obtained in previous studies of PFL mechanism has shown that having two intermediates in the reaction pathway is very likely to be the right reaction mechanism for this addition-elimination process.[56]

The relative energies shown in Figure 3.11 are obtained by geometry optimization at B3LYP/6-31+G(d) level of theory, where improved energies are calculated using G3(MP2)-RAD method. Black curve represents pure QM methods and grey curves are results of QM/MM (ONIOM) approach. The Figure 3.11 clearly shows that ONIOM method has the ability of reproducing the potential energy surface calculated with QM methods. The geometries optimized using ONIOM methods have properties very similar to analogous DFT optimized geometries. The most notable discrepancy has been found for distance between carboxylate group and methylguanidinium in salt bridge, where ONIOM values are in average 0.2 Å longer than DFT counterparts. The distribution of the spin density is very much alike the one calculated with pure DFT methods, only without the possibility of spin delocalization onto MM part of the system, i.e. methylguanidinium.

Comparison of the energies calculated with pure QM methods and combined ONIOM techniques shows that the biggest disagreement is associated with complexes of reactants and products. In other points, the agreement is rather reasonable for the energies calculated at given level of theory.

3.3.2.2 Radical Addition to Oxamate

For the analogous reaction with oxamate, on the calculated potential energy surface (Figure 3.12) no stable intermediates have been found, unlike for the reaction with pyruvate.

According to the calculations, this reaction is concerted and the S-C bond is formed and C1-C2 bond is broken in a concerted fashion. The energy barrier for that transformation has a value almost twice as high as the one required for the fragmentation of pyruvate. The addition of methylthiyl radical to oxamate is around 25 kJ mol^{-1} more endothermic than the same reaction with pyruvate. This information points to the conclusion that the main reason for inhibitory effect of oxamate lies in reaction pathway that goes over high-energy transition state, together with higher reaction endothermicity compared to reaction with pyruvate.

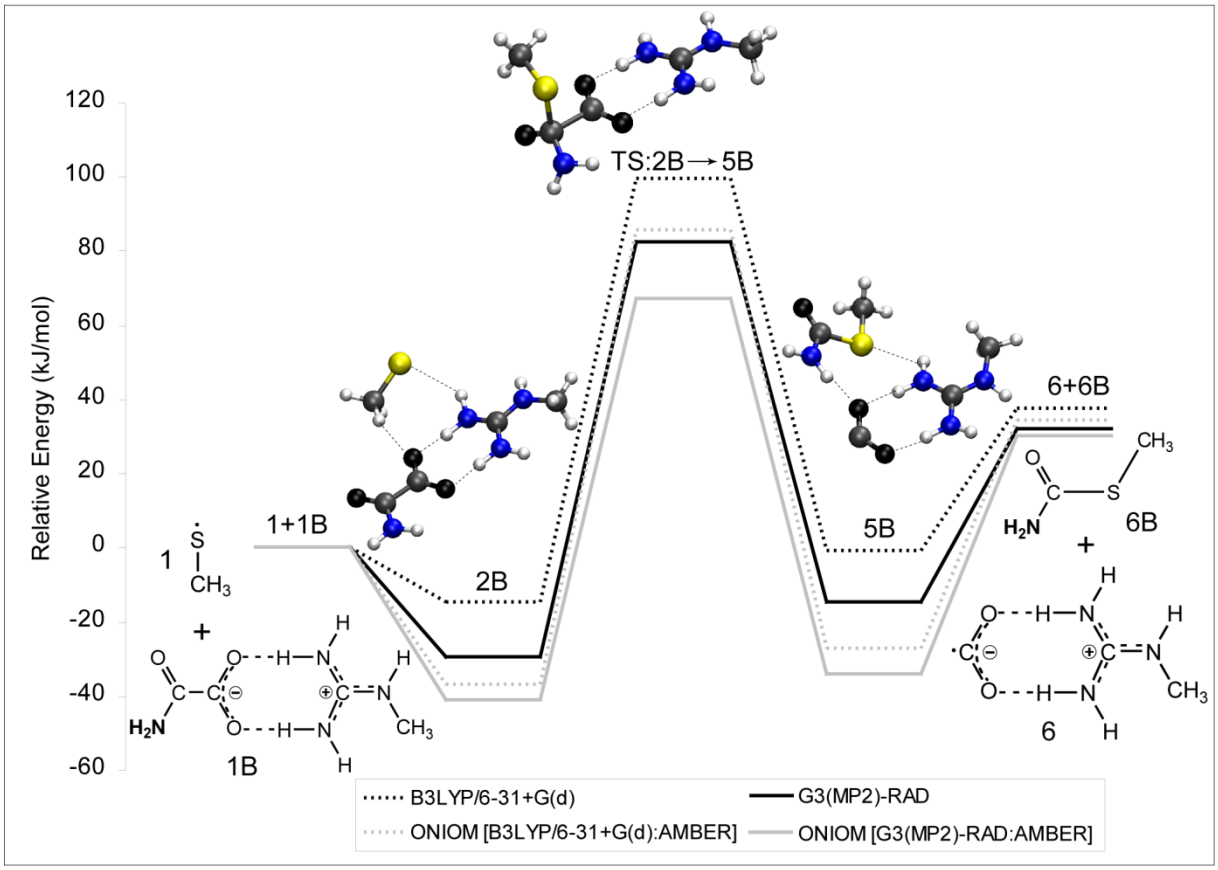


Figure 3.12 A comparison of QM and ONIOM results for addition of methylthiyl radical to carbonyl group of oxamate at oK (geometries presented were optimized at B3LYP/6-31+G(d) level of theory).

Almost half of the spin density in transition structure located on carboxyl group, while the other half is almost equally distributed between sulphur and carbonyl oxygen. Certain amount of spin density is transferred to methylguanidinium, as already established in case of pyruvate.

When it comes to energetics, ONIOM results (grey) follow pure QM results (black), except ONIOM tends to give lower relative energies than QM methods (Figure 3.12). This difference is the most obvious for complexes of the reactants and products.

As in a case of pyruvate fragmentation, geometries optimized with ONIOM are almost identical to those optimized with pure DFT method, except for the complex of the products. ONIOM geometry is slightly more bent than its DFT counterpart. ONIOM Mulliken spin analysis closely resembles to the spin density distribution provided by DFT analysis, with restriction of spin location only on the QM part of the system.

3.3.2.3 Hydrogen Abstraction from Pyruvate

As mentioned earlier, in mechanisms commonly employed by the gylcyl radical enzymes (or radical enzymes in general) substrate is usually activated by abstracting hydrogen from it. In order to understand the reason of unusual behaviour of PFL in that context, this alternative mechanism has been studied for both pyruvate and oxamate (Scheme 3-10).

The first step in this mechanism is hydrogen abstraction from pyruvate (Figure 3.13). This single-step reaction goes over transition structure that lies about 60 kJ mol^{-1} higher in energy than free reactants. In this endothermic reaction the final product is the pyruvylate radical, which, as expected, is stabilized by the partial delocalization of the unpaired electron from C3 (ρ_α(C3)=0.77) into neighbouring carbonyl group (ρ_α(O3)=0.23).

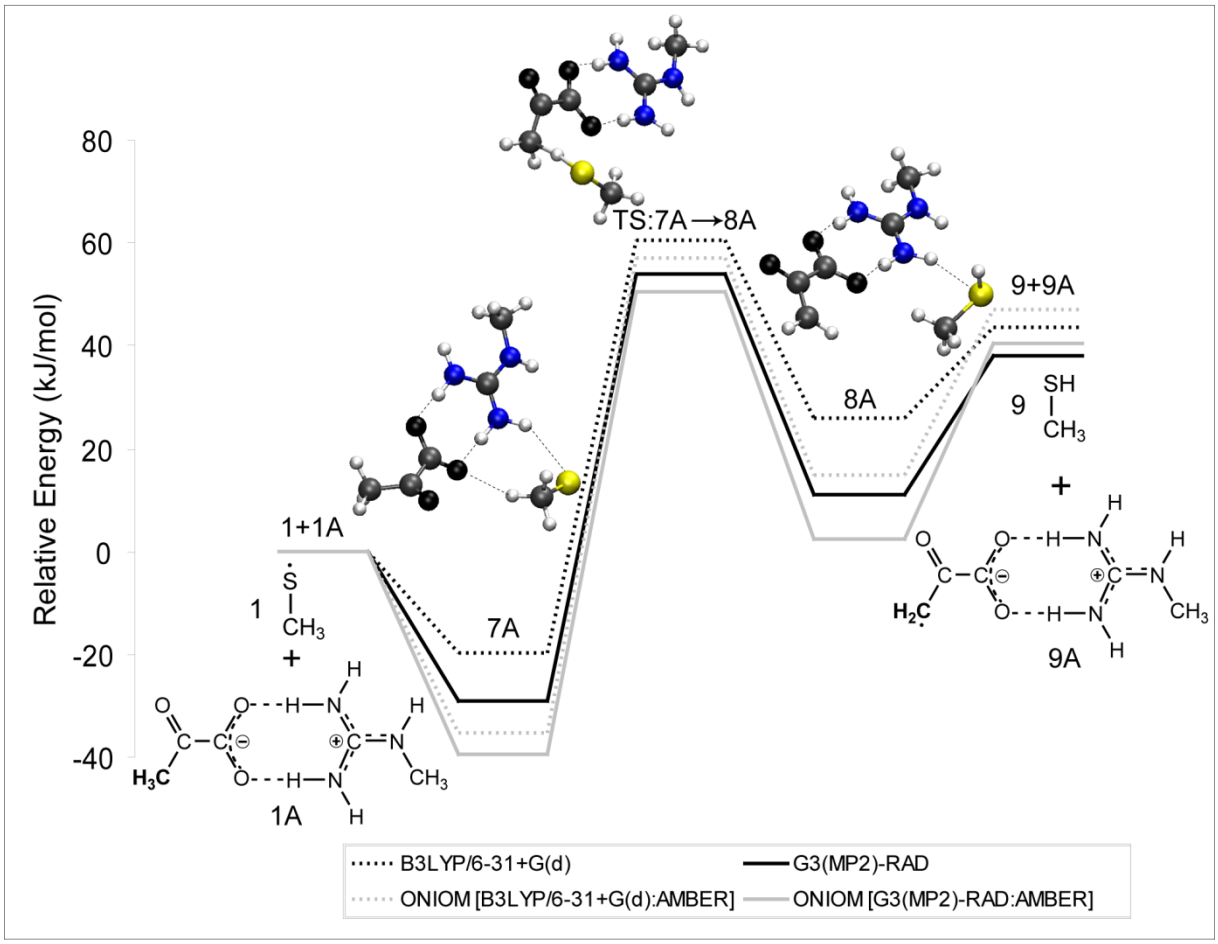


Figure 3.13 A comparison of QM and ONIOM results for hydrogen abstraction from pyruvate by thiyl radical at 0 K (geometries presented were optimized at B3LYP/6-31+G(d) level of theory).

The removal of hydrogen from pyruvate by thiyl radical described with ONIOM results with locating the equivalent stationary points as with QM methods, but ONIOM gives lower

energies than analogous QM methods, as observed in previous examples. Again, the biggest disagreement is associated with complexes.

The reactants and products in complexes optimized with ONIOM are slightly closer to each other than in corresponding DFT structures (7A, 8A). The Mulliken analysis results with similar spin density distribution at both levels of theory, except that the spin density is restricted only to QM part in ONIOM calculations.

Compared to the addition-elimination reaction of pyruvate, hydrogen abstraction has a higher energy barrier (about 10 kJ mol^{-1}) and more positive reaction enthalpy by 30 kJ mol^{-1}. Although the hydrogen abstraction is energetically less favourable than the addition of the thiyl radical to pyruvate, this energy difference is not so drastic to justify PFL's preference of the addition over the abstraction, at least in the gas phase where these reactions were modelled. When substrate is modelled with pyruvic acid only, calculations done on such a model show that the energy barriers heights for both the addition-elimination process and hydrogen abstraction have similar values. When complexed model is used, the reaction pathway for the addition reaction is lowered in energy, but the same effect hasn't been observed for the reaction of hydrogen abstraction. This observation leads towards the conclusion that the interactions of substrate with its environment prefer the addition reaction by stabilizing the reaction pathway, while this stabilization doesn't happen for hydrogen abstraction. Except this stabilizing effect, the protein environment has another important role in governing the possible pathways by imposing sterical conditions. Namely, cysteinyl residue is positioned almost ideally for radical attack on carbonyl group of the substrate.

The next step in the alternative mechanism is fragmentation of the radical formed by hydrogen abstraction from substrate (Scheme 3-10). From the reaction profile of conversion of pyruvylate radical into formylate and ketene (Figure 3.14) it can be seen that this step is highly endothermic. In addition, this process goes over the high-energy transition state.

As it is visible from Figure 3.14, ONIOM method successfully reproduces the potential energy surface, but the difference in QM and ONIOM energies is unusually high for the transition state, when compared to the reactions described earlier. This discrepancy is probably due to difference in C1-C2 bond length observed between DFT and ONIOM optimized geometries for TS: 9A→10A. It seems that C1-C2 bond in the ONIOM geometry is 0.24 Å longer than the same bond in the DFT analogue. This leads to the conclusion that the ONIOM geometry is a late transition structure, leaned more to the products side than the DFT optimized TS. The result is significant difference in energy of the structures compared.

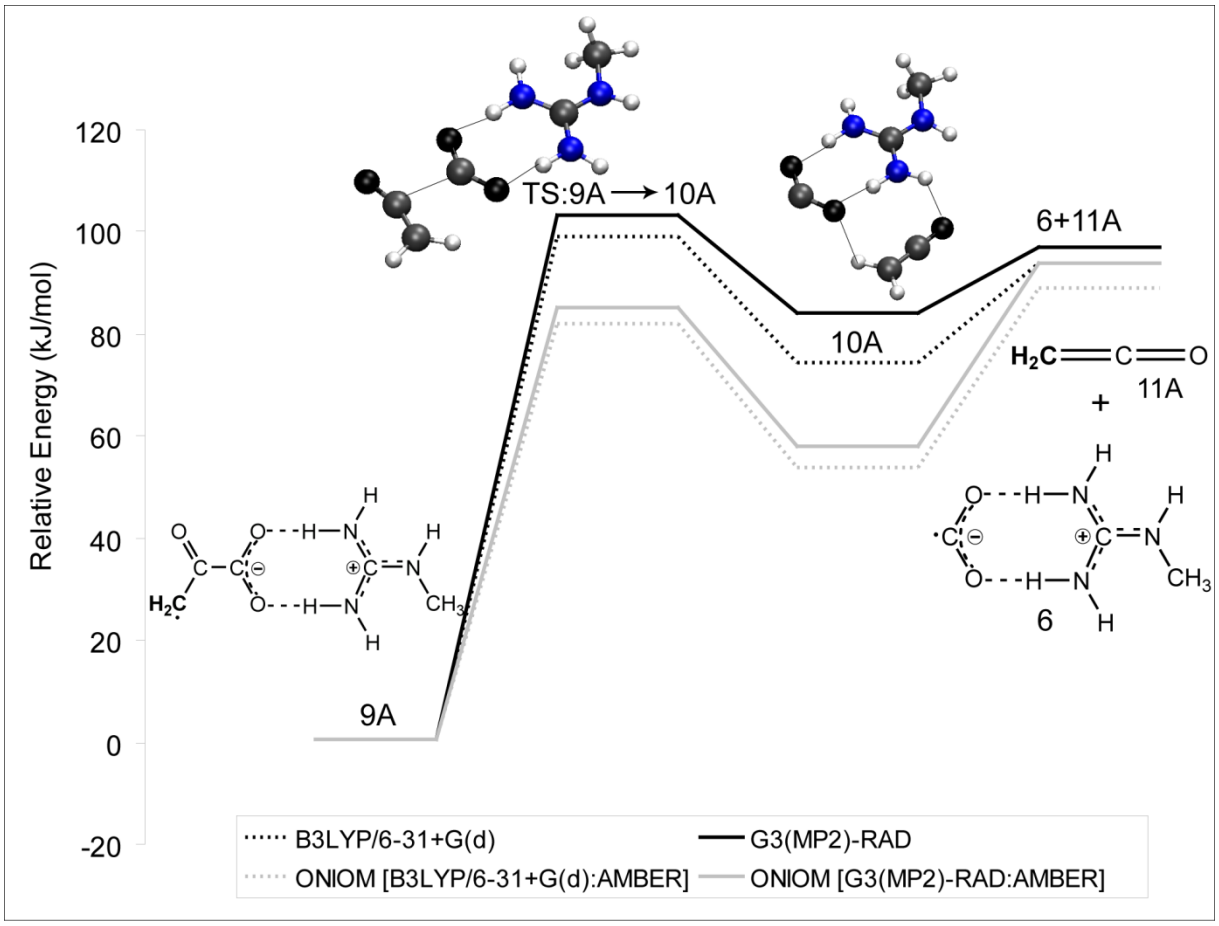


Figure 3.14 A comparison of QM and ONIOM results for fragmentation of pyruvylate radical at oK (geometries presented were optimized at B3LYP/6-31+G(d) level of theory).

While the step of pyruvylate formation still has reasonable chances to take place, fragmentation of radical formed requires high energy input to proceed. When both steps are summed together, the overall reaction mechanism is not likely to occur, both from kinetic and thermodynamic point of view.

3.3.2.4 Hydrogen Abstraction from Oxamate

The calculated potential energy surface for the abstraction of hydrogen atom from amino group of oxamate is shown in Figure 3.15. This reaction is energetically the least favourable of all of the examined processes.

A generation of oxaminylate radical has an energy barrier height similar to the one found for the addition-elimination process, but it is markedly more endothermic reaction. As stated earlier, high energy barrier is the most probable cause of inhibitory behaviour of oxamate. The

formation of the oxaminylate radical is process unfavourable from both kinetic and thermodynamic point of view.

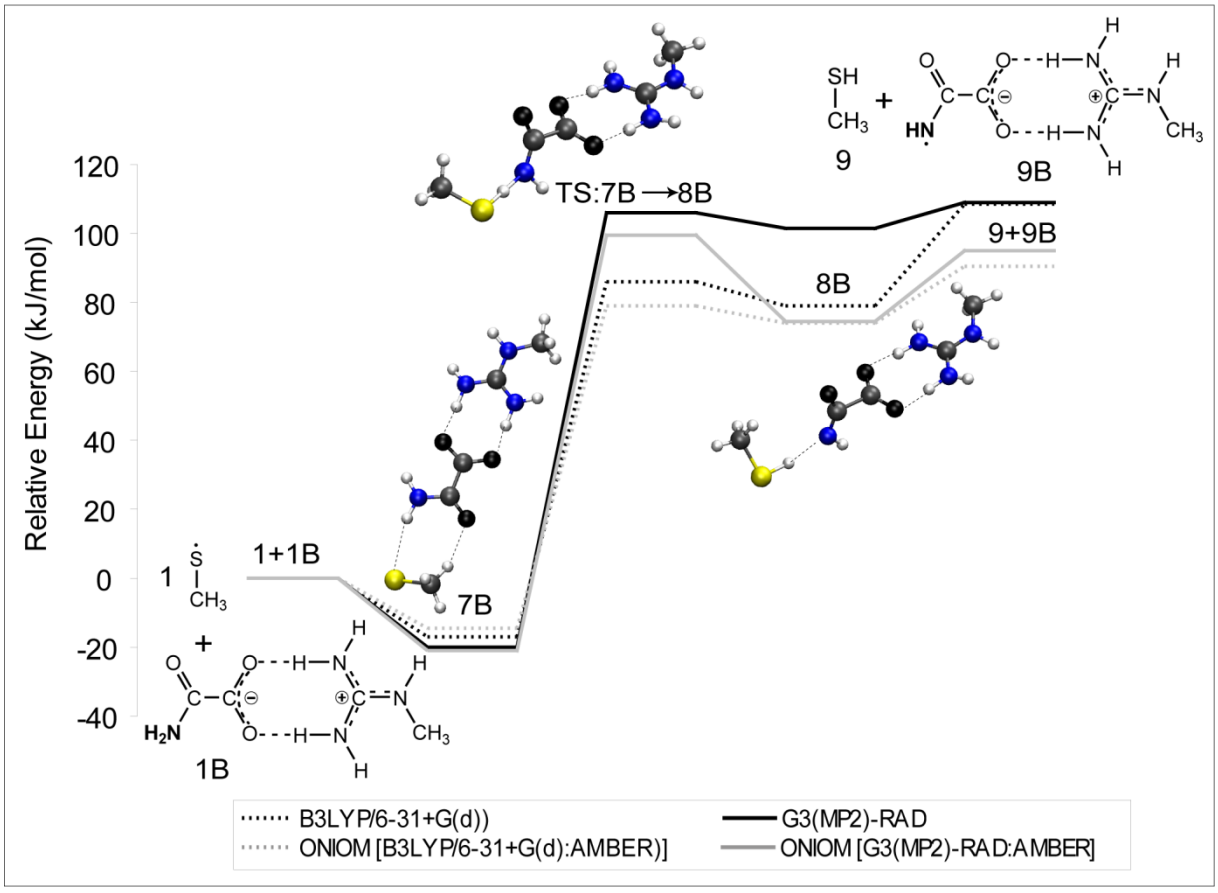
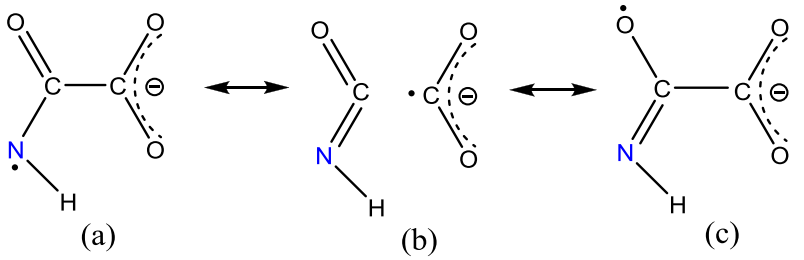


Figure 3.15 A comparison of QM and ONIOM results for hydrogen abstraction from oxamate by thiyl radical at oK (geometries presented were optimized at B3LYP/6-31+G(d) level of theory).

For hydrogen abstraction from oxamate ONIOM results give the same shape of the potential energy surface. Geometries optimized with ONIOM deviate only slightly from structures optimized at DFT level of theory. Hydrogen bond lengths present in the salt bridge are longer for ONIOM geometries than in their DFT counterparts, as observed for each geometry obtained in this study. Approximately, ONIOM bond lengths are 0.2 Å longer, most likely because the boundary that separates QM and MM part of the system intersects these bonds.

Interesting observation has been made about C1-C2 bond length. An elongation of C1-C2 bond in TS:7B→8B (1.655 Å) occurs, compared to the bond length in reactant oxamate (1.562 Å). This bond length is even longer in final product 9B, where it reaches the value of 1.729 Å. The elongation of this bond can be explained in terms of the delocalization of the unpaired

electron, depicted with the resonance structures in Scheme 3-12, where the resonance structure (b) is being preferentially favoured by the presence of the methylguanidinium.



Scheme 3-12 The resonance structures with the significant contribution to the overall structure of oxaminylate.

Calculated spin density distribution and bond lengths are in a good agreement with given explanation. In TS:7B→8B the significant percentage of the spin density is located on carboxylate group (0,27) and the rest is distributed between S, N and carbonyl O atoms. The electron delocalization is due to orbital overlap between amino and carbonyl group.[64],[65] As these results show, the abstraction of hydrogen from the amino group of oxamate can activate process of C1-C2 bond breaking to some extent . A certain C1-C2 bond elongation occurs in pyruvylate, but this elongation is negligible (0.02 Å) compared to oxaminylate.

The next step of the alternative mechanism is fragmentation of the radical formed in the first step. The energy barrier that has to be crossed for fragmentation of oxaminylate radical to occur almost doesn't exist, according to both ONIOM and DFT results (Figure 3.16). The reaction enthalpy shows the largest discrepancy between the methods used.

The fragmentation of the oxaminylate radical is significantly facilitated by the elongation of the C1-C2 bond, discussed above. Once the radical is formed, the barrier for the breaking of the C1-C2 bond is negligible and exothermicity notable.

When this reaction is modelled using neutral model, the reaction goes over higher energy barrier, but it is also more exothermic compared to the energies calculated with the complexed model. Although the fragmentation of the oxaminylate radical is spontaneous and probably occurs immediately after the radical has been formed, the removal of hydrogen from the amino group is energetically expensive step. Additionally, this reaction should take place inside of an enzyme, where the position of the reactants is relatively rigid. In this case, stereochemical conditions imposed by the active site prefer the addition of the cysteinyl radical to substrate.

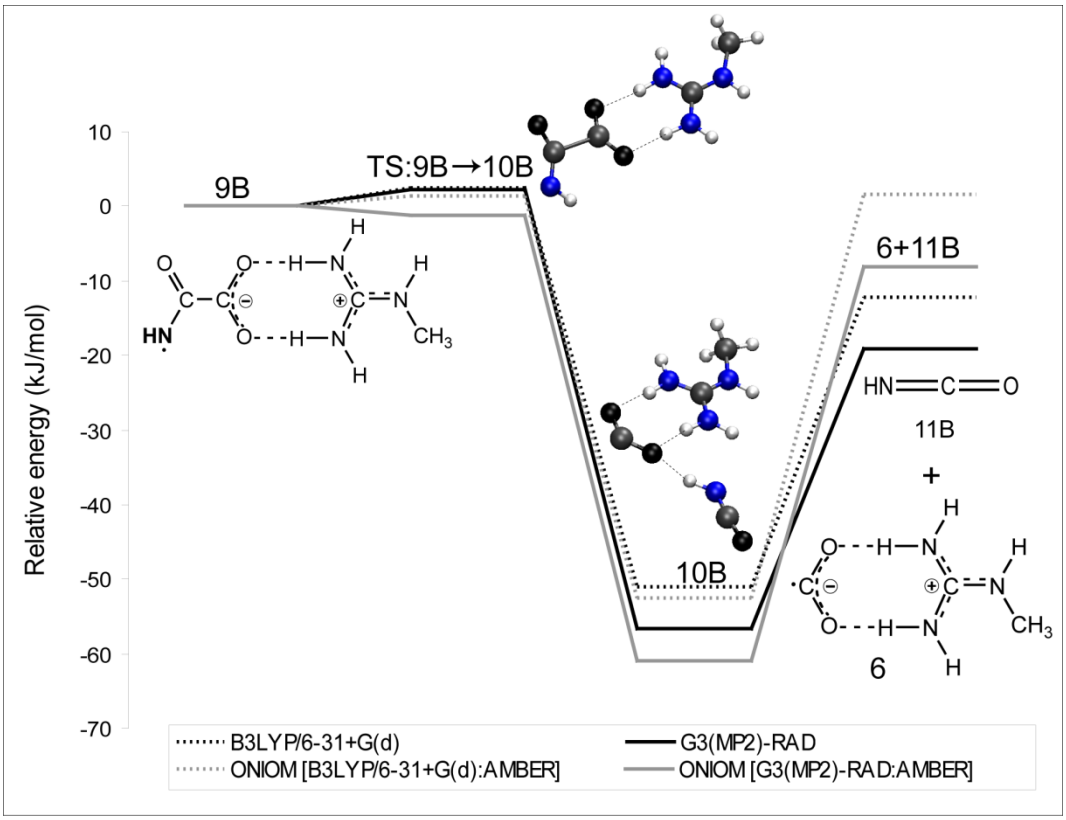


Figure 3.16 Comparison of DFT and ONIOM results for fragmentation of oxaminylate radical at oK (geometries presented were optimized at B3LYP/6-31+G(d) level of theory).

Note to be incorporated, HF charges in the gas phase are overpolarized to an extent and this is probably partially responsible for (or is mainly manifested in) the complexation energy problem. Notably, G3 handles this much better than B3LYP.

3.3.3 CONCLUSION

The results presented in this study show the ONIOM(G3(MP2)-RAD:AMBER) hybrid to be a viable means for incorporating a QM method of chemical accuracy into a QM/MM treatment. The agreement between the pure QM G3(MP2)-RAD treatment and the QM/MM approximation is impressive and, in most cases, exceeds that obtained from an all QM B3LYP/6-31+G(d) approach. This result suggests that such an approach may even be useful for small models, where a full DFT treatment might nevertheless be feasible. This final point is relevant to our previous study concerning the most appropriate treatment for the arginine-bound carboxylate motif. When using an all QM truncation, we found that the neutral form of the carboxylic acid would be the most satisfactory.[56] The results presented in the current study show that a high-level treatment of the substrate combined with an MM treatment of the arginine fragment would provide an even better result at a comparable cost.

We have validated the method on four different reaction sequences relevant to the substrate transformation catalyzed by PFL. In addition to serving for this purpose, these calculations reveal some interesting facts about the mechanism. The reaction barrier and endothermicity for sulfur radical addition to oxamate appear too high for such a pathway to prove viable in the enzyme. This result provides an explanation for the lack of observed turnover with oxamate, despite this inhibitor being ideally positioned to react.

We also considered H-atom abstraction as an alternative initial step in the PFL mechanism. If this mechanism were to occur, it would seemingly result in unwanted products. Interestingly, with neutral pyruvic acid as the substrate, the S-addition and H-abstraction barriers are very similar. When the salt-bridge model is used, the barrier for S-addition is reduced while that for H-abstraction remains virtually unaffected. On this basis it is tempting to speculate that PFL uses the active site arginine residues to avoid the unwanted H-abstraction pathway both by lowering the barrier for S-addition and by binding the substrate in a position where the methyl group is unable to approach the reactive thiyl radical.[66]

Naturally, the strength of the ONIOM(G3(MP2)-RAD:AMBER) method is its potential applicability to systems much larger than those presented in this study and such applications are indeed underway. Nevertheless, the validation of the approach on systems where high-level QM benchmark calculations are possible is an important step to establishing the viability of the technique for informed future use.

3.3.4 REFERENCES

1 Warshel, A.; Levitt, M. *J. Mol. Biol*. **1976**, 103, 227.

2 Singh, U. C.; Kollman, P. A. *J. Comput. Chem*. **1986**, 7, 718.; Bash, P. A.; Field, M. J.; Karplus, M. *J. Am. Chem. Soc*. **1987**, 109, 8092.; Gao, J.; Xia, X. *Science* **1992**, 258, 631.; Bakowies, D.; Thiel, W.; *J. Comput. Chem*. **1996**, 17, 87.; Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *J. Comput. Chem*. **2000**, 21, 1442.

3 See review articles: Senn, H. M.; Thiel, W. *Topics in Current Chemistry*, Vol 268. Edited by Reiher M. Berlin: Springer; **2007**, 173-290.; Lin, H.; Truhlar, D. G. *Theor. Chem. Acc*. **2007**, 117, 185; Senn, H. M.; Thiel, W. *Curr. Opin. Mol. Biol*. **2007**, 11 (2), 182; Warshel, A.; Sharma, P. K.; Kato, M.; Xiang, Y.; Liu, H.; Olsson, M. H. M. *Chem. Rev*. **2006**, 106, 3210; Bruice, T. C. *Chem. Rev*. **2006**, 106, 3119; Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem*. **2005**, 56, 389; Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, 303, 186.

4 Maseras, F.; Morokuma, K. *J. Comput. Chem*. **1995**, 16, 1170; Kerdcharoen, T., Morokuma, K. *Chem. Phys. Lett*. **2002**, 355, 257.

5 Das, D; Eurenius, K. P.; Billings, E.M.; Sherwood, P.; Chatfield, D. C.; Hodošček, M.; Brooks, B. R. *J. Chem. Phys*. **2002**, 117, 10534.

6 Antes, I.; Thiel, W. *J. Phys. Chem. A* **1999**, 103, 9290.

7 Zhang, Y.; Lee, T.-S.; Yang, W. *J. Chem. Phys*. **1999**, 110, 46.

8 DiLabio, G. A.; Hurley, M. M.; Christiansen, P. A. *J. Chem. Phys*. **2002**, 116, 9578.

9 Ferenczy, G. G.; Rivail, J.-L.; Surjan, P. R.; Naray-Szabo, G. *J. Comput. Chem*. **1992**, 13, 830.

10 Monard, G.; Loos, M.; Thery, V.; Baka, K.; Rivail, J.-L. *Int. J. Quantum. Chem*. **1996**, 58, 153.

11 Sironi, M.; Genoni, A.; Civera, M., Pieraccini, S.; Ghitti, M. *Theor. Chem. Acc*. **2007**, 117, 685.

12 Gao, J.; Amara, P.; Alhambra, C.; Field, M. J. *J. Phys. Chem. A* **1998**, 102, 4714; Gao, J.; Truhlar, D.G. *Annu. Rev. Phys. Chem*. **2002**, 53, 467; Pu, J.; Gao, J.; Truhlar, D.G. *J. Phys. Chem. A* **2004**, 108, 5454.

13 Bakowies, D.; Thiel, W. *J. Phys. Chem*. **1996**, 100, 10580.

14 Waszkowycz, B.; Hillier, I. H.; Gensmantel, N.; Payling, D. W. *J. Chem. Soc. Perkin. Trans*. **1991**, 2 2025; Sinclair, P.E.; de Vries, A.; Sherwood, P.; Catlow, C. R. A.; van Santen, R. A. *J. Chem. Soc. Faraday Trans*. **1998**, 94, 3401; Das, D.; Eurenius, K. P.; Billings, E. M.; Sherwood, P.; Chatfield, D. C.; Hodošček, M.; Brooks, B. R. *J. Chem. Phys*. **2002**, 117, 10534; Lin, H.; Truhlar, D. G. *J. Phys. Chem. A* **2002**, 109, 3991.

15 Sprik, M.; Klein, M. L. *J. Chem. Phys*. **1988**, 89, 7556.

16 Tubert-Brohman, I.; Acevedo, O.; Jorgensen, W. L. *J. Am. Chem. Soc*. **2006**, 128 (51), 16904; Jorgensen, W. L.; Tirado-Rives, J. *J. Comp. Chem*. **2005**, 26 (16), 1689.

17 Cummins, P. L.; Rostov, I. V.; Gready, J. E. *J. Chem. Theory Comput*. **2007**, 3 (3), 1203.

18 Sharma, P. K.; Chu, Z. T.; Olsson, M. H. M.; Warshel, A. *Proc. Natl. Acad. Sci. USA* **2007**, 104 (23), 9661.

19 Elstner, M. *Theo. Chem. Acc*. **2006**, 116 (1-3), 316; Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.;König, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, 110, 6458.

20 Zheng, J. J.; Wang, D. Q.; Thiel, W.; Shaik, S. *J. Am. Chem. Soc*. **2006**, 128 (40), 13204.

21 Rinaldo, D.; Philipp, D.M.; Lippard, S. J.; Friesner, R. A. *J. Am. Chem. Soc*. **2007**, 129 (11), 3135.

22 Claeyssens, F.; Harvey, J. N.; Manby, F. R.; Mata, R. A.; Mulholland, A. J.; Ranaghan, K.E.; Schutz, M.; Thiel, S.; Thiel, W.; Werner, H. J. *Angew. Chem. Int. Ed*. **2006**, 45 (41), 6856.

23 Warshel, A. *Annu. Rev. Biophys. Biomol. Struct*. **2003**, 32, 425; Garcia-Viloca, M.; Gao, J.; Karplus, M.; Truhlar, D. G. *Science* **2004**, 303, 186.; Min, W.; English, B. P.; Luo, G.; Cherayil, B. J.; Kov, S. C.; Xie, X. S. *Acc. Chem. Res*. **2005**, 38, 923.

24 A related issue is the accurate prediction of spectroscopic properties, which also requires sophisticated QM treatments. See, for example: Kirchner, B.; Wennmohs, F.; Ye, S.; Neese, F. *Curr. Opin. Chem. Biol.* **2007**, 11 (2), 134.

25 Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys*. **2007**, 126, 084108.

26 Curtiss, L. A.; Raghavachari, K. *Theo. Chem. Acc*. 2002, 108 (2), 61.

27 Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, 123 (12), Art. No. 124107.

28 Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.;Sieber, S.; Morokuma, K. *J. Phys. Chem*. **1996**, 100, 19357.

29 Vreven, T.; Morokuma, K. *J. Comput. Chem*. **2000**, 21, 1419.

30 Vreven, T.; Morokuma, K.; Farkas, O .; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem*. **2003**, 24, 760..

31 Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery J. A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, 2 (3), 815.

32 Froese, R. D. J.; Morokuma, K. *Chem. Phys. Lett*. **1999**, 305, 419; Froese, R. D. J.; Morokuma, K. *J. Phys. Chem. A* **1999**, 103, 4580; Vreven, T.; Morokuma, K. *J. Chem. Phys.* **1999**, 111, 8799; Vreven, T.; Morokuma, K. *J. Phys. Chem. A* **2002**, 106, 6167.

33 Li, M.-J.; Liu, L.; Fu, Y.; Guo, Q.-X. *J. Phys. Chem. B* **2005**, 109, 13818; Li, M.-J.; Liu, L.; Fu, Y.; Guo, Q.-X. *J. Mol. Struc. Theochem*. **2007**, 815, 1.

34 Coote, M.-L. *J. Phys. Chem. A* **2005**, 109, 1230; Izgorodina, E. I.; Coote, M. L. *Chem. Phys*. **2006**, 324, 96; Izgorodina, E. I.; Brittain, D. R. B.; Hodgson, J. L.; Krenske, E. H.; Lin, C. Y.; Namazian, M.; Coote, M. L. *J. Phys. Chem. A* **2007**, 111, 10754.

35 Henry, D. J.; Sullivan, M.B.; Radom, L. *J. Chem. Phys*. **2003**, 118, 4849.

36 Ponder, J. W.; Case, D. A. *Adv. Prot. Chem*. **2003**, 66, 27.

37 Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys*. **1999**, 110, 4703.

38 Knappe, J.; Blaschkowski, H. P.; Gröbner, P.; Schmitt, T. *Eur . J. Biochem*. **1974**, 50, 253.

39 Knappe, J.; Neugebauer, F. A.; Blaschkowski, H. P.; Gänzler, M. *Proc. Natl. Acad. Sci. USA* **1984**, 81, 1332.

40 Eriksson, L. A.; Himo, F. *J. Chem. Soc. Perkin Trans. 2*, **1998**, 305.

41 Frey, M.; Rothe, M.; Wagner, A. F. V.; Knappe, J. *J. Biol.Chem*. **1994**, 269, 12432.

42 Becker, A.; Fritz-Wolf, K.; Kabsch, W.; Knappe, J.; Schultz, S.; Wagner, A. F. V. *Nat. Struct.Bio*. **1999**, 6, 969.

43 Becker, A.; Kabsch, W. *J. Biol.Chem*. **2002**, 277 , 40036.

44 Himo, F.; Eriksson, L. A. *J. Am. Chem. Soc*. **1998**, 120, 11449.

45 Parast, C.V.; Wong, K. K.; Lewisch, S. A.; Kozarich, J. W. *Biochemistry* **1995**, 34, 2392.

46 Sawers, G.; Watson, G. *Mol. Microbiol*. **1998**, 29, 945.

47 Lehtiö, L.; Goldman, A. *Protein Eng. Des. Sel*. **2004**, 17, 545.

48 Selmer, T.; Pierik, A. J.; Heider, J. *Biol. Chem*. **2005**, 386, 15347.

49 Sofia, H. J.; Chen, G.; Hetzler, B. G.; Reyes-Spindola, J. F.; Miller, N. E.; *Nucleic Acid Res*. **2001**, 229, 1097.

50 Eklund, H.; Fontecave, M. *Struct. Fold Des*. **1999**, 7, R257.

51 Krieger, C. J.; Roseboom, W.; Albracht, S. P. J.; Spormann, A. M. J. *J. Biol.Chem*. **2001**, 276, 12924.

52 O'Brien, J. R.; Raynaud, C.; Croux, C.; Girbal, L.; Soucaille, P.; Lanzilotta, W. N. *Biochemistry* **2004**, 43, 4635.

53 Sintchak, M. D.; Arjara, G.; Kellogg, B. A.; Stubbe, J.; Drennan, C. L. *Nat. Struct. Biol*. **2002**, 9, 293.

54 Lucas, M. F.; Fernandes, P. A.; Eriksson, L. A.; Ramos, M. J. *J. Phys. Chem. B* **2003**, 107, 5751.

55 Guo, J.-D.; Himo, F. *J. Phys. Chem. B* **2004**, 108, 15347.

56 Čondić-Jurkić, K.; Perchyonok, V. T.; Zipse, H.; Smith, D. M. *J. Comput. Chem*. **2008**, 29, 2425.

57 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; and Pople, J. A. *Gaussian 03, Revision C.02*, Gaussian, Inc., Wallingford CT, 2004.

58 Knowles, P. J.; Hampel, C.; Werner, H-J. *J. Chem. Phys*. **1993**, 99, 5219; Erratum: *J. Chem. Phys*. 2000, 112, 3106.

59 Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Rauhut, G.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Hampel, C.; Hetzer, G.; Lloyd, A. W.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pitzer, R.; Schumann, U.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T. MOLPRO, version 2006.1, a package of ab initio programs. University College Cardiff Consultants Limited, Cardiff, 2008.

60 Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. *J. Am. Chem. Soc*. **1993**, 115, 9620.

61 Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem*. **1993**, 97, 10269.

62 Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc*. **1995**, 117, 5179.

63 Freindorf, M.; Shao, Y.; Furlani, T. R.; Kong, J. *J. Comput. Chem*. **2005**, 26, 1270.

64 Wood, G.P.F.; Henry, D. J.; Radom, L. *J. Phys. Chem*. **2003**, 107, 7985.

65 Rajendra, P.; Chandra, P. *J. Phys. Chem*. **2001**, 114, 7450.

66 Rétey, J. *Angew. Chem. Int. Ed*. **1990**, 29, 355.

3.4 A Molecular Dynamics Study of PFL Catalysis: Second Half-Reaction

Once pyruvate formate-lyase has been activated (see Section 3.1.1), it catalyzes decarboxylation of pyruvate under anaerobic conditions in many microorganisms, where the final products are formate and acetyl-coenzyme A. This reaction proceeds in two steps:

$$\text{PFL} + \text{pyruvate} \rightleftharpoons \text{acetyl-PFL} + \text{formate}$$

$$\text{acetyl-PFL} + \text{CoA} \rightleftharpoons \text{PFL} + \text{acetyl-CoA}$$

In the first step, the C-C bond in pyruvate is cleaved upon the attack of thiyl radical located on Cys418, producing formate radical and acetylated Cys418. A detailed dissection of this half-reaction is provided in Sections 3.2 and 3.3, where small model systems were used together with very accurate QM methods to describe properly the chemistry taking place in the active site of PFL. The resulting energy profiles strongly support the currently accepted mechanism of the first half-reaction. However, some open questions remain when it comes to the mechanism of the second step in which CoA undergoes acetylation. Several hypotheses were suggested in order to explain a possible reaction mechanism of the transacetylation process (see Section 3.1.2), but none of them were corroborated experimentally or theoretically. The initial hypothesis did not assume a radical involvement in the second half-reaction at all; it was rather regarded as a common acetyl exchange between sulfhydrils. The proposal that the second half-reaction also has a radical character came later, suggesting that the CoA reacts with acetylated Cys418 via radical mechanism. This mechanism involves hydrogen abstraction step from CoA thiyl group by a radical species in the PFL active site, followed by acetyl transfer and restoration of radical centre at Cys418. It remains vague, though, which radical species should abstract hydrogen from CoA in this scenario. Of course, all these events take place under the assumption that CoA has already reached the active site. This brings us to the other unresolved piece of the PFL mechanistic puzzle – how does CoA reach the buried active site in the first place, when its known binding site is located at the protein surface 30 Å away? It is clear that certain conformational changes are necessary, but what triggers those changes and how are they manifested are the questions of interest tackled in this section. The aim of the section is getting a better insight into these complex processes in the protein environment by utilizing the unrestrained molecular dynamics simulations and selected free energy methods.

As mentioned, in the early PFL studies it was assumed that only the first half-reaction involved radical species, while their participation in the second step was regarded as unlikely, since it was believed to be a common acetyl exchange between sulfhydryls. However, subsequently was observed that S-acetyl transfer depends on the radical state of the enzyme, proven by [^{14}C]CoASH/acetyl-CoA isotope exchange experiment.[1] This feature was attributed to conformational requirements for protein functionality. The effect of radical presence on acetyl transfer was additionally confirmed in the following experiments, where a ~10^5-fold rate increase of S-acetyl transfer with the radical form of PFL was measured.[2] Despite these observations, the idea of non-radical (heterolytic) mechanism of transacetylation between active-site cysteine and CoA remained and dominated for a long period of time, until Himo and Eriksson suggested a homolytic acetyl transfer in their theoretical study.[3] In this study, it was suggested that the formyl radical created from cleaving C-C bond in pyruvate abstracts a hydrogen atom from Cys419, which would in turn generate a radical on thiyl group of CoA. This would facilitate the transfer of acetyl group from Cys418 to radical CoA and restore radical at Cys418 for a new catalytic cycle. Calculations performed in the study showed that these mechanistic steps are energetically feasible. In the following computational study performed by Guo and Himo, an alternative approach in which formyl radical abstracts hydrogen directly from CoA was introduced.[4] The third option might involve storing the radical at Gly734 between two half-reactions until CoA reaches the active site, but this scenario was not examined theoretically. It is probably less likely option due to the higher number of mechanistic steps required in the process, but it cannot be excluded. All three possibilities remain at a speculative level, leaving the question of how the acetyl group is transferred from protein to CoA without a firm answer.

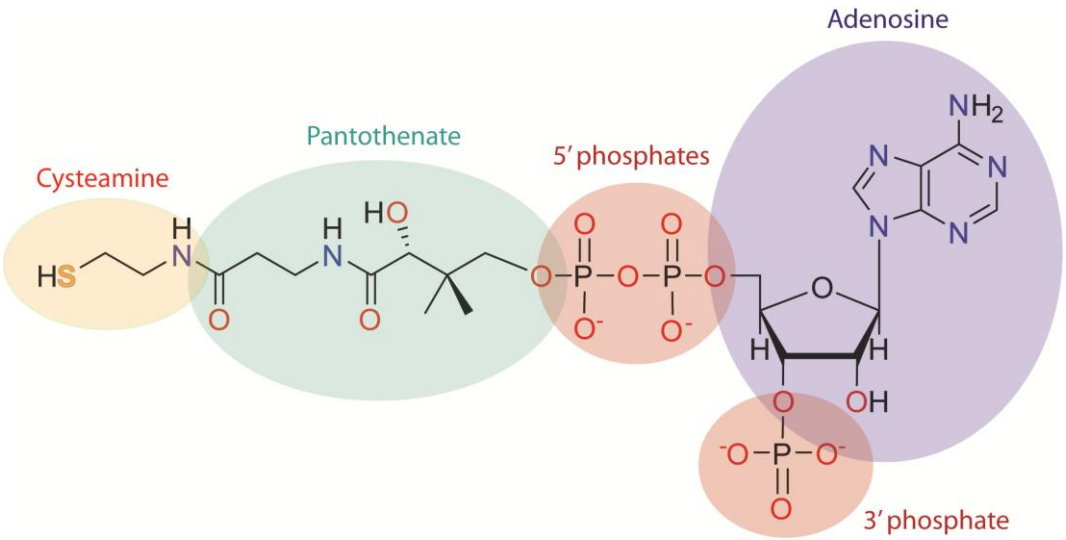


Figure 3.17 The building blocks of coenzyme-A.

This mechanistic issue is tightly coupled to the problem that arises from the fact that CoA binding site is located at the protein surface 30 Å away from the active site. The binding site was identified by solving the crystal structure of protein in complex with pyruvate (or oxamate) and CoA and it lies in the proximity of the interface between two constituent monomers. As described in the introductory Section 3.1.1, the amino group of CoA adenine moiety is hydrogen bonded to Asn145 and Gln146, while the imidazole ring makes stacking interaction with the neighbouring Phe149 (Figure 3.18). Additional stabilisation is achieved through salt bridge interaction formed between 3' and 5' phosphates and positively charged Lys161. Also, in the crystal structures containing CoA a density peak was detected next to the phosphates that was later interpreted as Mg^{2+}, although no divalent cations were included in the crystallization buffer, according to the authors.

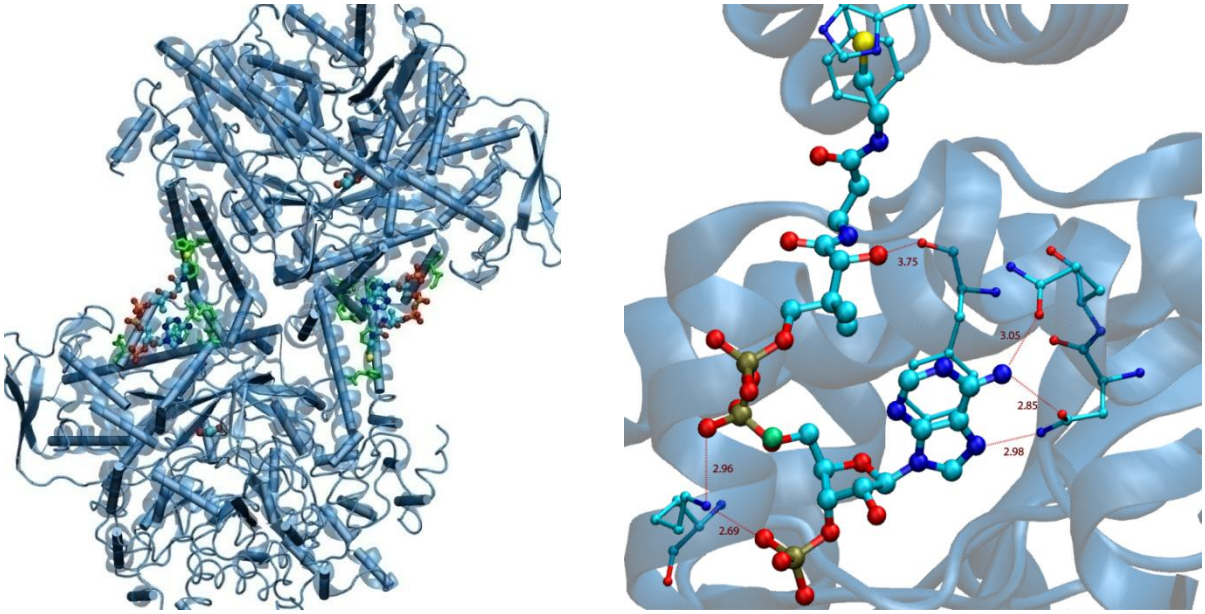


Figure 3.18 (a) Binding site of CoA in the PFL dimer spans two subunits and it is approximately 30 Å away from the active site; (b) Detailed view of the interactions between CoA and the protein residues in the binding site, with Mg^{2+} shown as a green sphere.

Namely, the buffers used in the crystallization process contained sodium and potassium ions, while CoA was added in the mixture in the form of lithium salt. Seven sodium ions with approximately 1300 water molecules per monomer were identified based on the crystallographic data, according to the available PDB files of the PFL crystal structures (1h16, 1H17). Bearing in mind that often is quite difficult to distinguish between metal ions based on the electron density maps, it is quite understandable that sometimes these ambiguous peaks are misinterpreted. This is especially the case among isoelectronic species such as Mg^{2+} and Na^{+} (and even water, which is almost isoeletronic).[5] Basically, identification of the metal ion

relies on the coordination number and distances between the ion and the surrounding ligands. A reliable interpretation is available only at high resolution (< 1.5 Å) and high occupancy. For example, if the ligand is water, then metal-water distance is what distinguishes Na^+ from Mg^{2+}. For PFL, the available data do not provide a straightforward and unambiguous discrimination between the two. Although Mg^{2+} ions are very common in biological systems, the reason for opening a discussion about validity of this interpretation is the possible strong influence of the Mg^{2+} ion and its charge on the conformational space of CoA. A proper exploration of the phase space available to CoA and the protein is a key aspect of this research, which makes the identity of that peak an important question for this topic.

In that context, the observation that CoA adopts unusual *syn* conformation in respect to the N-glycosidic bond when bound to PFL becomes even more interesting. Namely, the preferred conformation of free CoA in solution is *anti,* but this *syn* conformation allows the thiol group on pantothenate chain to settle in an aromatic "sandwich" formed by the side chains of Phe200 and His227 of the opposing monomer. Therefore, it is obvious that changes of CoA and protein conformations are necessary in order to bring CoA to the active site. But, what is the event that sets these changes in motion?

The available experimental information about the second catalytic step is rather scarce, since most of the experiments were oriented toward clarification of the first half-reaction and how glycyl radical is generated in the first place. It has been established that the presence of CoA is not obligatory during the first half-reaction, primarily due to the observation that ^{14}C exchange between formate and pyruvate-carboxyl group is CoA-independent.[1] If CoA is present, it is bound to PFL in a stand-by mode, as suggested by the authors who solved the crystal structure of PFL in complex with CoA.[6] This implies that the system should undergo some conformational changes to allow CoA to reach the active site and pick up the acetyl. It has been suggested that ribose and pantothenate moiety might rotate around *N*-glycosidic bond and change from *syn* to *anti* conformation, which would enable CoA to reach the active site. This transition is expected to be energetically favourable and this changed binding site would also allow binding of the free CoA from the solution in the *anti* form upon the completion of the first half-reaction. However, it is not clear what triggers this conformational change and which residues are involved in this motion. Finding the possible answers to these questions was the main aim of the research presented in this section. The research was conducted using molecular dynamics simulations, which are a powerful computational tool for addressing the problems of the molecular processes that are sometimes difficult to access by standard biochemical experimental procedures.

As a part of the efforts to unravel the series of events that bring CoA to the active site, the potential coupling of chemical changes taking place in the active site and large-scale protein motions was more closely examined. To capture this coupling, if present, the modern free energy methods were used to calculate potentials of mean force for the process of CoA entering the active site before and after the pyruvate cleavage. The comparison of the two free energy profiles corresponding to the two different points on the catalytic pathway might give a hint whether the cleavage of pyruvate has any influence on the accommodation of CoA in the active site. Certain complications arise in this approach from the unresolved mechanistic issues of the second half-reaction, described earlier in the text. Namely, the unknown location of the radical between two half-reactions complicates the choice of the molecular topology after the pyruvate cleavage. Three possible scenarios were covered earlier in the text, where radical was located either on the newly formed formate or it was transferred to Cys419 or even Gly734 as temporary radical storages. For the purposes of the study presented in this thesis, it was decided to follow the mechanism proposed by Guo and Himo and to leave the unpaired electron on formate, which was demonstrated as an energetically plausible mechanistic solution.

In summary, an extensive computational study of the PFL system was carried out in order to provide a better insight in the protein behaviour and conformational changes required for the accommodation of the second substrate CoA in the active site. The models representing the PFL system before and after the first half-reaction with pyruvate were used to examine the possible effect that acetylation of the enzyme has on the necessary conformational changes. This was achieved by estimating potential of mean force for the entrance of CoA into the active site before and after pyruvate cleavage using state-of-the-art methods for free energy calculation.

3.4.1 Computational Details

3.4.1.1 Structural Models

All the structural models were derived from a single crystal structure containing PFL in a complex with pyruvate and CoA (PDB entry: 1H16).[6] Before further processing, the original PDB file was adjusted by removing duplicate entries and assigning the protonation states of histidines by running *reduce*[7] algorithm implemented in AmberTools software package.[8] Assignments made by *reduce* were additionally verified and confirmed by visualizing the structure in VMD[9] and inspecting the local environments of the histidines. All the ions and water molecules crystallized with the protein and listed in the PDB file were retained.

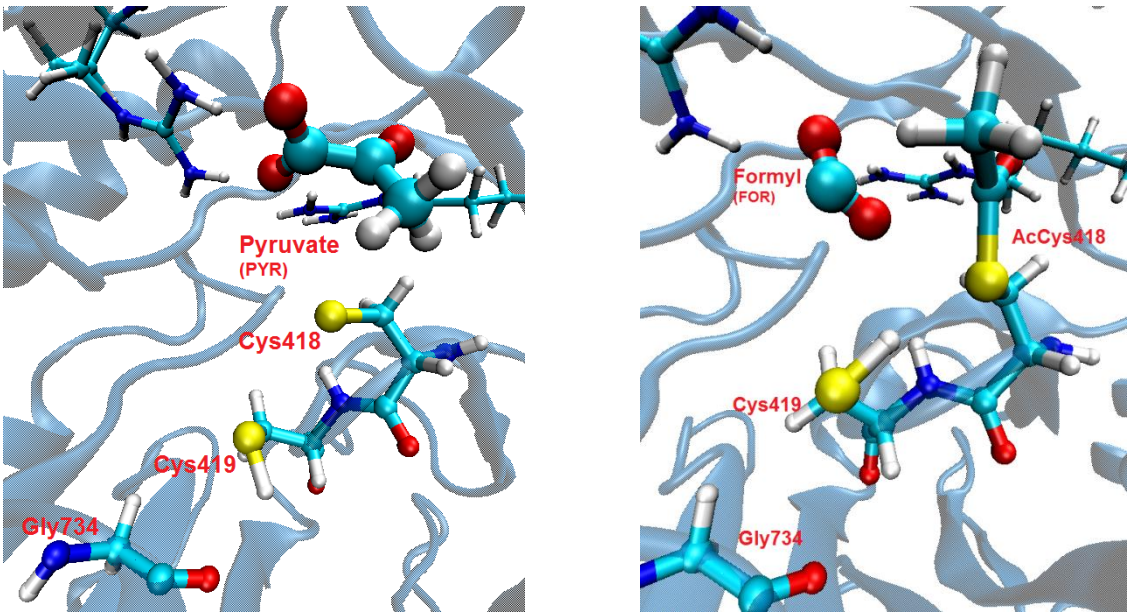


Figure 3.19 Models representing the active site of PFL before (a) and after (b) the first half-reaction of catalysis. Both models also contain CoA occupying its binding site at the protein surface.

Building the appropriate models of the system under investigation is one of the essential steps in every computational study. In this case, we are interested in the PFL system before and after the first half-reaction, during which C-C bond in pyruvate is cleaved and Cys418 undergoes acetylation. The model representing the system before reaction (PFL) contains the protein in the complex with CoA and pyruvate bound in the active site, while Cys418 is the radical carrier (Figure 3.19a). The crystal structure used for model building is solved for a non-radical form of the protein and the authors argue that the obtained crystal structure is a suggestive snapshot of that point in which radical transfer between Gly734 and cysteines has

already occurred, leaving Gly734 Cα as an sp^3 carbon. Basically, this model describes the system just in the moment before the Cys418 thiyl radical attacks pyruvate. The second structural model represents the point on the reaction pathway where the C-C bond in pyruvate is broken, resulting with acetylated protein at Cys418 and formyl radical still bound to Arg172 and Arg432 (AcPFL). This model is denoted as AcPFL (Figure 3.19). Both models contain CoA bound at the protein surface according to the crystal structure data.

A set of models was built using only a single monomer, while an additional set of models was derived from a full homodimer. Namely, the preliminary calculations were performed using smaller systems based only on a single subunit, but dimeric models were subsequently introduced to ensure that possible modulation of the important conformational changes by the other subunit is taken into account (Figure 3.20). In addition, the binding site of CoA spans both subunits so that the adenine and sugar moiety interacts with one monomer, while the thiol group is placed between a histidine and phenylalanine residue of the opposing monomer. Having that in mind, it was concluded that dimer is a more complete model of the PFL system. All dimeric models have only one active site activated in accordance with the experimental findings showing half-site radical occupancy. Dimeric models will be denoted by *d* (dPFL and dAcPFL) and monomeric by *m* (mPFL and mAcPFL).

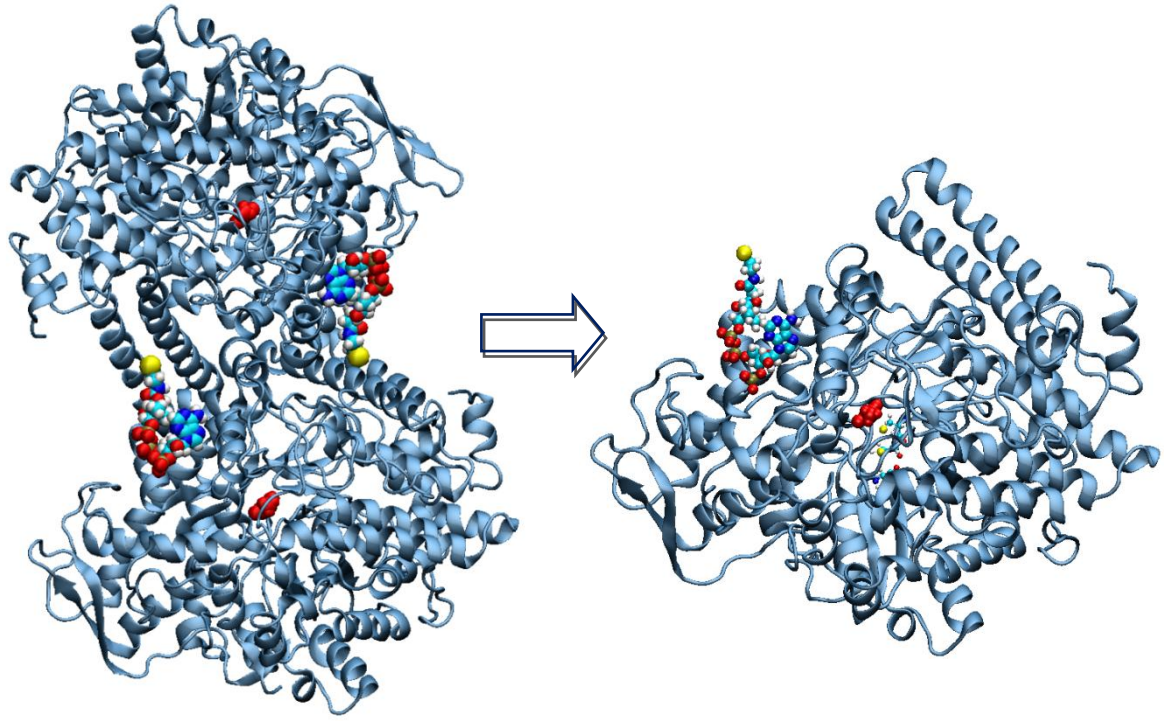


Figure 3.20 Crystal structure of the homodimeric form of PFL in complex with CoA and pyruvate bound in the active site (left) used for model building. A single subunit (right) was also used as a PFL model in molecular dynamics simulations to obtain preliminary results at lower computational costs.

3.4.1.2 Parameter Assignment

The parameters assigned to standard amino-acid residues in PFL were taken from AMBER03 (*ff03.r1*) force field developed by Duan *et al.*,[10] available within AMBER11[11] and AMBER12[12] software package. For the non-standard residues, the missing parameters were derived in a fashion consistent with the procedure used for parameterization of *ff03.r1* using Antechamber suite implemented in the AMBER11 package.[11] These non-standard residues include radical cysteine ($Cys^{\bullet}$), acetylated cysteine (AcCys), and enzyme substrates – pyruvate, formate and CoA. Specifically, it was necessary to provide the missing charges for these residues, which were obtained by following the standard RESP procedure.[13] Charges used in *ff03.r1* force field were derived from quantum mechanical calculations at B3LYP/cc-pVTZ//HF/6-31G(d,p) level of theory combined with an IEFPCM (ε = 4.335) continuum dielectric model mimicking solvent polarization. The same approach was used to derive charges for substrates and modified cysteines. The anionic form of pyruvate was used in parameterization, and formate was parameterised as a radical ($COO^{\bullet -}$) and closed-shell species ($HCOO^{-}$). In CoA, all the phosphate groups were fully charged. The procedure of charge derivation for non-standard amino-acid residues involves fitting charges to the electrostatic potential of two conformers, corresponding to α-helix and extended conformations. During the geometry optimizations of radical and acetylated cysteine residues, the dihedral angles (ϕ,Ψ) were fixed at (–60°, –40°) and (–120°, 140°) for the α-helix and the extended conformations, respectively. All the QM calculations were performed using the Gaussian 03 software package.[14] The other bonding and non-bonding parameters required to build full libraries for these molecules were taken from generalized Amber force field (GAFF).[15]

The topology files describing PFL system before and after pyruvate cleavage were built using the LEaP module of AMBER11 and all the necessary parameters were taken from the standard *ff03.r1* force field and custom-built libraries. Furthermore, each system was solvated with TIP3P waters in a truncated octahedron box. The edge length of the resulting box of solvent was ~95 Å for systems containing a monomer and ~137 Å for systems containing dimers. As mentioned earlier, all crystal water molecules and seven sodium ions per monomer present in the PDB file were retained, including the magnesium ion located close to the phosphates of CoA. An additional 8 sodium ions per monomer were required to neutralize the system. The number of waters added to monomeric systems was approximately ~ 21400 molecules, while dimers were surrounded by ~ 55600 solvent molecules.

As discussed in the introductory part, there is no strong evidence to support the interpretation of density peak close to CoA as an Mg^{2+} ion. To examine the possible effect of this ion on the conformational changes of CoA bound to a dimer, a new set of topology files were built in the same fashion as described above, except that the Mg^{2+} ion was removed and replaced by two Na^{+} ions (to keep the system neutral). The number of added solvent molecules is approximately 62 000.

Table 3-1. A summarized overview of the built model systems and the associated notation that will be used throughout the text. The notation is intended to encompass all the important information about the system at glance – a clear difference between topologies describing the system before (PFL) and after pyruvate cleavage (AcPFL). Models derived from a single subunit carry the prefix *m*, while those built from a full dimer are denoted by the prefix *d*. Dimeric systems are additionally divided into two subgroups based on the presence or absence of Mg^{2+} ion, with the latter systems carrying the suffix *X*.

	Monomer	Dimer	
	with Mg^{2+}	with Mg^{2+}	without Mg^{2+}
PFL	mPFL	dPFL	dPFLX
AcPFL	mAcPFL	dAcPFL	dAcPFLX

3.4.1.3 Molecular Dynamics Simulations

All the systems (Table 3-1) were treated within periodic boundary conditions. Long range electrostatic interactions were calculated with Particle-Mesh Ewald (PME) technique with the default non-bonded cut-off of 8.0 Å to limit the direct space sum. The temperature in all simulations was controlled by coupling the system with the Berendsen[16] or Langevin[17] thermostat with collision frequency set to 3 ps^{-1}. An integration time step of 2 fs was used and the SHAKE algorithm was employed to constrain bonds involving hydrogen atoms during dynamics.

Equilibration. Equilibration of the system was carried out in several steps: (*i*) Steepest descent minimization was applied to the protein-substrates complex (solute) with harmonic positional restraints on solvent molecules (5 kcal/mol $Å^2$). (*ii*) Minimization was repeated with restraints on the solute (5 kcal/mol $Å^2$) and no restraints on the solvent. (*iii*) Heating dynamics was performed with continued solute restraints at constant volume (NVT). Thereby, the

temperature was increased from 0 K to 300 K over 60 ps and kept at that value for another 40 ps. (*iv*) Minimization was carried out with reduced solute restraints (2.5 kcal/mole Å^2). (*v*) The system was again heated, as described above, with reduced solute restraints (2.5 kcal/mole Å^2). (*vi*) For monomeric systems 150 ps of constant pressure (NPT) dynamics at 300 K was performed, with isotropic position scaling at pressure of 1 bar and a pressure relaxation time of 0.2 ps. The analogous step for dimeric systems was extended to 300 ps. Harmonic restraints in this run were applied only to the CoA and Mg^{2+} in the system were Mg was present (2.5 kcal/mole Å^2) at 300 K. (*vii*) Finally, an unrestrained NPT simulation at 300 K and 1 bar was performed for a duration of 150 ps for monomeric systems and 500 ps for dimers.

Free molecular dynamics. The equilibrated systems were subjected to long unrestrained MD production runs for minimally 50 ns at constant volume and temperature (300 K), saving the snapshots every 5 ps. Simulations of the monomeric systems were carried out using the GPU code of *pmemd* algorithm available in AMBER 11 and the temperature was controlled by the Berendsen thermostat. The systems containing the dimer were simulated using both GPU and CPU versions of *pmemd* as provided in AMBER 12.[12,18] In these simulations, the temperature was controlled by Langevin thermostat. All the data obtained from the molecular dynamics simulations were subsequently processed and analyzed using the *ptraj* module of the AMBER12 program package.

Steered molecular dynamics. Steered MD simulations at constant volume and temperature (300 K) were used to generate trajectories that correspond to the process of CoA entering the active site. These trajectories can follow different pathways resulting with lower or higher amounts of work required to bring CoA from surface to protein interior. From the data collected by running a series of pulling experiments, it is possible to estimate PMF for a given process. The pulling was unidirectional for all the systems under observation and no reverse trajectories were computed, which implies usage of Hummer and Szabo estimator to obtain PMF from the resulting data.[19] This was achieved by employing a collection of MATLAB procedures written by D. Minh for analyzing biased molecular dynamics simulations and experiments and readily available at www.simtk.org.[20,21] A more detailed description of this method and theory behind it can be found in Section 1.3.4.2.

A reaction coordinate along which CoA was pulled into the active site was defined as a distance between the centres of mass of two groups of atoms; one group is formed by 4 backbone atoms connecting Cys418 and Cys419 (CA, N, C, CA), while the other group is made of 6 heavy atoms at the thiol end of CoA. Four atoms belong to cysteamine (S, C, C, N) and two

atoms come from pantothenate (C, O) moiety (Figure 3.17). The latter group is often referred to as "CoA tail" in this thesis. This distance was not always the same, depending on the choice of the initial set of conformations. The force constant of the harmonic potential used to drive the system along the points of the chosen coordinate was set to 5 kcal/mol Å^2. Duration of each pulling was 500 ps for monomeric systems. An additional set of trajectories with a slower pulling speed in duration of 2 ns was computed for mAcPFL system to make comparison between distributions of work values obtained with different switching speeds. There is a significant increase in size when a complete dimeric protein is used instead of just a single subunit, which reflects on the pulling speed. To reach comparable switching speeds to those achieved in monomeric systems, the pulling in the dimeric systems was done during 4 or 5 ns simulation.

The starting ensemble of conformations for the pulling experiments were mostly generated by running restrained MD dynamics during which the system was kept fixed to the chosen value of the reaction coordinate by harmonic potential (5 kcal/mol Å^2). These restrained simulations were started from an adequate snapshot taken from free dynamics, which had the most promising conformations of protein and CoA that would lead to low-lying trajectories. In case of dimeric systems, in addition to the restrained dynamics, it was also possible to sample the initial set of coordinates by taking snapshots directly from the free dynamics that possessed desired properties, i.e. the desired value of generalized coordinate. Usual duration of the restrained dynamics carried out with monomeric systems was 2.5 ns and snapshots were taken every 50 ps, while dimeric simulations lasted on average 5 ns and snapshots were collected every 100 ps.

Umbrella sampling. The umbrella sampling procedure was performed for monomeric systems only due to the large computational costs associated with this approach. The reaction coordinate remained identical to the one defined for steered dynamics simulations. In the AMBER implementation of this approach, sampling in the importance regions is achieved by placing centres of harmonic potential along the chosen coordinate to ensure sampling in the specified interval or "window"(see Section 1.3.4.1). Each window is characterized by the force constant and value of the generalized coordinate defined as a centre of the imposed restraint (Table 3-2). Starting conformations for all the window simulations were taken from the lowest-lying SMD trajectory and subjected to biased NVT simulation in duration of 2.5 ns. Data points were collected every 2 ps. The first 0.5 ns of the simulation were taken as equilibration period and data collected in this interval was discarded in data analysis. All the simulations were carried out in the AMBER 11 package.

Table 3-2. The placement of the umbrella potentials along the reaction coordinate, which is defined as a distance between two groups of atoms (four heavy backbone atoms connecting Cys418 and Cys419 and six heavy atoms of CoA thiol tail, see Figure 3.17), starting from 35.0 Å and decreasing to 7.2 Å. The table contains the force constants of the harmonic potentials imposed on the system and centered at the position given in the middle column. The third column provides the range of the reaction coordinate in which these force constants were applied.

Force constant (kcal/mol Å^2)	Window spacing (Å)	Reaction coordinate interval r(Å)
4.0	1.0	35.0-29.0
4.0	0.7	29.0-26.0
6.0	0.5	26.0-18.0
10.0	0.3	18.0-7.2

To compute PMF for the process of CoA entering the active site, the applied bias needs to be removed from the probability distributions obtained from each window. The unbiased distributions from all the windows are combined to retrieve the underlying PMF. There are several estimators that are able to perform this task and weighted histogram analysis method (WHAM) is by far the most popular.[22] A convenient code to analyze data obtained from AMBER simulation software was developed by A. Grossfield,[23] and it was used in this research. Umbrella integration (UI) is another unbiasing method used for PMF estimation, originally developed by Kästner and Thiel and Fortran script is available upon request.[24] Finally, the most recent method to estimate PMF from umbrella sampling approach is called multistate Bennet's acceptance ratio method (MBAR), developed by Shirts and Chodera.[25] Python implementation of this method can be found on www.simtk.org.[26] More about these methods and their theoretical background can be found in Section 1.3.4.1.

3.4.2 Results and Discussion

PFL catalyzes the break down of pyruvate into formate and the acetyl group upon the addition of a thiyl radical located at Cys418. The radical is initially stored at Gly734 is shuttled to Cys418 via Cys419. The addition of radical Cys418-S$^{\bullet}$ to pyruvate leads to C-C bond dissociation, resulting with formation of formyl radical and acetyl-Cys418. The latter species acts as a temporary acetyl carrier and a reactant in the subsequent half-reaction with the second substrate CoA to produce acetyl-CoA. Formation of acetyl-Coa, the final product, closes the catalytic cycle of PFL.

The investigated aspect of this mechanism concerns the process that allows CoA to enter the active site, which is a prerequisite for the second half-reaction. The problem with this step is that the binding site of CoA is located at the protein surface, while the active site is buried in the protein interior. In search for possible solutions to this problem, the PFL system was subjected to long unrestrained molecular dynamics simulations, followed by the free energy calculations used to estimate potential of mean force for the process of CoA approaching the active site while the pyruvate is still intact (PFL) and after it has been cleaved (AcPFL).

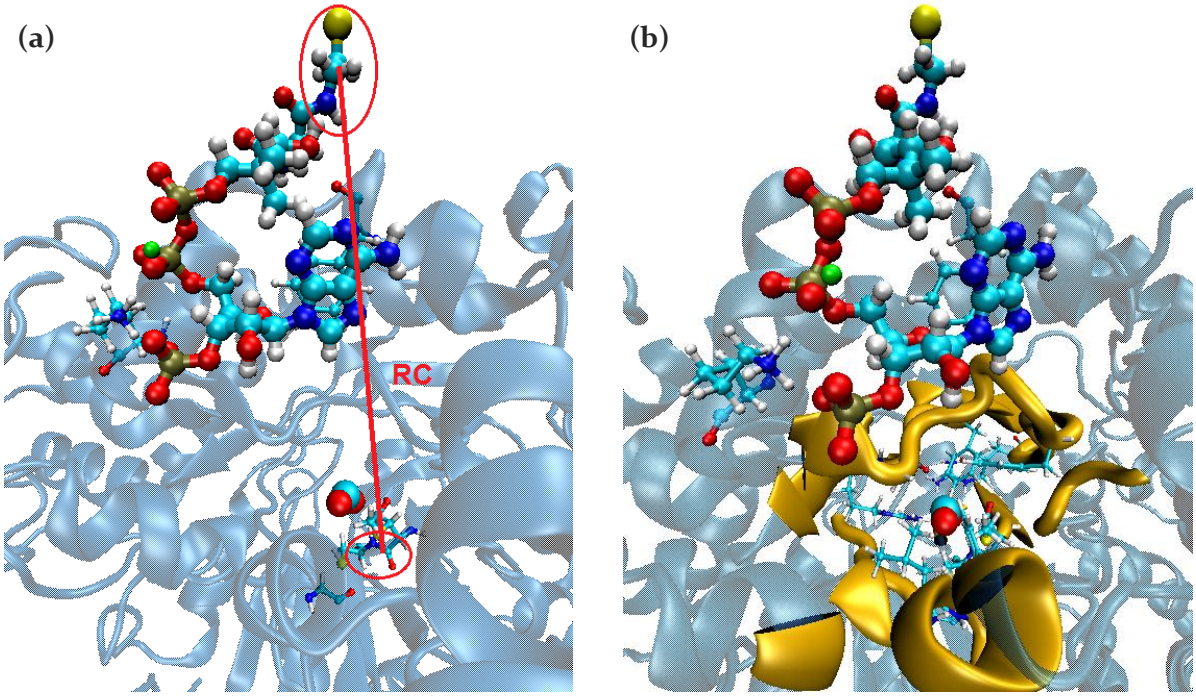


Figure 3.21 (a) The reaction coordinate (RC) used in the free energy calculations was defined as a distance between centers of mass of two atom groups – one group comprises of 6 heavy atoms in the CoA tail, while the other contains 4 backbone atoms connecting Cys418 and Cys419 (circled); (b) A possible channel that might accomodate CoA while approaching the active site along the chosen reaction coordinate (shown in mAcPFL system).

To undertake free energy calculations, one must choose an adequate reaction coordinate (r) that would be capable of capturing the key features of the process of interest in the free energy landscape. Definition of the suitable reaction coordinate is far from trivial and often there is a hidden variable that has a significant control over the process, but it is not easily recognized. In the free energy calculations done on the PFL system, it was necessary to find a suitable reaction coordinate to describe the pathway CoA follows from the protein surface to the buried active site in terms of free energy. A reasonable choice of this coordinate for the given problem was the distance between cysteamine moiety of CoA, which contains the reacting thiol group, and protein backbone atoms connecting catalytic residues Cys418 and Cys419 (Figure 3.21a). The choice was made based upon the expectation that this would be the variable that changes the most during the process in which the thiol group of CoA approaches the catalytic cysteines. This variable was used as a reaction coordinate in all the free energy calculations carried out in this study.

The models used in the simulations were built by using a single protein subunit, but also a full dimer. Two sets of dimeric models were created; one set contains sodium ions and a putative magnesium ion, while the other set contains sodium ions only (Table 3-1). The results obtained from the molecular dynamics simulations are divided into sections based on the models used in the simulations.

3.4.2.1 Monomeric Models

The results presented in this section come from the simulations carried out with models derived from a single monomer of otherwise dimeric PFL protein. Namely, PFL consists of two identical subunits interacting through the salt bridges formed at the interface, but only one of them becomes activated upon reaction with PFL-activase. The monomeric models could potentially provide an adequate approximation to the PFL system, while significantly reducing the concomitant computational costs of the free energy calculations. Basically, they were built as test models to obtain some preliminary results from free energy calculations before engaging into computationally demanding calculations based on dimeric models. The PMF for the process of CoA approaching the active site while bound to the protein surface was calculated using steered molecular dynamics and umbrella sampling technique, preceded by the extensive free dynamics simulations at constant volume and temperature (NVT).

Free dynamics. After equilibration, a 50 ns production run was performed for each of two monomeric models representing the PFL system before (PFL) and after (mAcPFL) the first half-reaction. In that period, none of the models displayed major structural changes compared to the crystal structure. The RMS deviations of the entire protein from the crystal structure during simulations are given in Figure 3.22, which clearly shows that during 50 ns of simulations no large movements of the protein domains in the examined model systems took place. The most pronounced fluctuations can be assigned to the sequences that form loops exposed to the solvent, but this is a typical feature of the more flexible units and the effect on the overall protein structure is almost negligible.

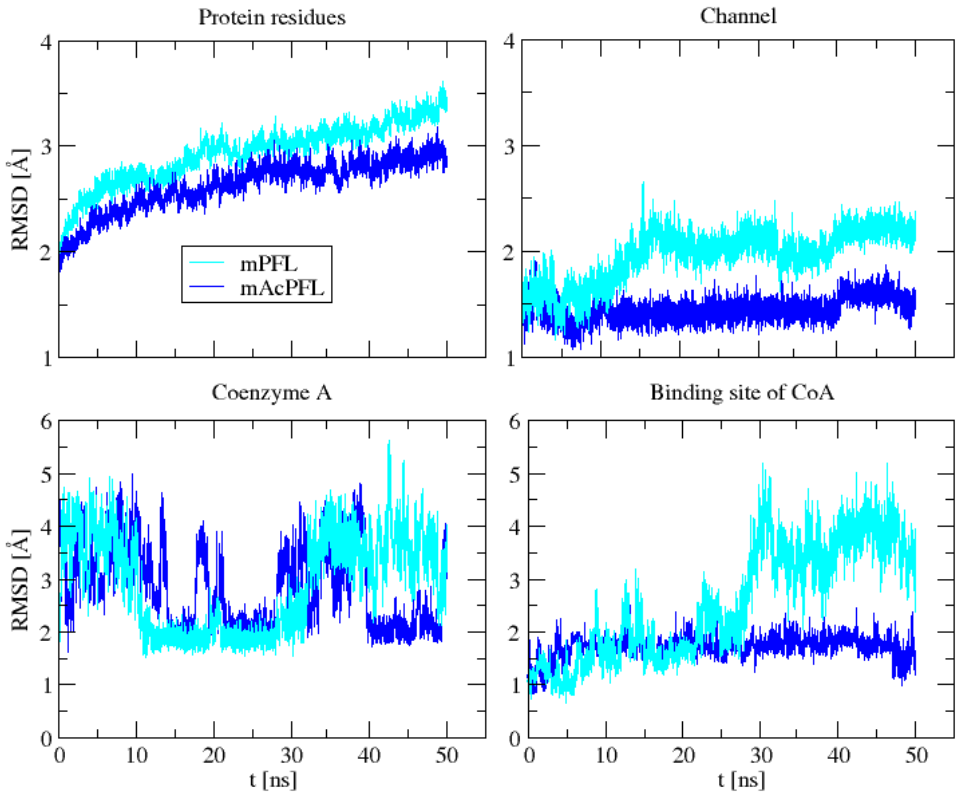


Figure 3.22 The RMS deviations of the selected items from the crystal structure. These items might play a role in the process under investigation, which is could bring the CoA initially bound to the protein surface into the active site buried in the protein interior. The possible channel is determined by 8 residues (Gly167, Tyr172, Arg176, Tyr323, Leu326, Phe327, Phe432, Arg435), while the CoA binding site by 4 residues (Asn145, Gln146, Phe149, Lys161).

The RMSD for movement of CoA and its binding site is also given in Figure 3.22 to monitor the behavior of the key residues for the investigated process during simulations. The RMSD values of CoA look quite similar for mPFL and mAcPFL, but the difference is visible in the motion of the binding site. The large deviations occurring after 30 ns of the mPFL simulation correspond to the process of CoA unbinding, which by the end of the simulation interacts only with Lys161

via 3' phosphate (Figure 3.24). The substrates in the active site, pyruvate and formate, are tightly bound during entire simulation.

Another parameter was chosen and plotted to provide an insight in the motion of the residues that form the most plausible channel for the entrance of CoA considering its binding mode (Figure 3.22). This channel is defined by three residues that participate in substrate binding – Arg176, Phe432 and Arg435; and five more residues placed above the active site, which form sort of a gateway: Gly167, Tyr172, Tyr323, Leu326, and Phe327. There are three residues with aromatic sidechains stacked together and forming a sort of spiral lid that covers the active site (Figure 3.23), while Gly167 and Leu326 are sitting at the top of two opposing loops, which keep the channel closed. These loops are part of two longer sequences that enclose this channel, ranging from residue 166-176 and 321-333. The "gate" is positioned approximately 15 Å above the catalytic cysteines.

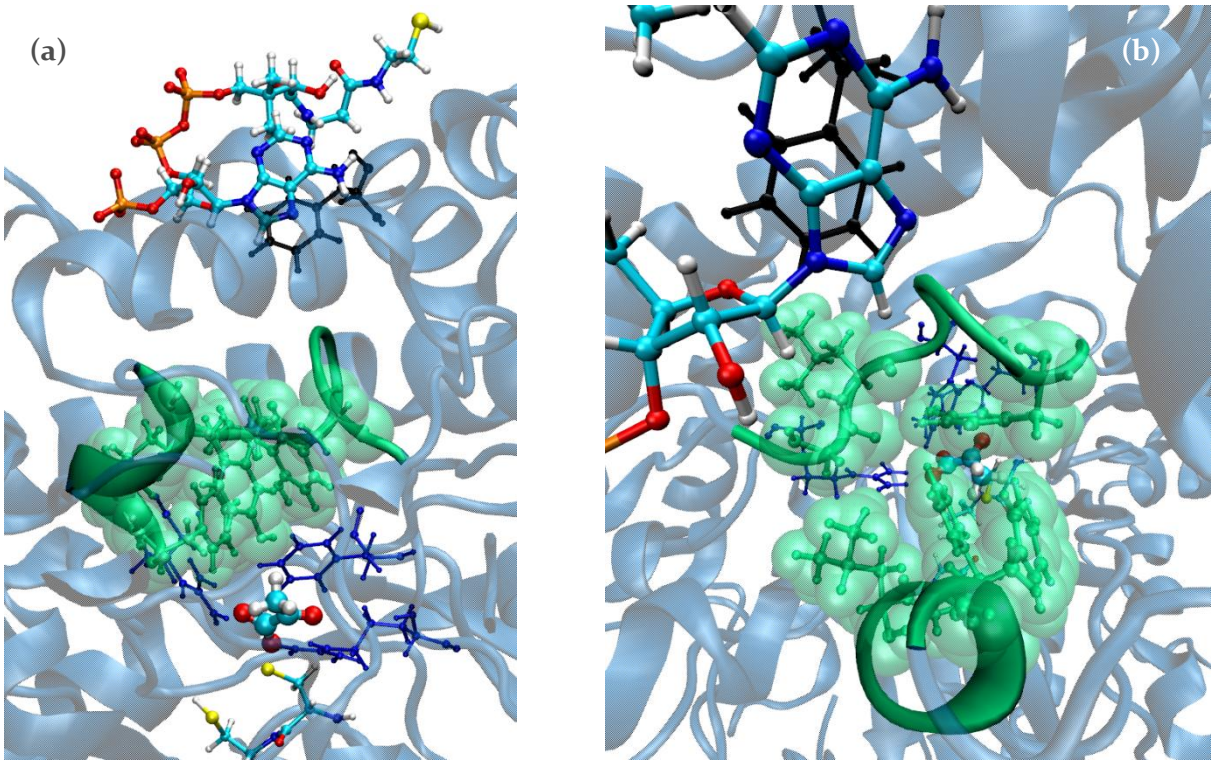


Figure 3.23 The "gateway" (Gly167, Tyr172, Tyr323, Leu326, Phe327 - green) and the binding residues (Arg176, Phe432, Arg435 - violet) that protect the active site keeping the prospective channel closed viewed sideways (a) and from top (b).

Based on the MD simulations, it seems that more pronounced movement of the residues takes place in mPFL system than mAcPFL, but the difference is not so extravagant. The most notable difference between the two systems is the unbinding of CoA that took place in the mPFL system, but not in the mAcPFL. A snapshot of the simulation endpoint is given in Figure 3.24, where the change in the binding of CoA in mPFL can be seen from a perspective looking

down the prospective "channel". The channel, however, remains closed in both model systems and the forming residues display only moderate deviation from their original positions in the crystal structure. The same conclusion can be basically applied to the motion of the entire protein. The lack of the expected structural changes can be attributed to several reasons, including the choice of the model system and the too short time scales represented by the performed simulation. Nevertheless, the free energy calculations were carried out to examine if there is any difference between free energy profiles resulting from driving the CoA down the prospective channel before and after the first half-reaction.

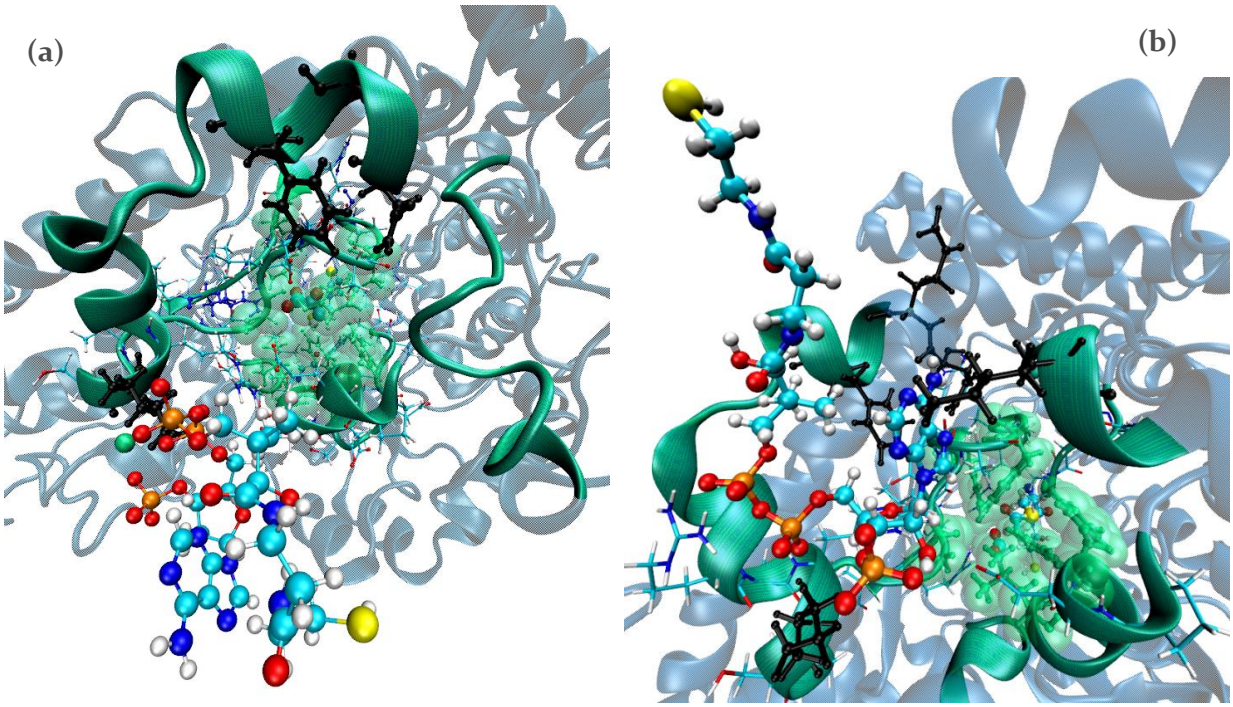


Figure 3.24 A view down the possible entrance channel after 50 ns free MD simulations from the protein surface: (a) mPFL model; (b) mAcPFL model.

Steered molecular dynamics. The theory underlying the steered molecular dynamics (SMD) is based on the Jarzyinski equality that relates the free energy difference (ΔF) between two equilibrium states with the amount of work (W) necessary to drive the system from the initial to the final state during a non-equilibrium process. The switching process must be repeated many times and the exponential average of the work values obtained from numerous trajectories should approximate the free energy difference. The average is closer to the exact ΔF if a large number of switching experiments (trajectories) is available for statistical reasons, but also if the switching is done at lower rates (Figure 1.6).

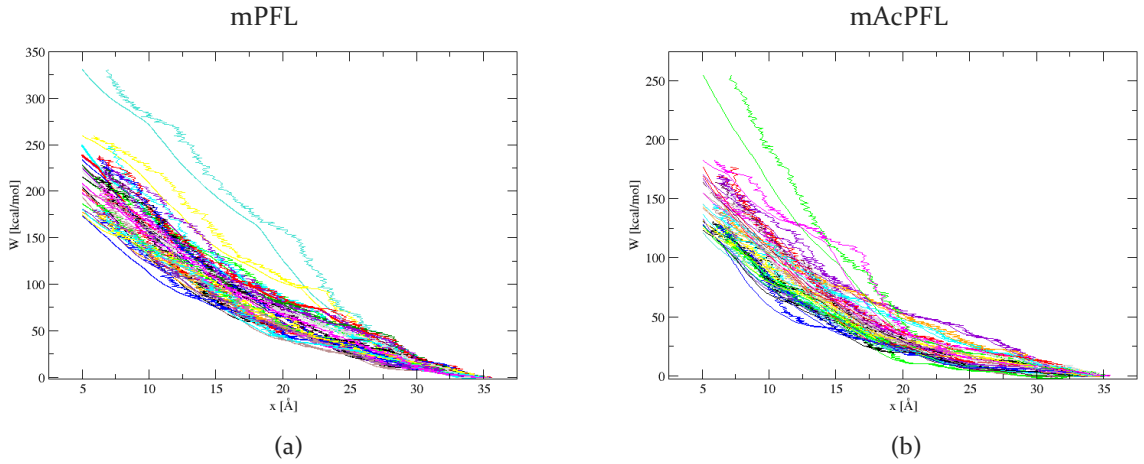


Figure 3.25 A set of 25 trajectories for each topology, mPFL (a) and mAcPFL (b), was obtained by running SMD simulations, in which CoA was pulled toward the active site cysteines during 500 ps. The smooth lines represent the amount of work as a function of the predefined values of the reaction coordinate (r) for each run, while the ragged lines depict work as a function of the actual values of the chosen variable (x) during the pulling experiments.

In this set of simulations, the cysteamine group of CoA was pulled from the protein surface closer to the active site in 25 pulling simulations for each monomeric system representing PFL system before and after pyruvate cleavage – mPFL and mAcPFL, respectively. CoA was pulled along the reaction coordinate in total length of 30 Å, starting from the initial distance at 35 Å to the final value of 5 Å, during 500 ps run. The initial ensemble of configurations was generated by running restrained dynamics, where restraint was given in the form of harmonic potential to keep system in a close region of the given initial value of reaction coordinate. The system was driven from the initial to the final state by applying a harmonic restraint (5 kcal/mol Å^2), which forces the system to explore the phase space around the chosen reaction coordinate. This restraint moves along the reaction coordinate in a time-dependent fashion, where each point on the reaction coordinate act as a centre of this harmonic potential. This allows to each pulling experiment to take a different transition pathway from the initial to the final state, resulting each time with a different amount of work required for that process. The trajectories obtained for mPFL and mAcPFL are shown in Figure 3.25, where the amount of work was plotted both as a function of the predefined points of the reaction coordinate (**r**) and the actual values of the chosen variable at that point (**x**). The plots show how the actual values of the variable are dissipated around the desired value, as the system explores the available neighbouring space as much the imposed restraint allows. Of course, a larger deviation from the targeted value corresponds to an increase in the amount of irreversible work invested to drive the system along the chosen path. By closer inspection of these plots, it can be noticed

that both systems stop at 6 Å of the reaction coordinate, unable to come closer. Basically, at this point the thiol group of CoA is already too close to Cys418 and Cys419 and there are sterical problems associated with closer approach. A better choice of the end point of reaction coordinate would probably be between 8-10 Å, where the thiol group is aligned with the substrate in the active site.

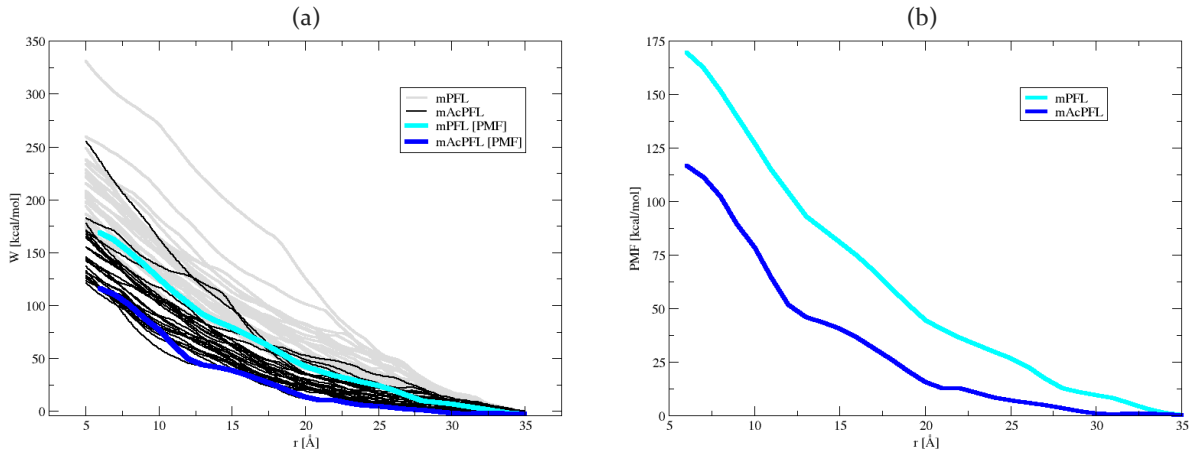


Figure 3.26 Trajectories obtained from the SMD simulations in which CoA is pulled toward the active site before and after pyruvate cleavage (a) overlapped by the estimated PMF. The final PMF was estimated by using Hummer and Szabo expression (b).

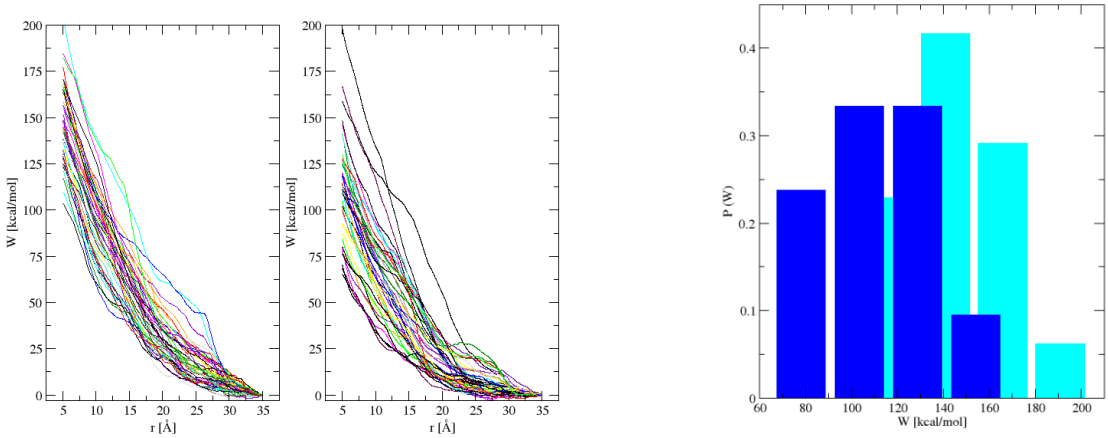
From the presented trajectories, it is visible that larger amounts of work are required to drive the CoA from the protein surface closer to the catalytic cysteines in the mPFL model that corresponds to the PFL system before the first half-reaction takes place in the active site and the pyruvate is still intact. In general, the trajectories obtained from SMD simulations with mAcPFL model are lying lower than those obtained with mPFL model. This is even more obvious when both sets of trajectories are overlapped, as in Figure 3.26a. The PMF for each topology was extracted from the given trajectories by using Hummer and Szabo's estimator for unidirectional pullings. Comparison of two curves clearly shows that bringing CoA into the active site is easier for mAcPFL topology, although no significant conformational changes were observed in any of the model systems during the free dynamics simulations that might assist that process. However, the difference between the conformations of CoA in mPFL and mAcPFL exists, despite the choice of the similar conformations for both topologies as starting points for restrained dynamics simulations. During the restrained simulations, in the mAcPFL system CoA adopted more extended conformation, while in mPFL system it was found in a partially bundled form. This form allows self-interaction between different parts of CoA when the molecule is pulled toward the active site, which in turn increases the amount of the required work. The extended form of CoA in mAcPFL system also had interactions with the

neighbouring protein residues, but those were easier to overcome by pulling. The difference in the starting conformations is reflected the most in the initial portion of PMF between 35-25 Å of the reaction coordinate, where more energy is required to pull the CoA tail from a bundled conformation in mPFL compared to the extended form found in mAcPFL. In the region between 20-25 Å, the CoA takes up a similar U-shaped form in both topologies, resulting with two curves that end with similar, but not entirely parallel slopes.

The presence of this moderate divergence between two curves indicates that there is also a difference between the systems once the CoA enters the channel which approximately corresponds to a distance of 20 Å distance along the reaction coordinate, hinted by a small minimum on mAcPFL curve. Namely, around that point, the thiol group of CoA starts to interact with the "gate" residues, while the carbonyl group of pantothenate moiety usually makes a hydrogen bond with Arg160. A second hint of minimum on mAcPFL curve corresponds to the interaction of thiol group with the aromatic sidechains of the "gate" residues. A possible explanation of these two minima and the overall lower free energy profile obtained from mAcPFL could be found in the additional flexibility introduced in the active site after the first half-reaction, in which the rigid C-C bond in pyruvate is replaced by a looser configuration involving a small formyl radical and acetylated cysteine. The new arrangement should allow the residues that participate in the substrate binding more flexibility, but possibly also facilitate the accommodation of CoA in the entrance channel and the active site. However, the channel was closed in both model systems, resulting with the uphill profiles as the CoA had to penetrate through the hindering residues. This result is also interesting in light of the free dynamics, where mPFL system displayed mildly larger deviations from the initial structure compared to mAcPFL system.

Not only that the profiles are uphill, the quantification of the free energy associated with the process of CoA approaching the active site also provides quite high values under the given conditions. The large amounts of work required to drive the described process in both model systems are to a certain extent a consequence of the switching rate used in the simulations. The decrease in the amount of work related to the process under examination can be attained by decreasing the switching rate, as shown in Figure 3.27. In this figure is presented the outcome of the SMD simulations in duration of 500 ps and 2 ns, performed on mAcPFL system. The histograms on the right panel right clearly show that the longer simulations decrease the average value of the total work for approximately 35 kcal/mol under given simulation conditions. The slower switching rate also allows better sampling of the phase

space around the chosen reaction coordinate, but the price of these improvements is paid in a form of the increased computational costs to obtain the same number of trajectories.



(a) (b)

Figure 3.27 (a) Comparison of trajectories resulting from SMD simulations in duration of 0.5 ns (left) and 2ns (right); (b) Distribution of the total work required to drive the process from the initial to final state obtained at different switching rates.

Either way, 25 trajectories are far away from being enough to provide an adequate statistical sample and the results presented thus far can only be viewed from a qualitative perspective. From that standpoint, there is a clear difference between two topologies, indicating that the pyruvate cleavage plays a role in the approach of the second substrate to active site.

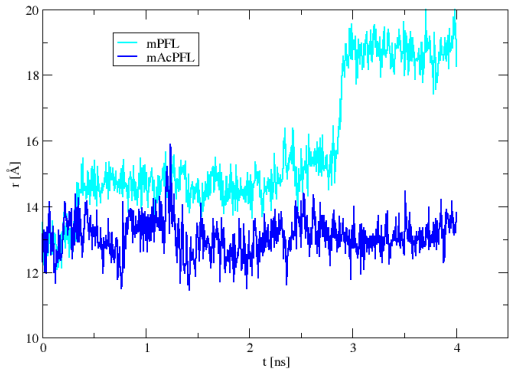
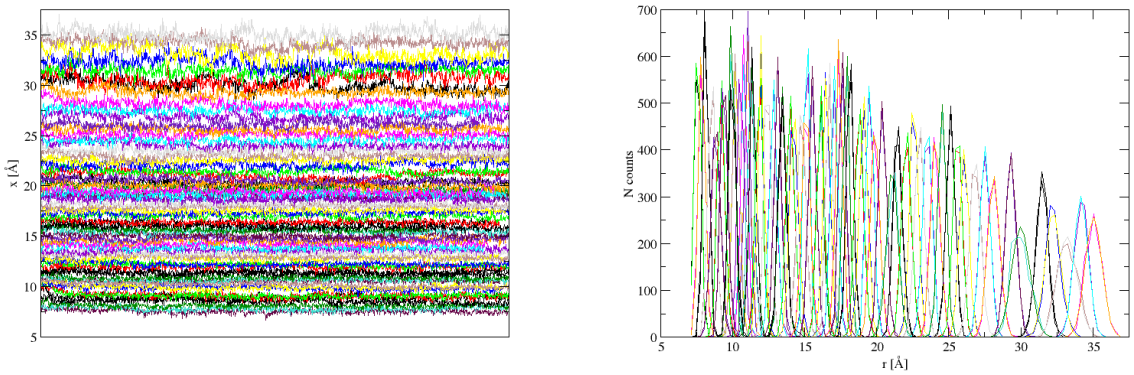


Figure 3.28 Comparison of the reaction coordinate values during 4 ns of free dynamics of mPFL and mAcPFL system started from an end point of SMD pulling experiment, where CoA was inside the channel.

A free dynamics simulation was performed for both topologies starting from the end point of the SMD pullings in duration of 4 ns to examine the stability of the system state with CoA inside the channel. It seems that CoA is able to stay inside of the protein during entire simulation if the pyruvate is cleaved, while the presence of the intact pyruvate leads to withdrawal of CoA back to surface after 3 ns (Figure 3.28). This is additional support to the claims that mAcPFL constitutes a more favourable environment for the entrance of CoA.

Umbrella sampling. The umbrella sampling approach also uses harmonic potential to restrain the system to explore the desired region of the phase space. In contrast to SMD, the centre of this harmonic potential is not time-dependent and remains fixed during the entire simulation. Each centre is called a "window" and they are adequately placed along the reaction coordinate to achieve overlapping probability distributions of the chosen variable between neighbouring windows (see Section 1.3.4.1). The lowest lying trajectory from each set obtained with SMD simulations for mPFL and mAcPFL systems was used as a source of the initial coordinates for each window used in this experiment. The snapshots corresponding to the selected values were used to initiate a series of the restraint MD simulation in duration of 2 ns (Figure 3.29). The sampled data presented in Figure 3.29 comes from the simulations carried out with mPFL system and almost identical plots are obtained for the mAcPFL model (not shown).



(a) (b)

Figure 3.29 A set of raw data collected during "window" simulations performed for mPFL system (a), where distances between CoA tail and active site cysteines were sampled in range between 7-35 Å. The resulting histograms from the raw data are given on the panel right (b).

The collected statistics for both model systems, mPFL and mAcPFL, were post processed to extract the potential of mean force. Several methods were used for that purpose and the most popular one is weighted histogram analysis method (WHAM). The underlying assumption of this approach is that the best estimate of the reduced probability distribution of the system would be a linear combination of the adequately weighted probability from each window. Some of more recent and not so widely used estimators include umbrella integration (UI) and multistate Bennett acceptance ratio method (MBAR). More about theoretical background of these estimators can be found in Section 1.3.4.1.

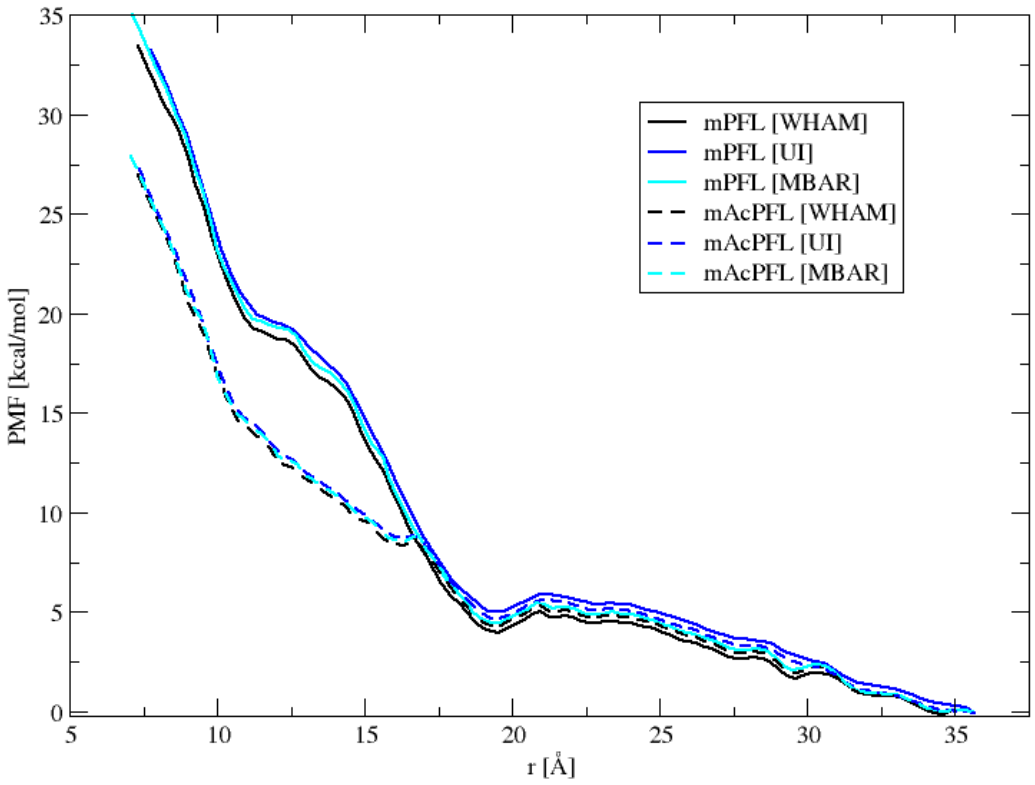


Figure 3.30 Potential of mean force for pulling CoA closer to the active site cysteines before (solid) and after (dashed) pyruvate cleavage obtained from the umbrella sampling experiment. The three curves correspond to three different estimators used to calculate PMF from the same datasets.

As shown in Figure 3.30, all three estimators give similar results in terms of the final potential of mean force for both topologies and all three agree that mAcPFL provides more favourable environment while CoA is approaching the active site. This is in agreement with the general outcome of the SMD simulations, but in comparison, the umbrella sampling method provided a free energy surface with a notable lowering toward more reasonable numbers. The interesting feature easily noticed is how two profiles basically overlap in the region between 17-35 Å, which roughly corresponds to the conformations in which thiol group of CoA still freely moves in the solvent (35-20 Å) or it has just started to interact with protein residues located at the top of the entrance channel (20-17 Å). The observed minimum corresponds to the point at which thiol group of CoA starts to interact with the aromatic sidechains of the "gate" residues. The minimum was spotted for mAcPFL curve obtained from SMD calculations, but not for mPFL. However, the umbrella sampling approach encompasses thorough sampling along the reaction coordinate compared to the SMD method (only 25 trajectories) and the minimum was captured in both cases. The better sampling also provides an explanation for the overlap of the curves in the mentioned region, surprising only at the first glance. Namely, the system

conformations sampled in this region (35-17 Å) are quite similar for mPFL and mAcPFL models, with free CoA tail and the interactions with the "gate" residues closing the channel. However, once CoA actually penetrates through the interfering gate residues, there is an obvious difference between two systems. Here comes firmer evidence that the reaction with the first substrate (pyruvate) mitigates the approach of the second substrate (CoA). The observed difference is in free energy profiles is not drastic (~ 5 kcal/mol) and considering that the both profiles are uphill, it makes one wonder if there is something else that affects this process, but that is not entirely captured in this setup. This was part of the motivation to engage into simulations employing dimeric models, with an assumption that the presence of the other subunit might influence conformational behaviour of PFL. Of course, conformational changes are of crucial importance for the investigated problem.

3.4.2.2 Dimeric Models

Dimeric models were devised to make a better representation of the PFL system. Namely, PFL appears as a homodimer in which two subunits are related by a two-fold symmetry axis, but only one is catalytically active. The two subunits are kept together via electrostatic interactions. In addition to the existing salt bridges formed between the residues close to the interface, an important observation is that the binding site of CoA extends over both subunits, according to the available crystal structure (Figure 3.20).

Namely, the nucleotide moiety is anchored by interaction with Phe149, Asn145, Gln146 and Lys161 residues, while the thiol group rests between Phe220 and His227 sidechains of the opposing monomer. Considering the fact that both subunits participate in the CoA binding and the possibility that the other subunit somehow affects the conformational behaviour of the activated monomer, it seemed reasonable to engage into simulations based on dimeric models.

The models were again built to represent the system before and after the first half-reaction in the same way as for the monomeric models. Two sets of dimeric models were built, differing only in the presence of a putative magnesium ion identified in the proximity of the CoA phosphates. There are several reasons that raise a reasonable doubt regarding the correct assignation of the density peak located nearly to the CoA phosphates. The buffers used in the protein crystallization contained only monovalent cations (Na^+, K^+, Li^+), while Na^+ and Mg^{2+} are very difficult to distinguish in crystallography. The identification relies on the coordination

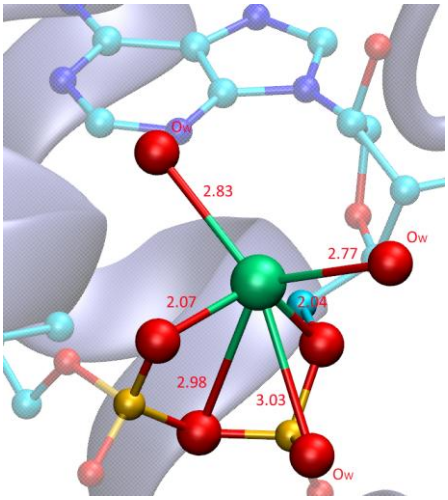


Figure 3.31 Coordination of Mg^{2+} ion with 3 phosphate oxygens and 3 water oxygens (Ow), according to authors that solved the PFL crystal structure (1H16).

number of ligands and the distance between the metal and a ligand, usually water. The Mg^{2+} ion is typically characterized by octahedral coordination (in 70 % cases) and an average Mg^{2+}...O distance is approximately 2.1 Å, according to the server designed to browse PDB structures for metals and generate statistics.[27,28] The coordination of an ion presented in Figure 3.31 can hardly be considered as octahedral. The tetrahedral coordination might be a better description of the given structure, where two ligands are phosphate oxygens and two ligands are water molecules at distance 2.77 and 2.83. On the other hand, the Na^+ ions show more diversity in the coordination number, although the most common numbers of ligands are 5 (26 %) and 6 (30%). Less common coordination numbers are 2 and 3 (both ~10%) and 4 (17 %). The average distance with oxygen atoms is 2.5 Å. In the most cases the ligand is water so the major contribution comes from the distance between metal and water oxygen. There is also quite possible that this is potassium ion, based on this criterion, where preferable coordination numbers of K^+ are 4 and 6 and typical K^+...O distance is around 2.7 Å (see Figure 3.31). However, it is reasonable to expect that K^+ would be easier to differentiate from Na^+ and Mg^{2+} based on the electron density. In either case, the idea was to replace a highly charged Mg^{2+} ion with a monovalent cation and it was decided to proceed with Na^+, because the rest of the ions present in the PDB file were characterized as Na^+ and the same ions were used to neutralize the system in the model building procedure.

This results with one set of the models includes the Mg^{2+} ion (dPFL and dAcPFL), while the second set (dPFLX and dAcPFLX) contains only sodium ions originating from the PDB file or those added to enforce the system neutrality. All the models were subjected to rather long free dynamics simulations (NVT), followed by a series of steered MD simulations.

3.4.2.2.1 PFL Model Set (with Mg^{2+})

Free dynamics. Free dynamics simulation with dimeric models provided some interesting results, especially in case of dAcPFL model, but they also opened a door to some new

difficulties. One of the first observations that can made by visualizing the resulting trajectories is the conformation of CoA, now deprived of the ability to move as freely as in monomeric models. The loss of motility is a direct consequence of the interactions of CoA thiol group with the aromatic sidechains of Phe220 and His227 belonging to the inactive subunit. This interaction is quite strong under the given conditions and CoA remained bound to this aromatic "sandwich" in all of the free MD simulations carried out with both model systems (dPFL and dAcPFL). Occasionally, CoA would briefly disengage only to become recaptured again between two aromatic rings. This opened a question of how CoA is released from this position, considering the stability of this interaction. Several variables were selected to monitor the system dynamics, usually via RMS deviations from the crystal structure position, which was chosen as reference to keep the values in both systems comparable. The results were plotted for the active and inactive subunit. The inactive monomer was used as a control of the system dynamics, as it provides a correct description of the state in which protein was crystallized and in which no chemical changes took place.

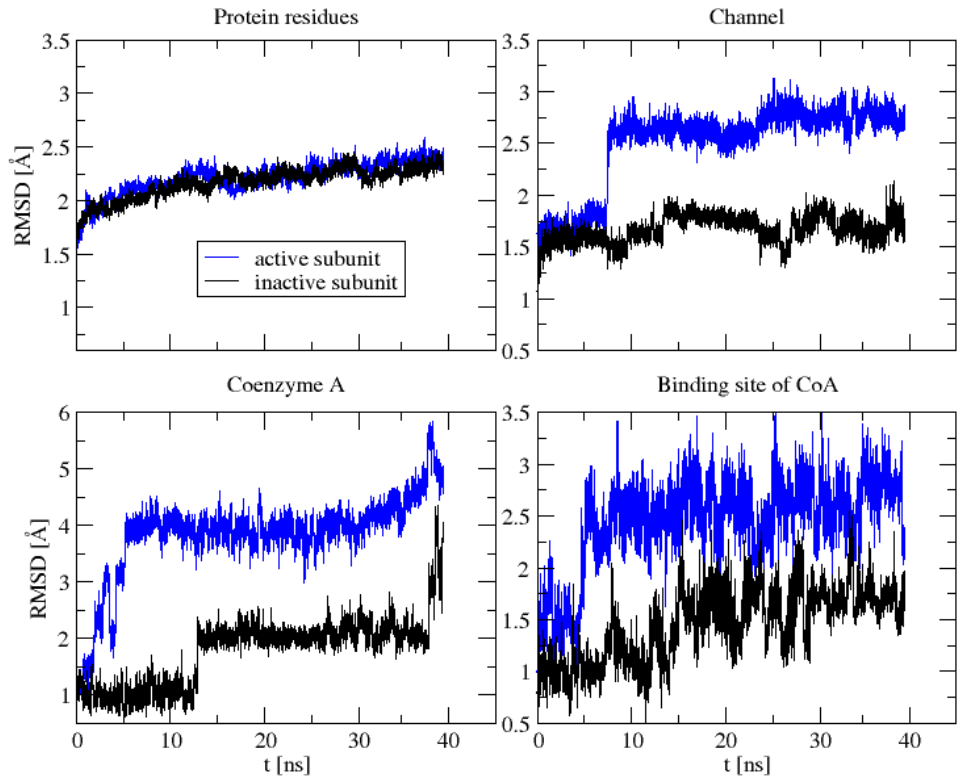


Figure 3.32 The RMS deviations of the selected variables in the active and inactive subunit of dAcPFL model during the simulation 1 (40ns) compared to the crystal structure.

The simulations containing dimeric PFL showed to an interesting development in a comparison with their monomeric counterparts. In the 40 ns long simulation (1) carried out on

dAcPFL model, an interesting rearrangement of the sidechains of the "gateway" residues was observed. It is visible in the RMSD plot given in Figure 3.32 that the behaviour of the protein residues is generally alike in the active and inactive subunit. However, the mentioned gate sidechains in the crystal structure are positioned in a way that prevents breach of the unwanted molecules close to the active site. In one of the simulations performed with dAcPFL system, a certain rearrangement of the hydrogen bond network occurred after 10 ns, resulting with new positions of these hindering sidechains (Figure 3.32). This is not, by any means, a dramatic conformational change involving drastic movement of helices, but it may be enough to help in accommodation of CoA on its way in. The average value of RMSD for this parameter in the active subunit is around 2.7 Å, compared to the average value of 1.5 Å resulting from the simulation involving monomeric mAcPFL.

In an attempt to reproduce this phenomenon, two additional independent simulations were initiated and ran for 30 ns and 55 ns – simulations 2 and 3, respectively. The shorter simulation did not result with similar rearrangement, although the dAcPFL system again displayed significant fluctuations of the channel residues (Figure 3.33). However, in the longer simulation 3 an incident leading to the channel widening occurred again, but in somewhat different fashion. Namely, this time the two neighbouring loops that carry Gly167 and Leu328, which belong to "gateway" residues, decoupled for approximately 10 ns before they came back closer again. In that interval, the active site was left more exposed and easier to access. This is visible as a bump between 30 ns and 40 ns in the RMSD plot for the channel residues.

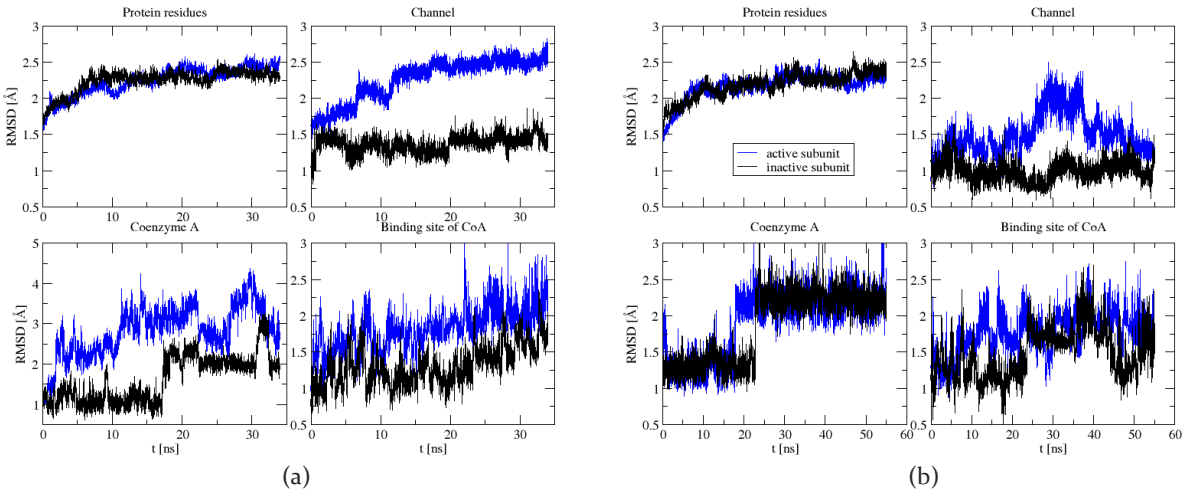


Figure 3.33 The RMS deviations of the chosen variables in the active and inactive subunits of dAcPFL model during free dynamics simulation 2 (a) and 3 (b), in comparison to the crystal structure.

Another interesting observation concerns the movement of the CoA itself, which considerably changes the binding mode of its nucleotide moiety during the first simulation

with dAcPFL model. Basically, the Lys161 ends up bound to only 5' phosphate group, while the 3' phosphate starts to interact with another lysine residue, Lys115, which belongs to the neighbouring helix. As a consequence, the adenine base is pulled away from Phe149 and hydrogen bonds with Asn145 and Gln146 are broken, resulting with an S-shaped conformation of CoA (see Figure 3.36). All these changes are reflected in the RMSD plot of CoA, but not so much in the plot of RMSD for the binding residues, which remain in the similar position.

The subtlety of the observed changes makes it quite difficult to find a single parameter that would be able to capture this rearrangement and quantify it in this way. As visible from Figure 3.33, the shorter simulation exhibits greater fluctuations of the channel residues than the longer one, but it is difficult to assign that motion as being related to the channel opening. A similar observation can be made for the results of 55 ns long simulation performed with dPFL, where there exists significant increase in channel residues fluctuations, but there is no channel opening of any kind (Figure 3.34).

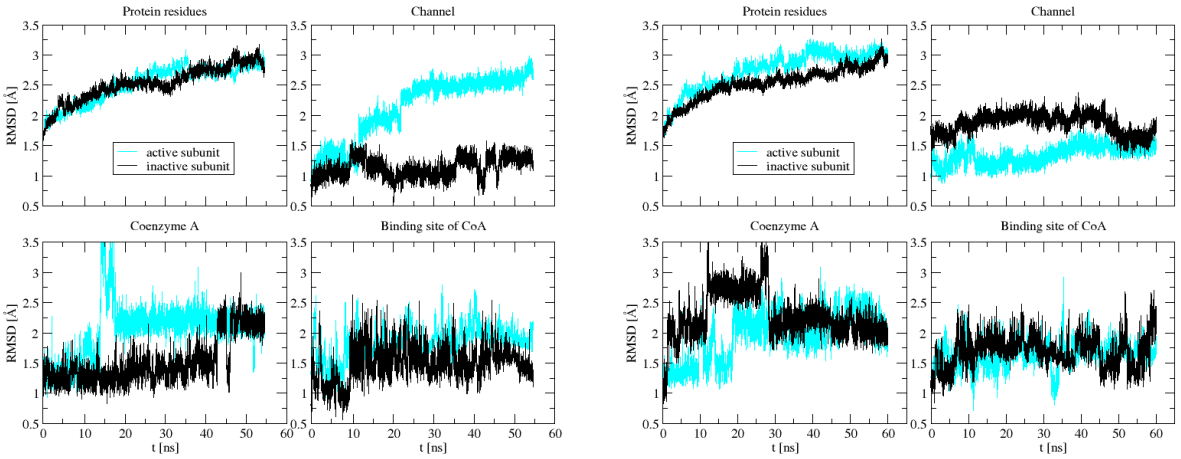
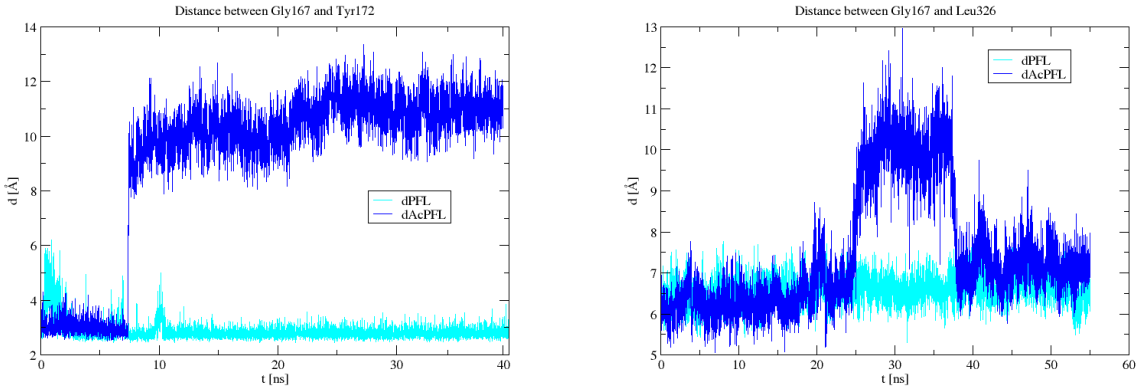


Figure 3.34 The RMS deviations of the chosen variables in the active and inactive subunits of dPFL dimer model during 55 ns (a) and 60 ns (b) of free dynamics simulations, in comparison to the crystal structure.

Actually, the only significant difference between the active and the inactive subunits occurs in the behaviour of the channel, whereas all the other monitored variables display comparable fluctuations. This is even more obvious in the plot obtained from another independent simulation in duration of 60 ns. The latter simulation was performed to verify the results of the first simulation, considering that in the monomeric models mPFL system was more prone to fluctuations than mAcPFL. In the end, all the performed simulations employing dimeric models show that the active subunit of dAcPFL system undergoes stronger structural changes than its counterpart in dPFL.

To further illustrate the channel openings noticed in two independent dAcPFL simulations, additional parameters that changed markedly during the simulations are given in Figure 3.35 and compared to the corresponding values in dPFL system. Both of these additional parameters present a distance between the selected residues that constitute the prospective channel that reflect the observed rearrangements. One of them is distance between Gly167 and Tyr172, which are usually connected via hydrogen bond formed between carboxyl oxygen of Gly167 and the hydroxyl group of Tyr172. That bond is lost during the channel opening in the first simulation and Tyr172 forms a new hydrogen bond with the hydroxyl group of Thr320. The channel opening taking place in the simulation 3 is best characterized by plotting a distance between Gly167 and Leu326. Both of the residues sit on the top of two opposing loops that are part of two longer sequences enclosing the channel (166-176 and 321-333) and the bulkiness of Leu326 sidechain helps in keeping the channel closed. The opening and closing of the channel is quite obvious from depicting this parameter.



(a) (b)

Figure 3.35 (a) The change in the hydrogen bond distance connecting backbone oxygen of Gly167 and OH group of Tyr172 during channel opening in simulation 1 (40 ns) of the dAcPFL sytem; (b) The change in the distance between two "gateway" residues, Gly167 and Leu326, where each residues is positioned at the top of two loops that are part of the longer sequences surrounding the defined channel during simulation 3 (55 ns) of dAcPFL model.

The results from the simulations based on dimeric models present a considerable advance compared to the simulations performed with monomeric test models. Two out of three independent simulations with dAcPFL recorded some kind of channel "opening" and the one where the hydrogen bond between Gly167 and Tyr172 is lost seems to be more stable. However, the conspicuous changes involving some larger parts of the protein were not noticed in any of the simulations, reflected by the comparable RMSD values of the protein residues for all the examined systems. These results indicate that only delicate changes and minor rearrangement

of the sidechains in the channel might results in enough space to accommodate CoA, illustrated by Figure 3.36. Another important outcome of the performed simulations is related to the observation that structural changes take place in the PFL system after the reaction with pyruvate. This implies that the successful completion of the first half-reaction might serve as a signal for the necessary structural changes that lead the second substrate into the active site. It might also be important to mention the ostensible influence of the inactive subunit, as none of the changes were spotted in the simulations based on monomeric models. The effect of the second monomer is subtle and it is not entirely clear how the behaviour of the active subunit is affected by its presence.

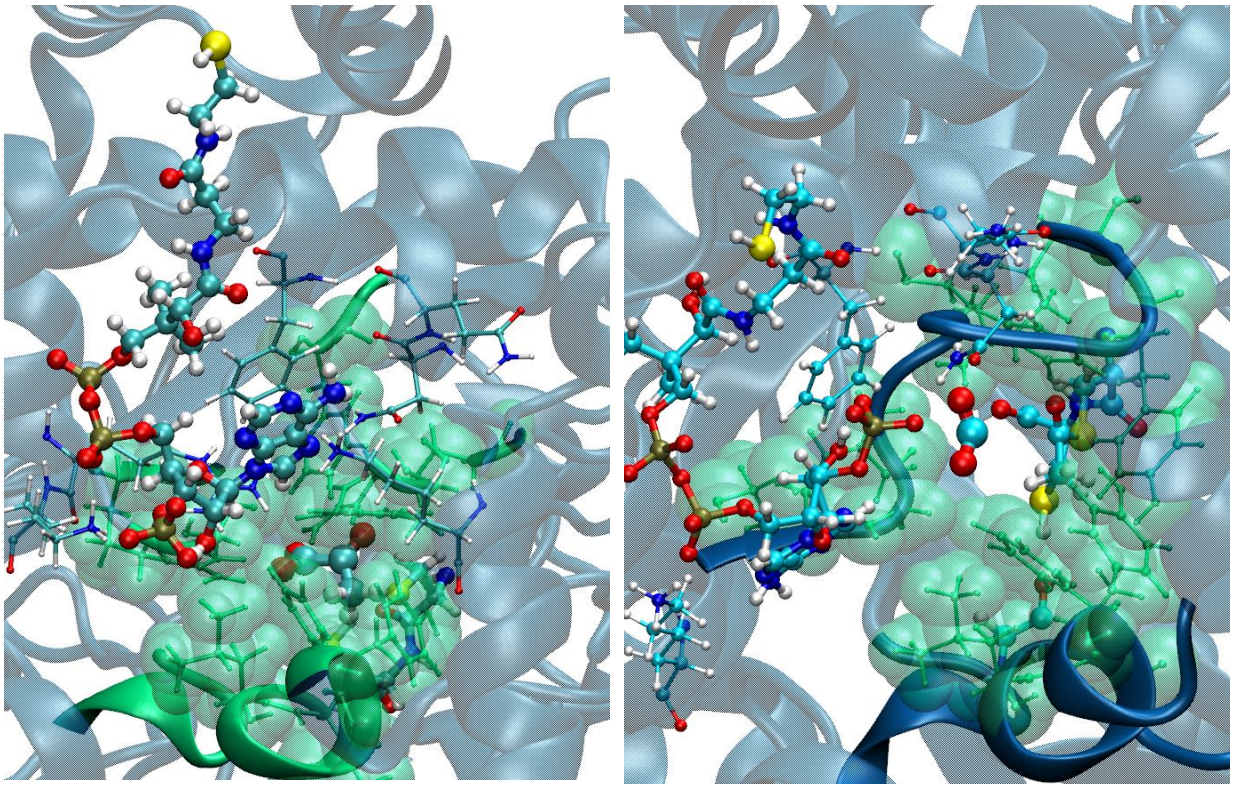


Figure 3.36 A view down the channel at the end of the free dynamics simulations of dPFL (left) and dAcPFL (right) systems. The dAcPFL system experienced some structural rearrangement of the sidechains in the channel leading to the channel "opening" (simulation 1), whereas dPFL system lingered to the hindered conformation found in the crystal structure.

Steered molecular dynamics. The free dynamics simulations were followed by a series of SMD simulations, in which different ensembles of the initial state were used to generate trajectories for pulling CoA inside of the protein. The first representative ensembles for dPFL and dAcPFL systems were generated by taking snapshots from free dynamics simulations after 30 ns that matched the targeted initial value of reaction coordinate (35 Å). The final value of reaction coordinate was 5 Å. The snapshots of dAcPFL system were sampled from the

conformations in which the entrance channel is "open" (simulation 1). Although some changes in the binding mode of CoA occurred during the simulation with dAcPFL described earlier in the text, the common feature of the CoA conformations adopted in both systems is the trapped thiol group between the residues of the inactive subunit. In the SMD simulations, that thiol group of CoA was pulled toward the active site cysteines during 5 ns by applying time-dependent potential (5 kcal/mol Å^2), resulting with a set of trajectories for each topology given in Figure 3.37.

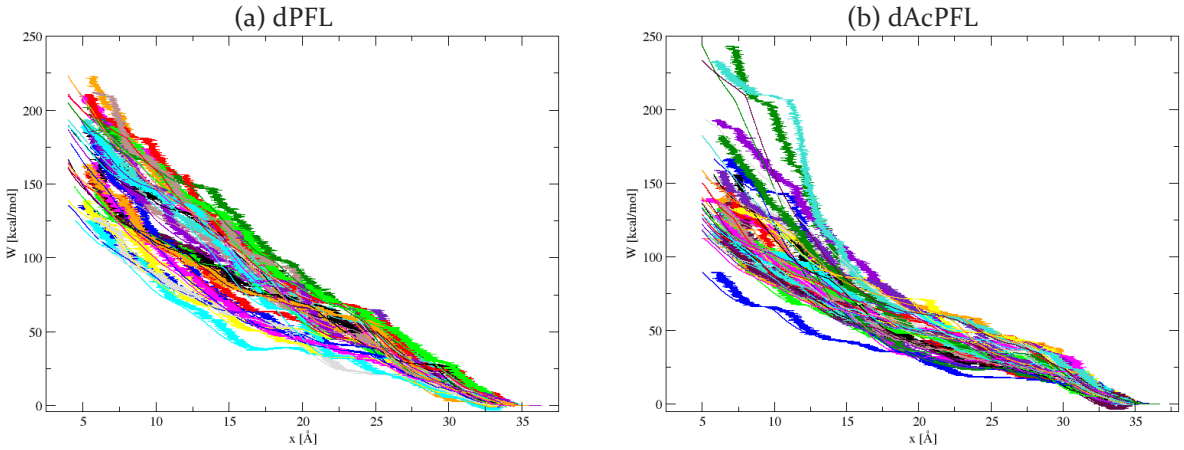
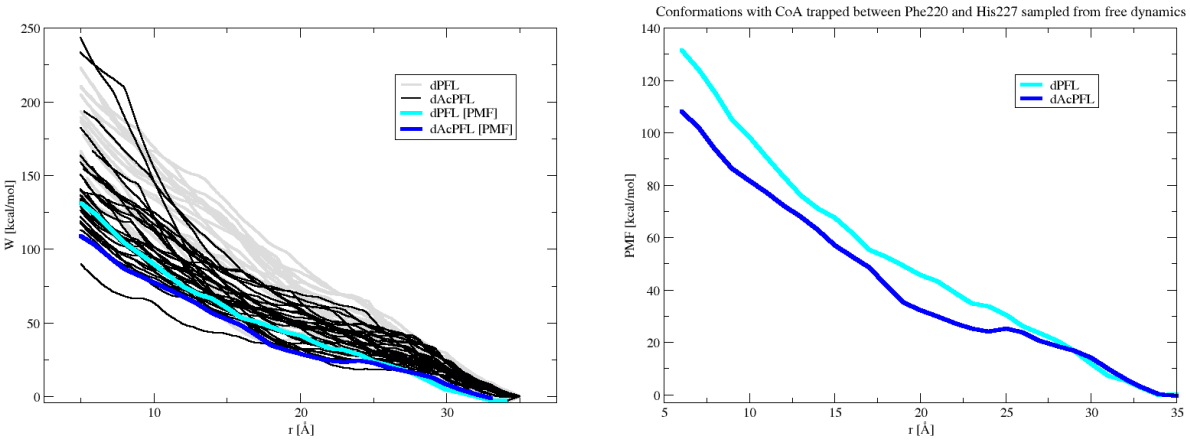


Figure 3.37 A set of 26 trajectories for each topology, dPFL (a) and dAcPFL (b), was obtained by running SMD simulations, in which CoA was pulled toward the active site cysteines during 5 ns. The smooth lines represent the amount of work as a function of the predefined values of the reaction coordinate (r) for each run, while the ragged lines depict work as a function of the actual values of the chosen variable (x) during the pulling experiments.

Again, the resulting free energy profiles computed from the simulated trajectories show that CoA takes a lower energy pathway to active site after the first half-reaction (Figure 3.38). The two curves practically overlap in the first part of the pulling experiment (35-25 Å), where the CoA tail adopts similar conformations in both systems with a trapped CoA thiol. After that point, the CoA enters the channel and two curves clearly diverge. This divergence is most likely a consequence of the channel opening in dAcPFL system, as the pulling runs were started in both systems from practically equal initial conformations of CoA, also confirmed by the overlap, and the first significant change occurs at the point in which CoA starts to interact with the channel. It was expected a stronger difference between two PMFs due to the presence of a more spacious channel in dAcPFL than in dPFL, but the absence of this effect might be due to the insufficient number of trajectories, resulting from insufficient sampling.



(a) (b)

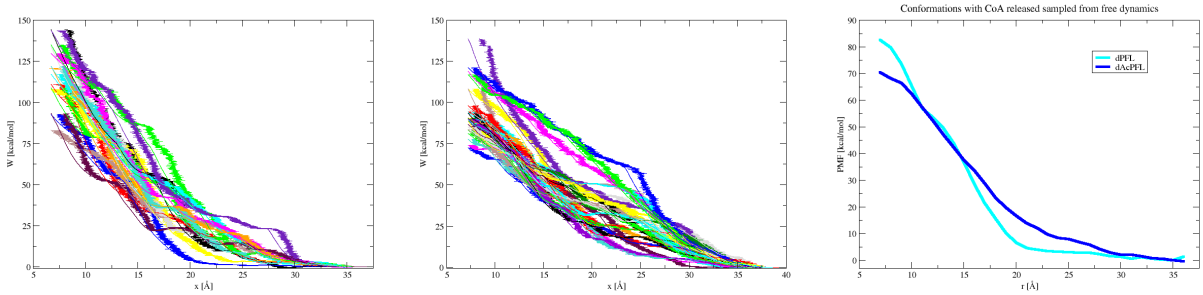
Figure 3.38 Trajectories obtained from the SMD simulations in which CoA is pulled toward the active site before and after pyruvate cleavage (a) overlapped by the estimated PMF. The final PMF was estimated by using Hummer and Szabo expression (b).

The insufficient sampling also leads toward high numerical values of the computed free energies, even higher than the numbers obtained from SMD simulations of monomeric system. This is partly due to the increase of the system size, but there is an additional problem reflecting on the total amount of work and pathways taken down the channel. This problem originates from the conformational deadlock of the CoA thiol group between two aromatic rings. First, an additional amount of work is required to overcome this interaction to release the thiol before proceeding down the channel. Second, as shown by now, there is a strong influence of the initial conformation on the resulting pulling pathway. The captured thiol group is positioned almost exactly over the adenosine moiety, which results with CoA curling into a bundle during the pulling experiment. The side-effect is significant amount of self-interaction before CoA is able to assume an extended form, which is necessary to reach the active site. Similar behaviour was observed in the experiments carried out with the monomeric mPFL, but it is more pronounced in the case of dimeric models. The bundling of CoA then often leads to the choice of less favourable pathways down the channel, which is reflected in the final free energy profiles. Namely, the SMD procedure drives the system along the chosen coordinate in a rather linear fashion, thus providing the shortest, but, not the optimal pathway.

To potentially overcome this limitation, a new set of pulling simulations was performed for both systems, but only to displace CoA thiol group away from its crystallographic position. Once the CoA tail was disentangled, the end point of each pulling simulations was taken as a starting point for free dynamics simulations. There were overall 5 releasing simulations and corresponding free dynamics in duration of 6 ns. From these free runs conformations were

sampled to create a new initial set for pulling CoA to the active site for each topology. The obtained SMD trajectories together with the final free energy profiles are given in Figure 3.39.

From the presented trajectories is visible that the release of CoA had some effect in lowering the overall trajectories in both systems, implying that significant amount of work was spent on overcoming the interactions between CoA and the inactive subunit, but also on the formation of the self-interacting bundle. The latter was not entirely avoided in the case of dAcPFL, while dPFL system was more successful in that sense. Namely, in this setup, dPFL system takes a lower pathway in the beginning of the pulling on average, which is a direct consequence of the starting conformations characterized by freely moving tail of CoA. On the other hand, in dAcPFL system interactions between CoA tail and proximal protein residues are present. This is another illustration of the influence of the initial ensemble of conformations on the final SMD trajectories, but it is an even better example of the influence of the channel opening on the resulting free energy profiles. The computed PMF curves show that once CoA enters the channel, the energy steeply rises for dPFL system where the channel is closed. The slope of the dAcPFL curve is less steep and results with a lower overall free energy change for the given process. This finding supports the statement that breaking the bond in pyruvate facilitates the approach of CoA.

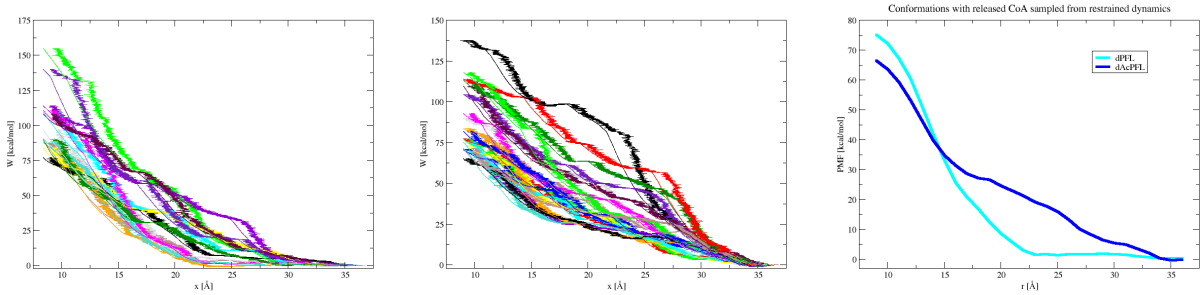


(a) dPFL (b) dAcPFL (c)

Figure 3.39 A set of 16 trajectories for each topology, dPFL (a) and dAcPFL (b), was obtained by running SMD simulations, in which CoA was pulled toward the active site cysteines during 5 ns. The smooth lines represent the amount of work as a function of the predefined values of the reaction coordinate (r) for each run, while the ragged lines depict work as a function of the actual values of the chosen variable (x) during the pulling experiments. The final PMF was computed by using Hummer and Szabo's estimator (c).

A repeated experiment with SMD simulations was performed from the initial conformations generated in the restrained simulations, where CoA tail was kept at a fixed distance from the catalytic cysteines (36 Å). A starting conformation for restrained dynamics was a snapshot containing non-interacting CoA tail for each topology. Although these simulations started from similar points, the imposed restraint still allowed CoA to adopt various conformations

including those that interact with the protein residues. That was the case in dAcPFL, while dPFL system kept a non-interacting form, which resulted with different ensembles for each system. These new ensembles were used to initiate subsequent SMD simulations and the calculated trajectories are given in Figure 3.40. The estimated PMFs closely resemble to those obtained from the previous set of calculations started from the snapshots generated during the free dynamics. This result provides an additional support to the claim that the average value of total work required to drive CoA from protein surface to the interior is lower for dAcPFL than for dPFL.



(a) dPFL (b) dAcPFL (c)

Figure 3.40 A set of 20 trajectories for each topology, dPFL (a) and dAcPFL (b), was obtained by running SMD simulations, in which CoA was pulled toward the active site cysteines during 5 ns. The smooth lines represent the amount of work as a function of the predefined values of the reaction coordinate (r) for each run, while the ragged lines depict work as a function of the actual values of the chosen variable (x) during the pulling experiments. The final PMF was computed by using Hummer and Szabo's estimator (c).

3.4.2.2.2 PFLX Model Set (without Mg^{2+})

Free dynamics. Based on the observation that the thiol group of the CoA was prone to being trapped between aromatic sidechains belonging to the opposing monomer, it was decided to investigate if the presence or absence of the Mg^{2+} ion might affect the flexibility of the coenzyme and hence the work required to bring it in contact with the substrate. This resulted in the simulations of the dPFLX and dAcPFLX.

The removal of this divalent cation indeed resulted with more flexibility and different behaviour of the CoA compared to all the previous simulations, but again only in the case of CoA bound to the active subunit of dAcPFLX system. Namely, a change in the binding mode occurs after approximately 15-20 ns in a 60 ns simulation, where the 3' phosphate starts to interact with the highly conserved Arg160. This newly formed salt bridge starts to pull the entire nucleotide part of CoA away from its initial binding site leading toward the loss of the

stacking interaction between adenine and Phe149. This change, in turn, allows formation of an additional salt bridge between 3' phosphate and neighbouring Lys115 and an indirect interaction with Glu324 via bridging sodium ion. All these changes in binding interactions result with the release of the thiol moiety from the aromatic trap made by the second subunit. After that point, CoA is free to explore the conformational space more thoroughly, occasionally forming interactions with another lysine residue – Lys617.

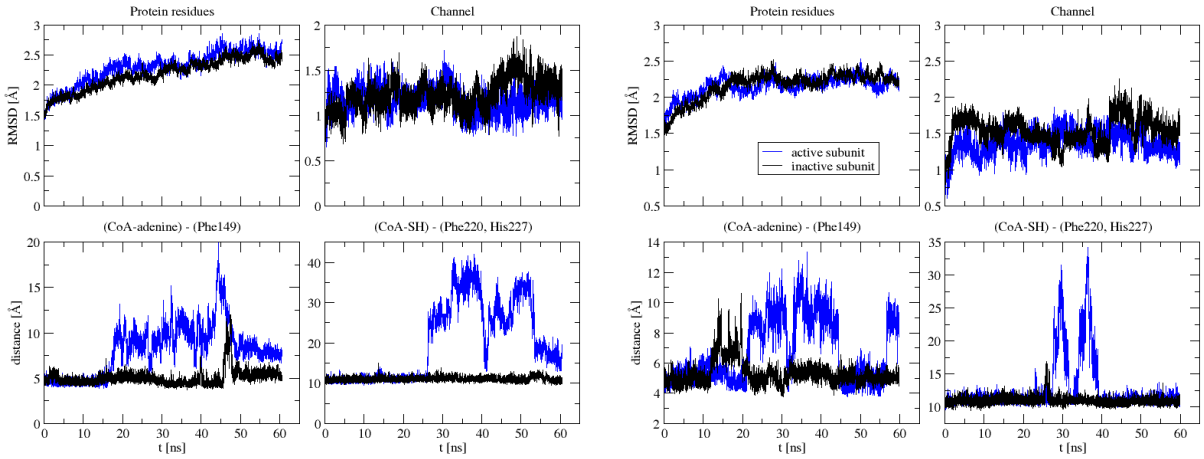


Figure 3.41 Data collected during two independent simulations (60 ns) carried out with the dAcPFLX system for the following variables in the active and inactive subunit: RMSD of the protein residues; RMSD of the channel residues; distance between adenine moiety of CoA from Phe149; distance between thiol group of CoA and two aromatic residues of the inactive subunit, Phe220 and His227.

The disentanglement of CoA was reproduced in a repeated simulation (60 ns), but after 20 ns the thiol group returned back to the initial position between two Phe220 and His227 of the inactive subunit. In that context, it is important to mention that CoA belonging to the inactive subunit remains bound in a mode similar to that in the beginning of simulation, although it fluctuates more strongly than in presence of Mg^{2+}. This CoA is also capable of making a salt bridge with Arg160, but the adenine base remains close to the Phe149 maintaining the stacking interaction. Because of that and the general increase in mobility of CoA, the RMS deviations of the CoA and its binding site proved to be not as informative as they were in the previous presentations of the system dynamics. Instead, these two variables were replaced by potentially better descriptors – the distance between adenine and Phe149, and a distance between thiol group of CoA and residues Phe220 and His227 of the opposing monomer (Figure 3.41). By looking at the distances over the simulation time, it is visible how adenine drifts away from the Phe149 as the binding mode changes and how the CoA tail managed to escape the resting

position between Phe220 and His227. At the same time, the CoA molecule bound to the inactive monomer preserves the stacking interaction with Phe149 and the thiol group remains between the two aromatic residues as found in the crystal structure. Interestingly, the channel opening events that were observed in the dAcPFL system did not occur in the simulations of the dAcPFLX system depicted in Figure 3.41.

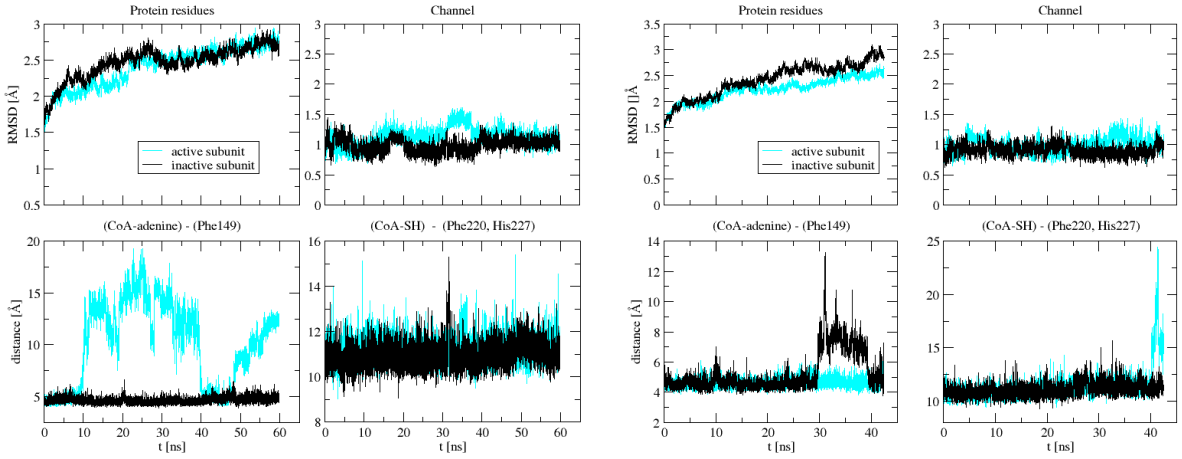


Figure 3.42 Data collected during two independent simulations (60 ns and 45 ns) carried out with the dPFLX system for the following variables in the active and inactive subunit: RMSD of the protein residues; RMSD of the channel residues; distance between adenine moiety of CoA from Phe149; distance between thiol group of CoA and two aromatic residues of the inactive subunit, Phe220 and His227.

The observations made for CoA bound to the inactive subunit of dAcPFLX are very similar to those made for the simulations carried out with dPFLX model, in which both CoA molecules remain close to their initial binding site with certain deviation. For this model two independent simulations providing rather consistent results were also performed, as shown in Figure 3.43 and Figure 3.42. As for dAcPFLX, this system also lacks any notable modifications in the positioning of the channel residues, but CoA exhibits increased flexibility as expected. This is reflected in the loss of the stacking interaction between adenine and Phe149 in the first simulation, while at the end of the second simulation a release of the CoA tail occurred in this system as well. The similar pattern was followed as described in dAcPFLX case, where interaction of 3'-phosphate and Arg160 serves as a trigger for further changes. In general, though, the binding of CoA on the active subunit of dPFLX system is considerably more stable than seen in dAcPLFX system.

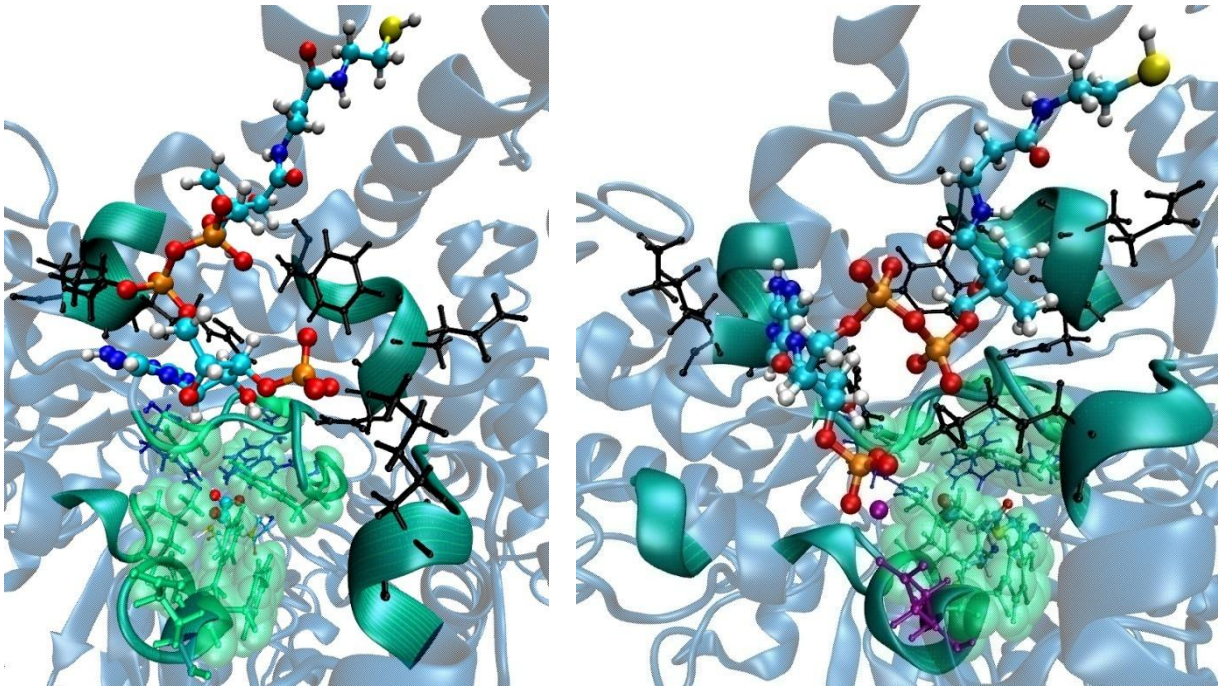
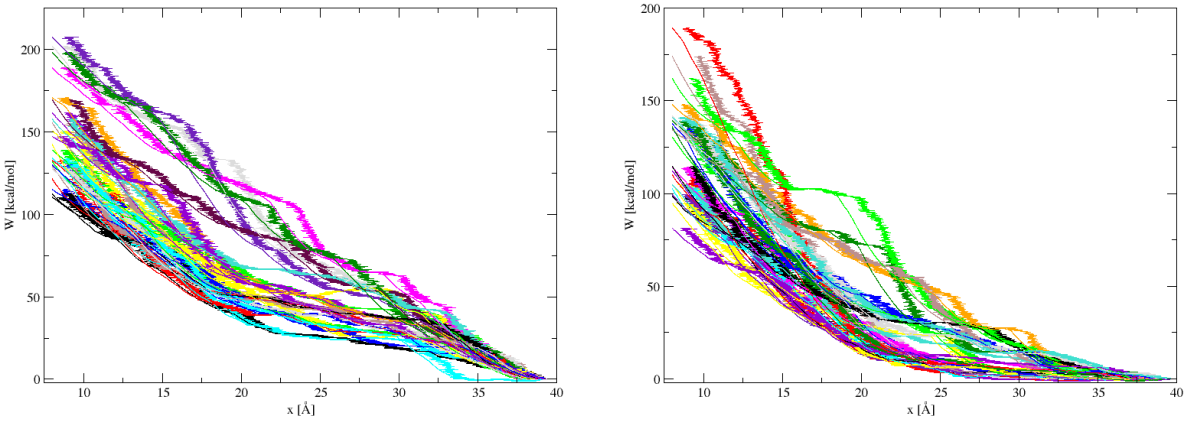


Figure 3.43 The final snapshot of the simulation 1 performed with dPFLX (left) and dAcPFLX (right) models. In both systems the adenine base is not stacked to Phe149 anymore; it rather makes a new π-π interaction to Arg160. The 3' phosphate is strongly bound to Lys118 in both systems, while in dAcPFLX it makes additional indirect bond to Glu324 via bridging sodium ion (purple).

The free dynamics simulations performed with dPFLX and dAcPFLX system support the assumption that the removal of Mg^{2+} would affect the flexibility of CoA. The increased mobility enables it to escape the trapping interaction with the residues of the opposing monomer. However, the channel opening observed in the simulations carried out with dAcPFL system was not reproduced in the absence of the Mg^{2+}. Despite this outcome, the configurations were sampled for the subsequent SMD simulations for both topologies. It is indicative, though, that the acetylated system is again more susceptible to structural changes than the dPFLX model representing the system before the reaction with pyruvate.

Steered molecular dynamics. Configurations corresponding to the initial state were sampled from free dynamics simulations performed with dPFLX and dAcPFLX in the interval between 30-50 ns. The snapshots with a suitable value of the reaction coordinate (39 Å) were selected from that interval and used in the SMD simulations in duration of 4 ns. The snapshots taken from dAcPFLX simulations correspond mostly to the conformations in which the CoA tail does not interact with the inactive subunit, in contrast to the dPFLX system, where a binding mode of CoA similar to that in the crystal structure was preserved.



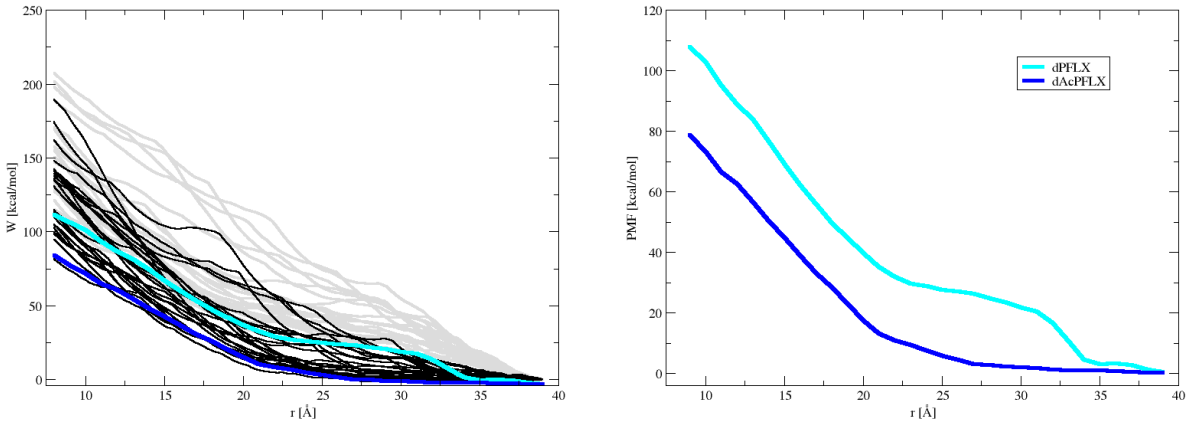
(a) dPFLX (b) dAcPFLX

Figure 3.44 A set of 25 trajectories for each topology, dPFLX (a) and dAcPFLX (b), was obtained by running SMD simulations, in which CoA was pulled toward the active site cysteines during 5 ns. The smooth lines represent the amount of work as a function of the predefined values of the reaction coordinate (r) for each run, while the ragged lines depict work as a function of the actual values of the chosen variable (x) during the pulling experiments.

Trajectories obtained for both systems are shown in Figure 3.44, where the evident effect of the CoA conformation on the shape of the trajectory is again displayed. Again, there is an obvious difference in the beginning part of the pulling simulations originating from the different CoA conformations in two systems. In dAcPFLX system, where CoA tail is free, the initial part of the trajectories requires a lesser amount of work to drive the process compared to the dPFLX system. The curves obtained for dAcPFLX model resemble to those calculated with dPFL model, which was also represented by an ensemble of conformations with non-interacting CoA tail. On the other hand, dPFLX trajectories resemble more to the dAcPFL counterparts, because in both cases a certain amount of energy was needed to overcome the interactions of CoA tail with the protein residues, resulting with the higher energy pathways for the entering portion of trajectories. However, it seems that it gets almost equally difficult to push CoA tail closer to the catalytic cysteines once it approaches the “gate”. This becomes more evident in Figure 3.45b presenting the estimated potential of mean force for the given process for both topologies, where two curves run almost in parallel from the gate point further on. Here the conformation of CoA does not play such an important role anymore, but rather the protein sidechains hindering the active site.

Nevertheless, the final free energy profiles are consistent with the previous findings that breaking the C-C bond in pyruvate facilitates the entrance of CoA in some way. In the model containing Mg^{2+} ion (dAcPFL), the approach of CoA was mitigated by moving the sidechains of the residues forming the “channel” away from their initial hindering position. The same

behaviour was not repeated by the dAcPFLX system; instead the help was provided in a form of disentanglement of CoA tail from the interactions with the opposing subunit.



(a) (b)

Figure 3.45 Trajectories obtained from the SMD simulations in which CoA is pulled toward the active site before and after pyruvate cleavage (a) overlapped by the estimated PMF. The final PMF was estimated by using Hummer and Szabo expression (b).

Either way, the performed free dynamics simulations resulted with some structural changes in dAcPFL and dAcPFLX systems, which were not observed in dPFL and dPFLX systems. This indicates that the chemical changes taking place in the active site during the first half-reaction play an important role in the regulation of the approach of the second substrate closer to the active site. The free energy calculations further corroborate this finding, in which AcPFL system in all model types consistently provides a lower energy pathway for moving CoA thiol group from protein surface all the way to the catalytic cysteines.

3.4.3 CONCLUSION

A series of extensive molecular dynamics simulations was carried out to investigate possible entrance pathways of CoA to reach the buried active site of PFL from its binding spot at the protein surface. The CoA is the second substrate of PFL and its role is to pick up the acetyl group located temporarily at the Cys418 after the cleavage of the first substrate, pyruvate. Acetyl-CoA and formate are the final products of PFL catalysis. However, the pathway that might bring CoA from the surface into the active site is not obvious from the crystal structures and it is reasonable to assume that the protein should undergo certain structural modifications to accommodate CoA.

The models representing the PFL system before and after the first half-reaction with pyruvate were used to examine the possible effect that acetylation of the enzyme has on the necessary conformational changes. In addition to the NVT free dynamics simulations, a series of free energy calculations was carried out to calculate potential of mean force describing the approach of CoA to the active site before and after the first half-reaction. The PFL protein comes in a homodimeric form and two sets of models were derived from the available crystal structure; one set of models was built using a single subunit (mPFL, mAcPFL), while the other contains the full dimer. The latter set of models was divided into two additional subsets, where one subset contains a putative Mg^{2+} ion (dPFL, dAcPFL), which is removed from the second subset (dPFLX, dAcPFLX) and replaced by two additional sodium ions.

In summary, most of the observed structural changes occurred in the acetylated enzyme (AcPFL), while the enzyme in complex with pyruvate undergoes only minor changes compared to the initial crystal structure. These findings support the hypothesis that the chemical changes in the enzyme probably serve as a trigger for necessary conformational changes and in that way control the approach of the second substrate. Further evidence comes from the free energy calculations, which consistently show that driving CoA from the protein surface down the putative channel is facilitated after the pyruvate cleavage.

However, the resulting free energy profiles may only be interpreted from the qualitative aspect, because the numbers are unreasonably high. The numbers obtained from the umbrella sampling method are significantly lower than those coming from non-equilibrium SMD approach, which is mostly due to the better statistics collected in the umbrella sampling experiments. A rather low number of trajectories generated with SMD was used to estimate the

PMF, which corresponds to insufficient sampling. A larger number of trajectories should certainly help to improve the results and also to avoid the observed problem of strong influence of the initial ensemble on the final trajectories. The latter problem is just a part of the limitations of SMD method in its execution, because the choice of the reaction coordinate and the switching rate also play a very important role in production of a representative set of trajectories for the investigated process. The slower pulling rate would enable the system to explore the phase space around the chosen coordinate more thoroughly by granting it more time to adapt to its new position and relax. That would also help to avoid the observed self-interaction within CoA. The SMD simulations should be significantly extended to achieve some reasonable values of the resulting free energy.

In the context of the proper sampling, it would be important to point out that longer free dynamics simulations may be required to capture the necessary changes. The observed changes with AcPFL system may only be the onset of more serious structural changes that come with further chemical events in the active site, such as radical transfer between formate and Cys419, or further to Gly734. The turnover number of PFL is approximately 1000 s^{-1}, meaning that the single catalytic cycle is taking place on the microsecond scale. Therefore, the simulations performed in this research are probably too short to record the key events that eventually lead CoA toward the active site. In any case, a certain progress in that direction has been made and the coupling of the chemical changes in the active site with the structural rearrangements of the protein is largely substantiated by the simulation results.

In the simulations based on the monomeric models, mPFL and mAcPFL, the expected structural rearrangement or any kind of appreciable conformational departure from the initial crystal structure was not observed. The major change was manifested as a partial CoA unbinding in the mPFL system after ~30 ns. However, the free energy calculations demonstrated that the difference between two systems exists when it comes to driving the CoA down the possible channel to reach the active site. This possible channel represents the shortest and the least crowded pathway when the system is driven along the chosen reaction coordinate, which corresponds to the distance between the cysteamine block of CoA and two catalytic cysteines, Cys418 and Cys419. The protein residues forming a sort of a gate inside this channel that protect the active site include Gly167, Tyr172, Tyr323, Leu326 and Phe327. An important role is also played by the binding residues in the active site – Arg176, Phe432 and Arg435. Although the channel is closed in both systems during the free dynamics, the free energy profiles resulting from SMD and umbrella sampling simulations indicate that the CoA

follows a lower energy pathway in mAcPFL system. This implies that scission of the C-C bond in pyruvate facilitates the entrance of CoA to the active site in some extent.

In contrast, the simulations with the dimeric models have introduced more dynamics in the system. Namely, during the free dynamics with dAcPFL system a channel opening in two of three independent simulations was observed. Since this type of structural rearrangement was not captured with the monomeric models and the possible explanation might be that the presence of the inactive monomer has an influence on the dynamics of the active subunit. If present, this influence is rather subtle and quite difficult to quantify, but the effect of the second subunit on CoA and its conformational space is evident. The thiol group of CoA is placed between the sidechains of Phe220 and His227 of the opposing monomer and it basically does not spontaneously disengage from this interaction on the given time scale. The unfavorable initial conformations of CoA affect the resulting trajectories obtained in SMD simulations for both systems, as they usually involve a lot of self-interaction during the pulling. Nevertheless, the free energy profiles extracted from SMD simulations corroborated the previous findings with monomeric models that the acetylated enzyme provides a more favourable environment for the accommodation of CoA inside the protein.

An additional set of calculations using dimeric model, but without the putative Mg^{2+} ion, was carried out to examine the possible effect of this ion on the conformational behaviour of CoA. Indeed, the changes in the binding mode of CoA take place in these simulations which were not observed in the previous runs. The changes in the binding mode lead to the loss of the interaction between CoA thiol and the opposing monomer, but interestingly, this again occurs only in dAcPFLX system. The channel opening was not reproduced in this set of simulations for any of the used topologies. The computed potentials of mean force from SMD simulations again resulted with the lower free energy profile for dAcPFLX compared to dPFLX system, but this lowering in energy can mainly be attributed to the conformational difference of CoA between two systems.

All the presented results support the claim that acetylation of the enzyme serves as a signal that is time for CoA to approach the active site and to finish the catalytic cycle. The importance of this signalling might be closely related to the efforts of the enzyme to preserve its catalytic activity. Namely, the premature presence of CoA might meddle with the first half-reaction by quenching the radical from the catalytic cysteines. It was shown in an experimental study that the small thiol molecules reversibly deactivate PFL, where the larger thiol molecules and non-thiolic compounds are not able to promote inactivation. According to the study, the

smallest thiols are the most efficient inactivators, implying that these molecules have to be able to access the active site to successfully inactivate PFL. As the size of thiol molecules increases, the approach to the active site becomes much harder task, and this includes the second substrate – CoA. By keeping the channel closed during the first half-reaction, the enzyme decreases the risk of futile side reactions with wandering thiols or any other potential quenchers. The acetylation of the enzyme could then be interpreted as a signal that the first half-reaction was successful and that the active site is ready for the entrance of the second substrate. In this context, the observed channel opening in the dimeric dAcPFL system might be on the right track for the solution of the investigated problem. It could be that only minor changes are sufficient to accommodate CoA by creating a rather narrow channel that also shows a certain level of selectivity, preventing other larger thiol molecules to enter the active site and interfere with the catalysis. Namely, there is a large number of positively charged residues surrounding the entrance in the putative channel (Lys118, Lys159, Arg160, Lys615, Lys617) that are able to interact with negatively charged phosphates of CoA and serve as an anchor point for CoA. On the other hand, it has been observed in the performed simulations that the pantothenate and cysteamine blocks of CoA are able to make various hydrogen bonds with the protein residues forming the putative channel. This would allow stable binding of CoA with its thiol group in the active site and a tight fit of the second substrate and the protein at the same time, preventing the access of the non-native substrates to the active site and thus improving the enzyme efficiency. This would be an amazing example of coupling between the chemical and structural changes in an enzyme that acts upon two substrates in consecutive fashion by allowing one substrate in the active site at the time, while preserving the radical necessary for the catalytic activity. The results presented in this thesis are definitely a harbinger of this kind of cooperativity to optimize the system efficiency.

3.4.4 REFERENCES

1 Plaga, T. W.; Frank, K.; Knappe. J. *Eur. J. Biochem*. **1988**, 178, 445.

2 Knappe, J.; Elbert, S.; Frey, M.; Wagner, A. F. V. *Biochem. Soc. Trans*. **1993**, 21, 731.

3 Himo, F.; Eriksson, L. A. *J. Am. Chem. Soc*. **1998**, 120, 11449.

4 Guo, J.-D.; Himo, F. *J. Phys. Chem. B* **2004**, 108, 15347.

5 Auffinger, P.; Grover, N.; Westhof, E. *Met. Ions Life Sci*. **2011**, 9, 1.

6 Becker, A.; Kabsch, W. *J. Biol. Chem*. **2002**, 277 , 40036.

7 Word, M. J.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol*. **1999**,285, 1735.

8 Case, D. A.; Darden, T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. AMBER 11, **2010**, University of California, San Francisco.

9 Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, 14, 33.

10 Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang,W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J. W.; Wang, J.; Kollman, P. A. *J. Comput. Chem*. **2003**, 23, 1999.

11 Case, D. A.; Darden, T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. AMBER 11, 2010, University of California, San Francisco.

12 Case, D. A.; Darden, T. A.; Cheatham, III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. AMBER 12, 2012, University of California, San Francisco.

13 Bayly, C.I; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem*. **1993**, 97, 10269.

14 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian 03, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004

15 Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem*. **2004**, 25, 1157.

16 Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys*. **1984**, 81, 3684.

17 Uberuaga, B.P.; Anghel, M.; Voter, A.F. *J. Chem. Phys*. **2004**, 120, 6363.

18 Goetz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. *J. Chem. Theory Comput*. **2012**, 8 (5), 1542.

19 Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. USA* **2001**, 98, 3658.

20 Minh, D.; Adib, A. *Phy. Rev. Lett*. **2008**, 100, 180602.

21 https://simtk.org/home/ferbe

22 Kumar, S.; Bouzida, D.; Swendsen, R.H.; Kollman, P.A.; Rosenberg, J.M. *J. Comput. Chem*. **1992**, 13, 1011.

23 http://membrane.urmc.rochester.edu/

24 Kästner, J.; Thiel, W. *J. Chem. Phys*. **2005**, 123, 144104.

25 Shirts, M. R.; Chodera, J. D. *J. Chem. Phys*. **2008**, 129, 124105.

26 https://simtk.org/home/pymbar

27 http://metals.zesoi.fer.hr/metals/

28 Tus, A.; Rakipović, A.; Peretin, G.; Tomić, S.; Šikić, M. *Nuc. Ac. Res*. **2012**, 1-6, doi:10.1093/nar/gks514

4. (6-4) PHOTOLYASE

4.1 INTRODUCTION

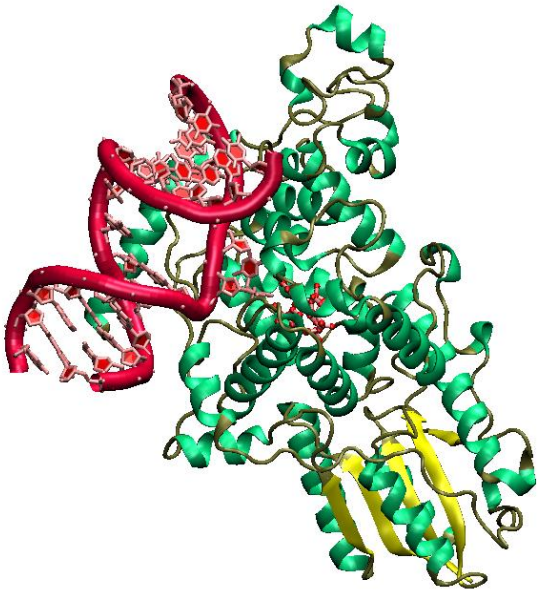
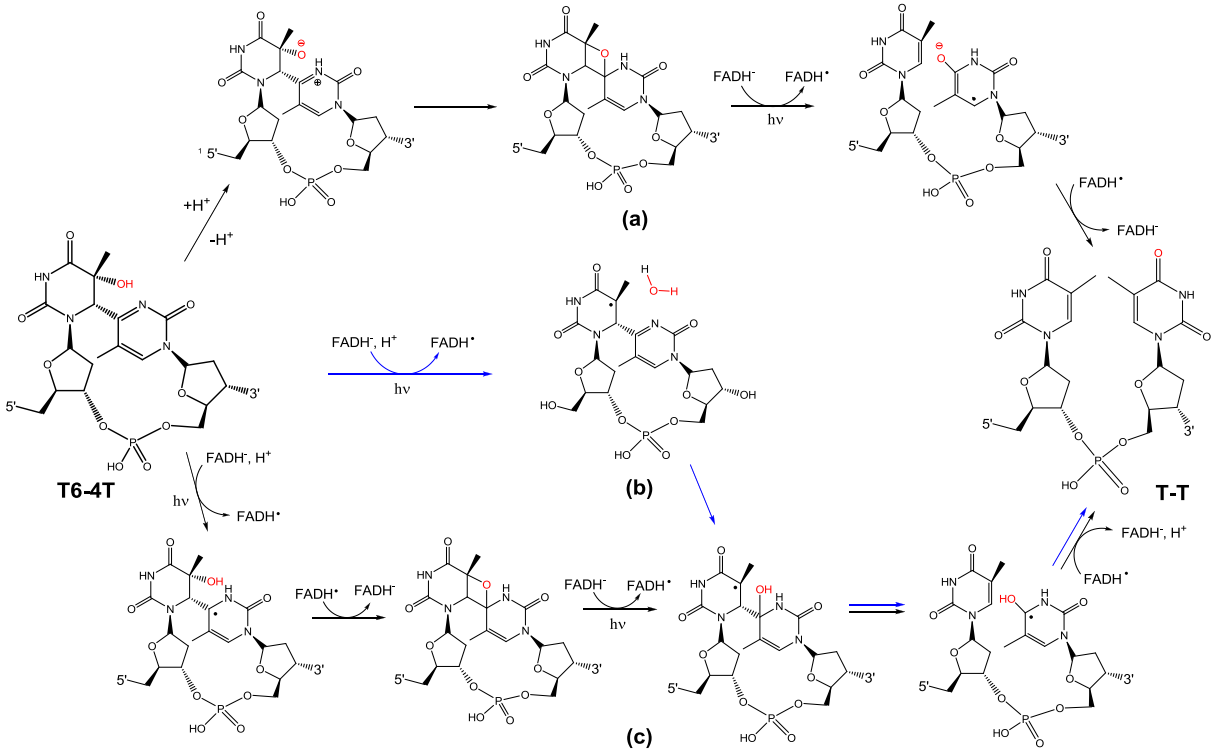


Figure 4.1 Crystal structure of (6-4) photolyase in complex with DNA strands and lesion flipped in the active site.

Ultraviolet radiation can cause harmful damage to DNA. For example, the exposure of two adjacent pyrimidine bases, such as the thymine-thymine (T-T) pair shown in Scheme 4-1, to UV light (in the 200-300 nm range) may result in dimerization. The most common forms of the corresponding lesions are the cyclobutane pyrimidine dimers (CPD) and pyrimidine(6-4)pyrimidone photoproducts (see, e.g., T6-4T in Scheme 4-1).[1] Because of their high mutagenic and carcinogenic potential, the repair of these lesions is of utmost importance for cell survival. The group of light-dependent enzymes capable of reforming the initial monomers by photo reactivation is known collectively as the DNA photolyases. The formation of the CPD lesion and its repair by CPD photolyase has been well studied over the years, both experimentally[2] and computationally.[3,4] Much less is known, however, about the mechanism of repair of (6-4) photoproducts by (6-4) photolyase. While both types of enzymes show certain sequence similarities and require reduced $FADH^-$ in their active site for catalysis,[5] their repair mechanisms are believed to differ.[6]

It has been widely accepted that the formation of the (6-4) photoproduct is a result of a Paterno-Büchi cycloaddition reaction of the C5=C6 (5') and C4=O/NH (3') double bonds of the two neighbouring pyrimidines (for T-T or T-C), which proceeds through an oxetane (for T6-4T) or an azetidine (for T6-4C) intermediate.[7,8] The initial proposal of the repair mechanism of the (6-4) photolyase included the thermal formation of this same cyclic intermediate as the first step,[9] catalyzed by two highly conserved histidines in the active site.[10] One histidine was proposed to act as a base, deprotonating the migrating functional group for the nucleophilic

attack on C4', while the other should act as an acid, donating the proton to the acylimine (Scheme 4-1a).[11] Subsequent electron donation from the excited FAD cofactor should enable the cycloreversion of the radical anion of the oxetane/azetidine,[12] a process which appears feasible on the basis of some experimental[13] and computational[14] studies. The final step involves the return of the electron to FAD. The energy gap between the T6-4T lesion and its oxetane isomer, predicted by theoretical calculations to be between 14.5-16.6 kcal mol^{-1},[15] has usually been identified as a major disadvantage of this mechanism. Similarly, the oxetane intermediate has not thus far been experimentally observed.



Scheme 4-1 (6-4) photolyase repair mechanisms proposed by (a) Hitomi *et al.*;[11] (b) Maul *et al.* (blue arrows);[16] (c) Sadeghian *et al.*[19] Protons are donated or accepted by His365 and His369, respectively, but the residues are omitted for clarity. For the same reason the water molecule participating in mechanism (c) is not shown.

The determination of the crystal structure of (6-4) photolyase from *Drosophila melanogaster*, complexed with the T6-4T lesion, was a major step forward in the understanding of the repair.[16] Based on structural and biochemical data, a modified mechanistic proposal lacking the oxetane intermediate has emerged. Oxetane formation requires the protonation of the acylimine, a step that was reportedly not supported by the positions of the histidine residues. Instead, it was suggested that one of the histidines (His365) should protonate the migrating functional group, rather than deprotonating it, thus making it

a better leaving group after the electron injection from FAD (Scheme 4-1b). The resulting water molecule could attack the acylimine to form a radical intermediate, which would then rapidly fragment into the initial pyrimidine bases. In this scenario, the loss of a proton along with electron transfer back to FAD would complete the catalytic cycle.

A subsequent computational study addressed the repair mechanism of the (6-4) photoproduct from an energetic point of view.[17] Using a model system that omitted the enzyme, the study investigated the possible fates of the radical anion of the T6-4T dimer, which is formed after electron capture from the excited FAD cofactor. It was found that pathways on the electronic ground state of the radical anion, either with or without an oxetane intermediate, were associated with high activation energies. The non-oxetane pathway, which involves a direct hydroxide transfer (analogous to the Scheme 4-1b without the addition or removal of a proton), was found to have a low-lying excited state whose relaxation could lead to lesion repair. In order to make productive use of this state, however, the reaction would require either a re-excitation (an additional photon) or, as the authors seemingly prefer, a directed non-adiabatic relaxation of the hot radical-anion state towards the product configuration. This study also highlighted the potential importance of the (non)formation of the intramolecular O5H-N3' hydrogen bond. According to the spectroscopic analysis of the pyrimidine(6-4)pyrimidone photoproduct, the formation of this intramolecular hydrogen bond is responsible for the remarkably low pK_a value for the N3' atom.[18]

During the preparation of the present article, an additional computational investigation appeared in the literature.[19] In agreement with the previous results,[17] the most recent study failed to identify a low-energy rearrangement mechanism on the ground state surface of the radical anion. Rather, through the extensive use of QM/MM calculations that explicitly included the protein environment, the authors advocate what is essentially a two-photon process. That is, the electron transfer from the FAD to the lesion is proposed to lead to the formation of an oxetane intermediate, which transfers its excess electron back to the cofactor (Scheme 4-1c). The oxetane formation is catalyzed by a protonated His365 and the additional stabilization is gained from a hydrogen bond formed between a water molecule and the hydroxyl group of the lesion. In the second step, another electron transfer from FAD, which presumably needs to be re-excited, is required to split the oxetane intermediate into the original monomers.

Despite the fact that the existing theoretical studies[17,19] seem to prefer a mechanism involving the excited state of the lesion or a re-excitation of either the lesion or the cofactor,

two independent investigations have recently provided evidence that the repair reaction takes place entirely on the ground state of the reduced lesion. In an experimental study,[20] the repair photocycle of (6-4) photolyase was investigated by means of ultra fast spectroscopy, applied to the wild-type enzyme as well as inactive mutants. The authors report that, according to their results, a proton transfer to the anionic lesion is a key step in the repair pathway. They find that, after this step takes place, the ordinarily rapid back electron transfer (50 ps) from the lesion to the cofactor is completely blocked and the subsequent ground-state repair occurs with 100% efficiency. Indeed, this back electron transfer has been suggested to be the main reason for the low repair efficiency of (6-4) photolyase. In a complementary study,[21] TD-DFT calculations have demonstrated that the initially absorbed photon does not carry sufficient energy to initiate the electron transfer and to simultaneously excite the radical anion of T6-4T. Similarly, using arguments based on the photon flux density of solar radiation at the Earth's surface in the range where $FADH^-$ and the lesion absorb, the authors argue against the involvement of a second photon in the repair mechanism. These arguments are likely to be more relevant to a mechanism of the type proposed in reference 17, where the intermediate is a short-lived radical anion, than a mechanism like that presented in reference 19, where the intermediate is a relatively stable closed-shell system. Namely, the latter species could have a life-time that is sufficiently long for it to absorb a second, lower-energy photon and thus constitute a viable reaction intermediate.

Despite the success of previous studies, further work is required to reach a consensus as to the repair mechanism of (6-4) photolyase. One of the issues with the computational studies[17,19] may involve inadequate treatment of the protein environment. That is, even though the most recent study was performed with state of the art QM/MM techniques,[19] it is not entirely clear that the protonation states of the key active-site residues were assigned correctly. Given that all of the proposed mechanisms involve acid-base chemistry of one form or another and that the active site contains two conserved histidine residues, this is clearly an important issue. Indeed, one may argue that its resolution is a prerequisite for further mechanistic studies, especially those employing computational methods.

It is commonly thought that, in the (6-4) photolyase from *D. melanogaster* (see Figure 4.2, prepared with VMD[22]), His365 is protonated while His369 is neutral.[16,20] This assumption is usually based upon results from an EPR/ENDOR study performed on the (6-4) photolyase from *Xenopus laevis*, which exploited the formation of the $FADH^{\bullet}$ radical.[23] Based on changes in the principal components and the intensities of the hyperfine couplings of selected FAD protons (H8 and H1'), induced by changing pH values and introducing point mutations, the

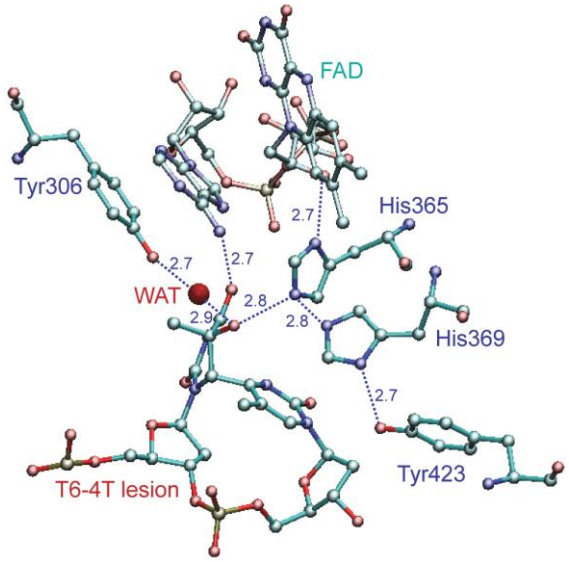


Figure 4.2 Hydrogen bond network connecting the T6-4T lesion, catalytic residues (His365, His369, Tyr423) and the FAD cofactor in the 3CVU crystal structure[16] of the active site of the (6-4) photolyase. Also shown are relevant interatomic distances in Ångstroms and the positions of Tyr306 and the oxygen atom of a structural water molecule.

authors argued that His354 (corresponding to His365 in *D. melanogaster*) is protonated, while His358 (corresponding to His369 in *D. melanogaster*) is neutral at pH 9.5. It is important to mention that (6-4) photolyase repair activity exhibits strong pH dependence, as shown by the experimental study of Hitomi *et al.*[11] According to their measurements, *X. laevis* (6-4) photolyase reaches its maximum activity at pH 8.5, which decreases significantly upon approaching pH 6. The optimal activity of *D. melanogaster* (6-4) photolyase has been measured at pH 7.8.[16] Another interesting observation in this context comes from Li *et al.*, who also studied the repair dynamics and the steady-state enzyme activity over a pH range of 7 to 9 and did not note any associated changes.[20] At the time of the aforementioned EPR study, however, the (6-4) photolyase crystal structure had not been solved. Given the relative positioning of the cofactor and the active site histidines apparent in the structure from *D. melanogaster*,[16] it is not entirely clear that the hyperfine coupling constants on the hydrogens of the cofactor are the most appropriate probes of the protonation states of the histidine residues.

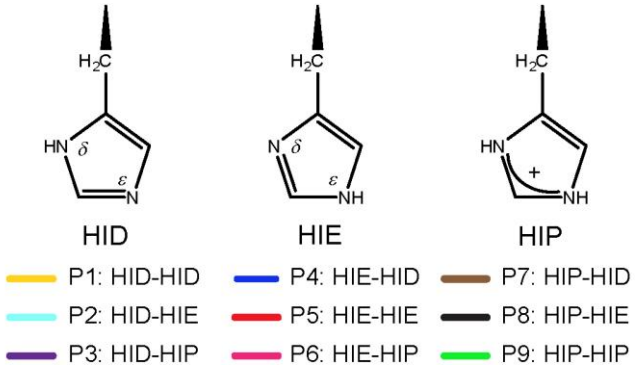
It is against this background that we present here a systematic investigation of the protonation states of the conserved histidine residues in (6-4) photolyase by means of various computational techniques. An individual histidine residue may adopt one of two neutral tautomers, with a proton on the Nε (HIE) or Nδ (HID) position, as well as one protonated form, with protons on both Nε and Nδ (HIP). With these three possible states for each histidine, there are a total of nine possible ways in which the protons can be distributed among two histidine residues. In the case of (6-4) photolyase, which has a complex H-bonding network connecting the lesion, catalytic residues and the cofactor, it is difficult to definitively discard any of the nine possibilities, purely on the basis of structural or intuitive criteria. With

this in mind, and in the interests of remaining systematic, we have explicitly taken all nine combinations into account.

As the first step in our analysis, we have examined the suitability of using the hyperfine coupling constants of selected hydrogens of the $FADH^{\bullet}$ radical as a quantitative indicator for the histidine protonation states. Specifically, using a combined QM/MM approach, we have calculated the relevant hyperfine coupling constants for all nine protonation states and compared them to the measured values. In the next step, we have applied a range of popular methods for estimating the pK_a values of the histidine residues in an effort to determine if one can arrive at a definitive assignment of the protonation state in this way. For this purpose we have investigated techniques of varying sophistication, ranging from basic web servers to more elaborate methods based on the solution of the Poisson-Boltzmann equation and free-energy cycles. Finally, we have carried out explicit molecular dynamics simulations on all nine possible combinations and monitored their structural evolution with respect to the crystal structure. Indeed, due to the interesting nature of the results of this latter aspect, we have applied an analogous approach to structures of the (6-4) photolyase complexed with variations of the T6-4C lesion[24,25] as well as of the repaired T-T product.[16] With the combination of these varied approaches, we are aiming to provide more definitive information concerning the protonation states of the conserved histidine residues and, in this manner, pave the way for further mechanistic studies.

4.2 Computational Details

4.2.1 Structural Models



Scheme 4-2.Possible combinations of protonation states (P*n*) of His365 and His369. Each state is defined as P*n*: HIX365-HIX369 (X=D, E, or P) and associated with a certain color. This notation will be used throughout the text.

Virtually all aspects of our study required the definition of a structural model. The initial configurations were taken from the appropriate crystal structure of (6-4) photolyase in complex with a DNA oligomer and the FAD cofactor. The relevant PDB entries were 3CVU, 3CVY,[16] 2WB2[24] and 2WQ7,[25] which contained the T6-4T lesion, the repaired T-T dimer, the T6-4C lesion, and the methylated mT6-4C lesion, respectively. Before further processing, the original PDB files were slightly modified by removing duplicate entries and assigning the protonation states of histidines, other than His365 and His369, based on their local environments. For each crystal structure, nine alternative versions were prepared according to the nine possible protonation states of His365 and His369 (Scheme 4-2).

4.2.2 Parameter Assignment

For standard nucleic acid residues, we assigned classical force-field parameters according to the AMBER99 (ff99) force field,[26] supplemented with the refined *parmbsco* parameters (ff99bsco).[27,28] For standard amino-acid residues, we employed the AMBER99SB (ff99SB) version of the AMBER protein force field.[29] For the non-standard residues, such as the DNA lesions and tautomers and the various oxidation states of the FAD cofactor (oxidized FAD, $FADH^{\bullet}$ radical and fully reduced $FADH^{-}$), missing parameters were derived using the Antechamber[30] suite and the GAFF force field[31] available in the AMBER9 program package.[32]

Initial coordinates for the auxiliary FAD and DNA lesions were extracted from the crystal structures and the geometries, with appropriately added hydrogen atoms, were optimized in the gas phase with Gaussian 03[33] at the HF/6-31G(d) level of theory (UHF for radical species). The electrostatic potential was determined for the optimum structures at the same level of theory. The final atom-centred point charges were fit to the resulting potential using the RESP procedure.[34]

Although the composite thymine monomers of the (6-4) lesions appear as separate residues in the PDB files, for ease of charge derivation, they were treated together as a single residue. We adopted a similar approach when treating an alternative tautomer of the repaired TT dimer. Accordingly, the ESP charges for the (di)nucleotides were derived such that the 5'-phosphate of the dimer was capped with a methoxy (O-CH_3) group, while the terminal O3' of the dimer was capped with a proton.[35] During the RESP fitting, the charges on the capping groups were required to sum to zero and were later excluded from the final library to allow the correct connectivity with the rest of the DNA molecule.

While the use of electrostatic potentials calculated using HF/6-31G(d) could be considered somewhat outdated, we have chosen this approach primarily to arrive at a balanced and consistent treatment of the electrostatic parameters for the protein, the lesions, the DNA fragments and the FAD cofactor. Indeed, the same reasoning lies behind the often employed combination of ff99SB and ff99bsco (both based of HF/6-31G(d) RESP charges) for systems that contain both amino and nucleic acids.[36] Force-fields based on more sophisticated electrostatic potentials are available for proteins. For example, AMBER03 uses charges derived from B3LYP/cc-pVTZ calculations in a polarizable continuum, which incidentally correlate quite well with those present in ff99.[37] However, analogous charge sets are not yet available for DNA fragments, where HF/6-31G(d) charges are still widely employed.[28,38] Thus to avoid mixing different charge-derivation protocols within our system, we have chosen to use charges based on the HF-RESP combination, throughout our study.

Using the standard (ff99SB, ff99bsco) and custom-built libraries, a topology file containing all necessary parameters for each system was built using the LEaP module of AMBER9. For those systems that were further subjected to molecular dynamics simulations in explicit solvent, each system was solvated with TIP3P waters in a truncated octahedron box, with edge length of ~40 Å. All crystal water molecules present in the PDB file were retained. Sodium ions were added to neutralize the charge. The number of added ions (15-20) varied with the protonation states of the histidine residues, length of DNA strands and the oxidation state of

FAD for each system. Overall, the sizes of the simulated systems ranged from 61862 (3CVY) to 74279 (2WQ7) atoms. Prior to any calculations, hydrogen bonds were oriented so as to produce an optimal (or at least non-clashing) interaction between the residues of interest (His365, His369, Tyr423, FAD, and the 2 nucleotide residues).

4.2.3 Calculation of EPR Parameters

Given that the hyperfine couplings (hfcs) of selected protons of the $FADH^{\bullet}$ radical were considered as probes for the protonation states of the nearby histidine residues in the EPR/ENDOR study of the *Xenopus laevis* (6-4) photolyase, we set out to calculate these hfcs using, primarily, the Gaussian 03 software package.[33]

Previous calculations of hfcs relevant to the ribonucleotide-reductase-catalyzed reaction[39] had found that the B3LYP/TZVP level of theory provided acceptable accuracy at a reasonable cost. In light of the relatively large systems considered herein, we have chosen this approach as our quantum mechanical benchmark.

We used the isolated $FADH^{\bullet}$ radical as a simple model for the purpose of method validation. The initial geometry was extracted from crystal structure and hyperfine couplings were calculated directly with B3LYP/6-31G(d). The structure was subsequently optimized at the same level of theory and hfcs were again computed with B3LYP/6-31G(d) and B3LYP/TZVP.

The influence of the surrounding protein on the hfcs was investigated with a series of QM/MM calculations, carried out within the ONIOM[40] formalism available in Gaussian 03[33] for the P2 protonation state. Three different sizes of the QM region, which was treated with B3LYP/6-31G(d), were trialled in these calculations. The simplest model (QM1) consisted of only the $FADH^{\bullet}$ radical. Beyond this, we expanded the high-level model system to include the most important catalytic residues (His365, His369 and Tyr423) in QM2 and, additionally, the T6-4T lesion in QM3 (see Figure 4.3). Only the sidechains of the amino acids, which were terminated between the C_{α}-C_{β} bonds, were retained in the QM region. The connections between the lesion and the DNA chain were introduced by placing link atoms between C4' and C5' ribose carbon atoms on both ends of the lesion.

The MM layer of the QM/MM models, whose parameterization was described above, consisted of the entire protein and the DNA duplex molecules surrounded by the 3000 water

molecules closest to FAD. The starting geometry for the P2 state was extracted after the solvated, protonated crystal structure had been allowed to relax (see steps (*i*) and (*ii*) in Section 4.2.5). Thereafter, a mobile region, consisting of all residues partially within 15 Å of the Nε atom of His365, as well as all water molecules partially within 12 Å radius of the same atom, was defined. With the remainder of the system frozen, the structure was optimized with mechanical embedding[40] in combination with the appropriate QM region (QM1-3). Subsequent evaluations of the hfcs within the QM/MM framework were carried out within the electrostatic embedding formalism.[41] After settling on QM3 as the most appropriate QM region, the above protocol was applied to calculate the hfcs for all nine protonation states (Scheme 4-2).

4.2.4 Structure-Based pK_A Calculations

The initial estimation of the pK_a values of His365 and His369 was performed using the PROPKA 2.0 server. Apart from being fully automated and very fast, PROPKA is a structure-based empirical method for the pK_a prediction of the ionizable residues in the proteins.[42] pK_a emerges as a sum of two terms ($pK_a = pK_{a,model} + \Delta pK_a$). $pK_{a,model}$ is associated with the unperturbed value for each residue (6.5 for His), whereas the shift (ΔpK_a) is treated as an environmental perturbation. Furthermore, for each ionizable group, ΔpK_a is given as a sum of perturbations due to desolvation, hydrogen bonding and charge-charge interactions ($\Delta pK_a = \Delta pK_{Desolvation} + \Delta pK_{HB} + \Delta pK_{ChgChg}$).[43] A pdb file is used as input to the server, which provides pK_a values in a fully automated fashion.

For comparison, we also used the H++ server.[44] In addition to pK_a (strictly $pK_{1/2}$) values for ionizable residues, the server provides an output structure containing the missing hydrogens added according to the results of the pK_a evaluations.[45] The automated calculation is based on continuum solvent methodology,[46] within the framework of either the generalized Born (GB) or linearized Poisson-Boltzmann (PB) models, and accounts for each of the 2^N protonation microstates (when N ionizable residues are active).[47] In our calculations, we used the PB model and kept the default values for ionic strength (0.15 M) and the external dielectric constant (ε_{ext} =80). We used an internal dielectric constant (ε_{in}) of 6, while the pH value was set to 7.

In principle, a ligand bound to the protein and its net charge can be included in H++ calculation, which is then parameterized in an automated procedure, although only one ligand can be processed per run. Because of this limitation, and the fact that we did not obtain

satisfactory automatic parameterizations of the lesions or the FAD, we chose to complete our H++ calculations in three sequential steps.

The first step in each calculation involved the application of the recently introduced *reduce* algorithm,[48] which is used to identify the preferred orientation for ambiguously placed heavy atoms, as well as to differentiate and assign the most probable histidine tautomers. The second step involved the calculation of the pK_a values and the associated protonation states for all ionizable residues in the resulting structure, in the presence of an approximate parameterization for FAD (oxidized or reduced) but in the absence of the lesion (except for 3CVY where the T-T pair was present). The final refinement step involved re-calculating the pK_a values for the two residues of interest, in the explicit presence of the properly parameterized FAD (oxidized or reduced) and the lesion.

This final step was achieved with the use of appropriate PQR files, which were prepared according to the previously outlined parameterization (ff99SB, ff99bsco, etc.), without the crystal waters. Using these files, we evaluated the pK_as of the two histidines together with all other titratable residues as well as in calculations where only a single target residue was allowed to titrate. Importantly, the relevant neutral tautomers (HID and HIE) for use in the PQR-based calculations were assigned based on the results of the second step above.

A related approach for obtaining protein pK_a values is the Adaptive Poisson-Boltzmann Solver (APBS).[49] In cases of a single titration event, the estimation of the pK_a value of the chosen residue is based on a rigorous free energy cycle involving the transfer of the titratable group from solution to the protein. The transfer free energies (ΔpK_a,) for protonated and deprotonated species are derived from the numerical solution of the Poisson-Boltzmann equation[50] using the Finite Element ToolKit.[51] The requisite model pK_a values in water (6.5 for His) were taken from the table provided by the developers.[52]

Using APBS, we have evaluated the transfer free energies associated with all possible single titration events from the states listed in Scheme 4-2, each in the presence the relevant DNA lesion and either oxidized or reduced FAD. We used the PDB2PQR[53] server to generate a template input file for the electrostatic energy calculations with recommended grid parameters for a given system. The PQR files used for APBS were analogous to those used for H++, with the added modifications required to close the free energy cycle. All calculations were carried out at 300K with a solvent dielectric constant of 80, an ionic strength of 0.15 M, and a protein dielectric constant of 6.

In line with standard practice, all pK_a calculations were performed directly on the crystal structures in the absence of water molecules. According to reference 16, the crystals of T6-4T were grown in a reservoir buffer with a pH of either 7.0 or 8.6, after which they were rinsed in a cryoprotection solution of pH 7.8. In this sense, one could argue that the use of solutions with elevated pH might introduce a bias towards neutral un-protonated states. In our opinion, however, such a bias is likely to be very minor.

4.2.5 Classical MD Simulations

The nine fully built systems (Scheme 4-2) were treated within periodic boundary conditions. Long range electrostatic interactions were calculated with Particle-Mesh Ewald (PME) technique with the default nonbonded cutoff of 8.0 Å to limit the direct space sum. The temperature in all simulations was controlled by coupling the system with the Berendsen thermostat.[54] An integration time step of 2 fs was used and the SHAKE algorithm was employed to constrain bonds involving hydrogen atoms during dynamics.

The equilibration of the system was carried out in several steps: (*i*) Steepest descent minimization was applied to the enzyme-DNA-FAD complex (solute) with harmonic positional restraints on solvent molecules (5 kcal/mol Å^2). (*ii*) Minimization was repeated with restraints on the solute (5 kcal/mol Å^2) and no restraints on the solvent. (*iii*) Heating dynamics was performed with continued solute restraints at constant volume (NVT). Thereby, the temperature was increased from 0 K to 300 K over 60 ps and kept at that value for another 30 ps. (*iv*) Minimization was carried out with reduced solute restraints (2.5 kcal/mole Å^2). (*v*) The system was again heated, as described above, with reduced solute restraints (2.5 kcal/mole Å^2). (*vi*) 150 ps of constant pressure (NPT) dynamics at 300 K, with isotropic position scaling at pressure of 1 bar and a pressure relaxation time of 0.2 ps was performed. Harmonic restraints in this run were applied only to the DNA duplex (2.5 kcal/mole Å^2) at 300 K. (*vii*) The equilibration was finished with an unrestrained NVT simulation at 300 K (150 ps).

The MD production runs of 2 ns were carried out with constant volume at 300 K, saving the snapshots every 4 ps. All the data obtained through molecular dynamics simulations were subsequently processed and analyzed using the *ptraj* module of the AMBER9 program package.

4.3 Results and Discussion

4.3.1 The EPR Hyperfine Structure

Using information obtained from an EPR study of the stable FADH$^{\bullet}$ radical in the (6-4) photolyase from *X. laevis* the active-site histidine residues were assigned a protonation state that would correspond to either P7 or P8 (Scheme 4-2) in the *D. melanogaster* enzyme. Specifically, the principal values of the hyperfine couplings of four selected protons (H5, H6, H1', H8α) were extracted by performing simulation and deconvolution of the pulsed ENDOR spectra to fit the experimental data. The spectra of FADH$^{\bullet}$ radical were measured in the wild type enzyme and in the H354A and H358A mutants, in the pH range between 6 and 9.5 and in the absence of substrate. The strongest shifts and intensity changes were observed for the H1' and H8α signals in the H354A mutant, which was interpreted as being indicative of a change in the protonation state of His358 when going from pH 6 to 9.5. The spectra of H358A mutant resembled those of the wild type more closely at both pH values. Based on these observations, and assuming that both histidines were protonated at pH 6, it was suggested that His354 (corresponding to His365) remains protonated over the measured pH range, while His358 (corresponding to His369) becomes neutral at higher pH values. Indeed, it is at a higher pH value (8.5) that the *X. laevis* (6-4) photolyase shows its maximum activity.

To investigate the connection between the hyperfine couplings (hfcs) of the FADH$^{\bullet}$ radical and the protonation states of the nearby histidine residues, we set out to calculate the hfcs from first principles, as described in Section 4.2.3. Interestingly, the hfcs calculated for the isolated FADH$^{\bullet}$ radical in the gas phase (Figure 4.3a) show reasonable agreement with the experimental data. The results obtained with B3LYP/6-31G(d) for the geometry extracted directly from the crystal structure (*D. melanogaster*) show the largest deviations from the experimental values, which were obtained from the *X. laevis* enzyme at pH 8. The hfcs resulting from optimizing the geometry at this same level of theory, however, parallel the trends in the experimental data remarkably well. The use of the larger TZVP basis set in conjunction with B3LYP has been previously shown to be a reliable method for this type of calculation.[39] The results presented in Figure 4.3a clearly show that, in this instance, the smaller basis set (6-31G(d)) can be safely used without any notable loss of accuracy.

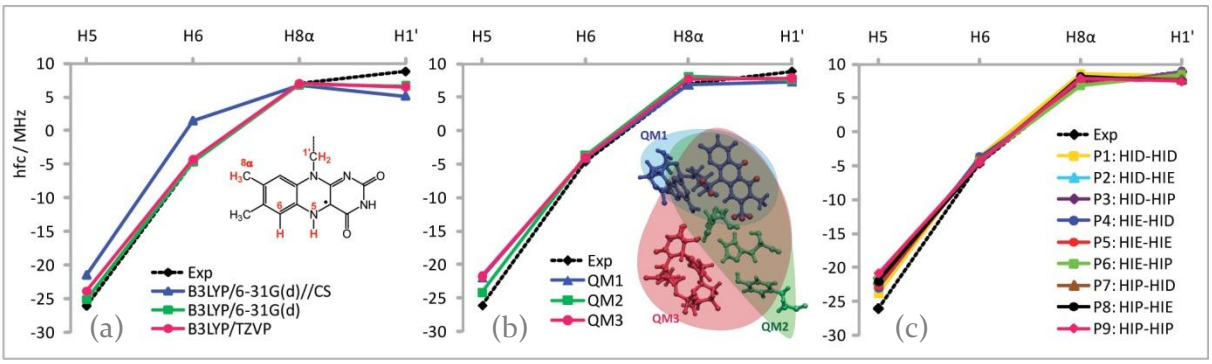


Figure 4.3 Calculation of hyperfine couplings of selected protons (colored in red) on FADH$^{\bullet}$ radical compared to experimental data: (a) gas-phase method validation (CS stands for crystal structure geometry); (b) QM/MM model validation for the P2 state; (c) calculated values for all 9 possible protonation states.

To establish a method that could potentially capture the effect of changing the histidine protonation states on the hfcs, we employed the ONIOM[B3LYP/6-31G(d):AMBER] method and tested it on three models with differing QM regions (QM1-3). The results in Figure 4.3b, which were obtained for protonation state P2 (Scheme 4-2), show that the calculated hfcs are not strongly dependent on the size of the QM region in such a treatment. Thus, while the smaller and more tractable QM1 could be expected to perform satisfactorily, we elected to use the more inclusive QM3 representation for reasons of completeness.

Figure 4.3c shows the hfcs obtained with the QM3/MM model for all nine possible protonation states of His365 and His369. From this figure, it is immediately clear that any of the nine combinations of HID, HIE, and HIP could equally well give rise to hfcs very close to the experimental values. This result indicates that any assignment of the protonation states of the active site histidines in (6-4) photolyase, made on the basis of the hfcs for the FADH$^{\bullet}$ radical, should be treated with some caution.

As outlined in Section 4.2.3, the results shown in Figure 4.3c were obtained for single protein conformations closely resembling the crystal structure. One could consider a more sophisticated comparison on the basis of protein conformations that more accurately represent the true equilibrium structure of each protonation state in solution. Indeed, we find that such an approach slightly improves the already close agreement with the experimental values, for the P2 protonation state (Figure 4.4a). A similar analysis across all nine possibilities (Figure 4.4b) confirms that, even when using relaxed protein structures, the hfcs do not constitute an adequate measure with which to distinguish the different protonation states. An extra level of sophistication, involving the calculation of the hfcs for numerous structures of a given conformational ensemble could also be considered.[55] In the current application, however, we

were primarily focused on comparing the different protonation states on an equal footing. In this sense, we wished to remove the static and dynamic structural variations from the comparison and emphasize the situation where the only difference between the nine examples shown in Figure 4.3c was, indeed, the protonation states of the active-site histidines.

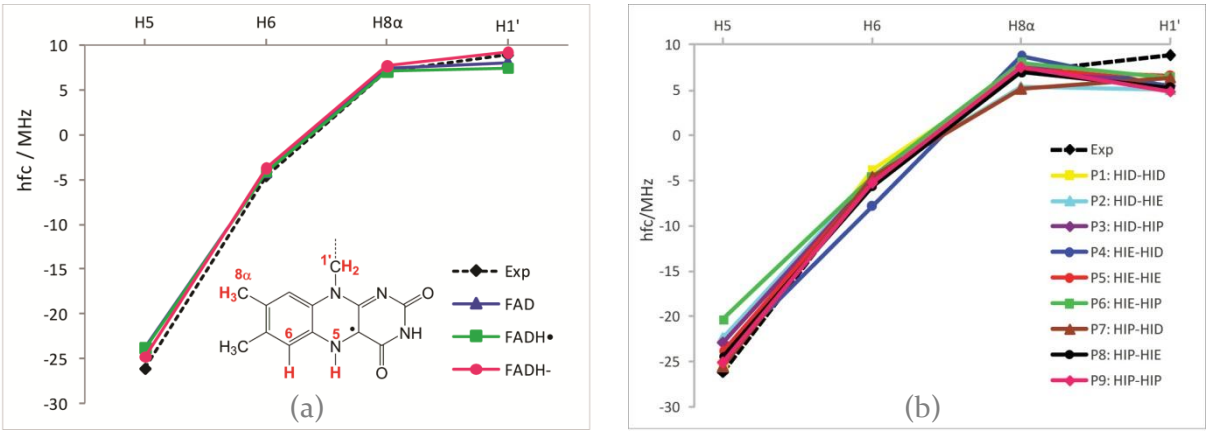


Figure 4.4 Hyperfine couplings calculated for selected protons of FAD cofactor with ONIOM[B3LYP/6-31G(d):AMBER], where QM region corresponds to the FAD cofactor (QM1). The structures for these QM/MM calculations were obtained by extracting the MD snapshot closest to the average structure of the production period of the MD simulations performed on: (a) the P2 state, for different oxidation states of FAD. The QM region ($FADH^{\bullet}$) was then optimized with mechanical embedding, followed by a *single-point* calculation of the hfcs within the electrostatic embedding formalism; (b) each P*n* state, carried out in the presence of the $FADH^{\bullet}$ form of the cofactor. The hfcs were obtained by a subsequent *single-point* calculation within the electrostatic embedding formalism.

Given the results arising from that analysis, it is rather difficult to assert that one can use the hfc values of the $FADH^{\bullet}$ radical to arrive at any definitive assignment of the protonation states of the histidine residues in the active site of (6-4) photolyase. On this basis, we feel that the investigation of alternative methods for the determination of the likelihood of these states is warranted and it is to this endeavour that we now turn our attention.

4.3.2 The pK_A Values

The assignment of proton positions for a structure derived from X-ray crystallography is a ubiquitous problem in the modelling community. Standard approaches range from the use of intuition to sophisticated algorithms designed to calculate pK_a values and automate the proton placement. In this section, we investigate the performance of several such approaches. In

addition to attempting to determine the implications for the active-site histidines in (6-4) photolyase, we also wish to exemplify the typical results, sensitivities, and limitations of these widely used procedures.

4.3.2.1 Empirical Approach

One of the fastest means to estimate the pK_a values of protein residues is through the use of the PROPKA server.[43] The empirical procedure estimates the pK_a values for all ionizable sites, on the basis of their environment. Table 4-1 shows the corresponding pK_a values for His365 and His369, obtained with several different crystal structures of (6-4) photolyase.

Table 4-1. Estimated pK_a values for His365 and His369 in (6-4) photolyase using PROPKA 2.0 server and different crystal structures available at the PDB.

PDB code: lesion	No ligands[a]		Ligands[a]	
	His365	His369	His365	His369
3CVU:T6-4T	-1.85	1.61	-8.09	-0.09
3CVY:T-T	-3.14	0.50	-8.20	-0.10
2WB2:T6-4C	-1.55	1.75	∞[b]	-1.45
2WQ6:T6(dew)4C	-1.77	1.58	-7.88	-0.14
2WQ7:mT6-4C	-1.93	1.53	∞[b]	-0.31
3FY4 (A)	-0.84	2.25	-7.57	-3.37

[a] The ligands are FAD and the DNA lesion, except for 3CVY:T-T, where FAD is the only ligand.

[b] Problems with determining contribution of the hydrogen bond between His365 and N6A atom of FAD to the p*K*a shift.

In the absence of the FAD and DNA-lesion ligands, the calculated pK_a values for His369 are uniformly low but positive. For His365, the values are even lower (Table 4-1, columns 2 and 3). Under these circumstances, the PROPKA method clearly predicts neither of the active site histidines to be protonated at a pH of ~ 7. The large deviations from the model pK_a value of 6.5 are due primarily to the desolvation term, which contributes $\Delta pK_{a,Des}$ ~ -4.5 for both residues. The remaining deviation arises from charge-charge interactions and hydrogen bonding, which have a stronger influence on His365 than on His369.

The effect of the FAD and DNA-lesion may be included in the PROPKA procedure, although one cannot distinguish between the oxidized and reduced forms of FAD. The

inclusion of these ligands causes the pK_a values of both monitored residues to shift towards even lower values (Table 4-1, columns 4 and 5). This effect, which is bordering on non-physical, is significantly stronger for His365, which participates in additional charge-charge interactions with the ligands. These additional interactions are not apparent in the case of His369, whose pK_a shift is mainly due to further increased desolvation effects. The relative importance of the DNA residues and the FAD cofactor can be gauged by comparing the results for 3CVY:T-T, where only FAD is considered as a ligand, to the remaining results.

Overall, the PROPKA methodology shows His369 to be more basic than His365. Although the absolute pK_a values are much lower than might be intuitively expected, it is clear that neither residue is predicted to be protonated at physiological pH. This preference for histidine neutrality persists irrespective of the absence or the presence of the DNA and the FAD ligand. While aspects of these results may be instructive, their reliability should be considered against the fact that the FAD charge cannot presently be explicitly accounted for. We therefore compare these findings to results obtained from more sophisticated approaches to obtaining pK_a values. The latter involve explicitly accounting for the electrostatic effects of the environment with the Poisson-Boltzmann equation as provided by the H++ server[44] and the APBS approach[49] (see Section 4.2.4).

4.3.2.2 Histidine Protonation States in the T6-4T Lesion

With its default settings, H++ provides pK_a values of all potentially charged residues in the protein, on the basis of self-consistent solution of the Poisson–Boltzmann equation for the lesion-related pdb structure of the protein. Since only one ligand is permitted in this way, the calculation is performed without the lesion and with an approximate parameterization of the FAD cofactor (Table 4-2, columns 2 and 3). The *reduce* algorithm,[48] can be used to estimate the protonation states of the system (Table 4-2, column 4). For the case of the T6-4T/FAD (3CVU), this yields pK_a values for His365 and His369 of -3.5 and 5.5, respectively. This result, which was obtained with all possible titratable residues active, is in agreement with the PROPKA predictions in that neither histidine is expected to be protonated at pH=7. Consistently, at this pH value, the most likely protonation state is found to be the neutral state P2 (HID-HIE).

Despite the fact that all the crystal structures were resolved with the FAD cofactor in the oxidized state (FAD), it is known that the reduced anionic form ($FADH^-$) is required for the

enzyme activity.[5] Following the same procedure discussed above and in Section 4.2.4, we have also calculated the pK_a values in the presence of the reduced cofactor. For the T6-4T structure (3CVU) in the absence of the lesion but in the presence of an approximate $FADH^-$ parameterization (Table 4-2, row 2), H++ yields pK_a values for His365 and His369 of 0.2 and 7.6, respectively. At a pH of 7, this result implies that His369 is likely to experience some degree of protonation and the most likely such state is found to be P3 (HID-HIP).

Table 4-2. Estimated pK_a values for His365 and His369 in D. melanogaster (6-4) photolyase obtained with H++ server and APBS software package. All the calculations were carried out at 300 K with an internal dielectric constant of 6 and an ionic strength of I=0.15 M.

Lesion/FAD	**Simultaneous titration of all residues**								
	H++ (pdb, reduce, no lesion)			H++ (PQR)		H++ (PQR)		APBS	
	His 365	His 369	State	His 365	His 369	His 365	His 369	His 365	His 369
T6-4T / FAD	-3.5	5.5	P2	-18.6	3.7	2.5	5.9	0.2	3.4
T6-4T / $FADH^-$	0.2	7.6	P3	-14.7	7.0	5.7	7.7	2.7	4.6
T-T / FAD	-6.6[a]	6.2[a]	P2[a]	-13.5	4.3	2.3	6.6	0.5	3.7
T-T / $FADH^-$	0.1[a]	10.0[a]	P3[a]	-4.7	7.7	5.4	8.2	2.7	4.4
T6-4C / FAD	-5.5	2.8	P2	-9.8	2.0	0.5	5.5	-0.4	2.9
T6-4C / $FADH^-$	-1.8	5.9	P2	-6.2	5.9	3.4	7.3	2.0	4.1
T6-4pC / FAD	-5.5	2.8	P2	-16.5	-2.1	-8.0	2.9	-8.4	1.8
T6-4pC / $FADH^-$	-1.8	5.9	P2	-12.3	0.2	-4.1	4.7	-4.6	2.5
mT6-4pC / FAD	/	/	/	-18.0	-2.7	-8.7	2.4	-9.3	1.0
mT6-4pC/$FADH^-$	/	/	/	-13.7	-0.3	-5.5	4.3	-6.8	2.1

[a]with the entire DNA

Once the initial characterization of the protonation states has been achieved, a more rigorous parameterization of FAD (oxidized and reduced) and the lesion can be included and pK_a values of *all* ionizable residues can be re-calculated. As can be seen from Table 4-2 (columns 5 and 6) this reduces pK_a values of both histidines in the active site, in a similar manner to that observed in Table 4-1. For the oxidized FAD, the reduction in the case of His365 (to -18.6) is more dramatic than that for His369 (to 3.7). The large, and presumably non-physical, negative value obtained for His365 is best interpreted to mean that, under these conditions, this residue is unlikely to be protonated at any accessible pH and that the histidines remain in the P2 state. Inclusion of the more realistic parameterizations of the reduced cofactor FADH- and the lesion again reduces both pK_a values. While His365 (-14.7)

again appears very unlikely to be protonated, His369 (7.0) is expected to be half-protonated at pH 7. Hence both the P2 and P3 states could be potentially observable.

Another informative measure of acidity is the intrinsic pK_a for a given residue. This value is obtained by calculating the free energy difference between the protonated and non-protonated forms, under the condition that all other residues remain in their standard protonation states. The intrinsic pK_a values for His365 and His369, given by H++ (Table 4-2 columns 7 and 8) for the T6-4T lesion (3CVU) in the presence of oxidized FAD and the lesion, are 2.5 and 5.9, respectively. Although these values are higher than those obtained when all ionizable residues are active, they still indicate the (neutral) state P2 to be the most likely at a pH of 7 (or higher). The APBS results for the equivalent system (Table 4-2, last two columns) are systematically lower than the H++ predictions (0.2 and 3.4 for His365 and His369) but concur that neither histidine is likely to be protonated under these conditions.

Calculations with H++ on the level of a single residue in the presence of FADH- give intrinsic pK_a values of 5.7 and 7.7 for His365 and His369, respectively. These values are again higher than those which account for changes in the other ionizable residues and place the pK_a of His369 above 7. The equivalent APBS results for the intrinsic pK_a values (2.7 and 4.6 for His365 and His369), however, are again systematically lower than those obtained from H++. In the case of T6-4T with the reduced lesion, the APBS approach does not find the pK_a of His369 to be greater than the pH of 7. On this basis, the APBS methodology favours the neutral protonation state P2 over the charged P3 state that is preferred by H++. Irrespective of this fact, the two methods agree that the presence of the cofactor in its reduced state increases the pK_a values of both residues. The magnitude of this effect is somewhere between 1 and 4 pK_a units but is uniformly smaller with APBS than with H++.

4.3.2.3 Histidine Protonation States in the T-T Lesion

The results obtained in this case (Table 4-2, rows 3 and 4) largely reproduce the trends discussed for the T6-4T lesion although the effect of including the ligands appears to be somewhat reduced. This is largely due to the fact that the two key thymine residues are already present in the direct analysis of the crystal structure. In the presence of the oxidized FAD cofactor, the neutral protonation state P2 (HID-HIE) is preferred for the T-T dimer and the pK_a of His369 is below 7 in all cases.

The introduction of the reduced $FADH^-$ cofactor shows an initial preference for the charged P3 state (HID-HIP), which persists for all results obtained with H++. The intrinsic pK_a of His369 obtained with APBS in the presence of $FADH^-$ (4.4) is, however, below 7 indicating a preference for the P2 state. Interestingly, the APBS pK_a for His369 (4.4) is actually lower for T-T than for T6-4T (4.6).

4.3.2.4 Protonation of Histidines in the T6-4C Lesions

The structure containing the T6-4C lesion (2WB2) carries with it the added complication that the primary amine group of the lesion may be in the protonated or non-protonated form. This factor does not influence the initial H++ calculations (Table 4-2, columns 2-4), where the P2 state is predicted to be the most favourable for both forms of the FAD cofactor. Inclusion of realistic ligand parameterizations continues this preference. The only pK_a value to (marginally) exceed 7 corresponds to the H++ result for the intrinsic pK_a of His369 (7.3), in the presence of the neutral amine and $FADH^-$. Protonation of the amine (T6-4pC) significantly suppresses the pK_a values of both histidines and is associated with a very clear preference for the P2 state. For the methylated T6-4C lesion (2WQ7), where we only considered the protonated form of the lesion (mT6-4pC), we were unable to obtain a result directly from H++ with the pdb file. The direct inclusion of the lesion and the FAD cofactor, however, produced results very similar to those obtained without the methyl group.

Perspective of the pK_a study. The majority of the results presented in this section point to an overall neutral protonation state of the two histidine residues at a pH of 7. The *reduce* algorithm finds the P2 combination to be the most likely such state. Certain circumstances lead the H++ procedure to predict the potential occurrence of the protonated P3 state at neutral pH. However, it is interesting to reiterate that, in all such cases, the intrinsic pK_as from the more rigorous APBS procedure are uniformly below 7. Furthermore, the APBS method systematically produces lower pK_a values than H++, even when the two are directly comparable.

Despite the apparent semi-definitive nature of these results, several cautionary statements are in order. It is evident from Table 4-2 that the results can vary widely depending on the way in which the calculations are performed. Indeed, certain circumstances can result in disturbingly large negative pK_a values for His365, the physical meaning of which is difficult to

fathom. There is also a marked sensitivity of the results to input parameters such as the internal dielectric constant and the ionic strength (see, for example, Table 4-3).

Table 4-3 Estimated pK_a values for His365 and His369 in D. melanogaster (6-4) photolyase obtained with APBS software package. The values were calculated for T6-4T lesion and for all possible combinations for deprotonation of two histidines (HIP365-HIX369 → HIY365-HIX369 and HIX365-HIP369 → HIX365-HIY369, X=D, E, P and Y=D, E). All the calculations were carried out at 300 K with different sets of parameters given by the internal dielectric constant (PDIE) and the ionic strength (I).

PDB code: Lesion	Cofactor oxidation State	Int. dielectric constant/ Ionic strength	HIP365-HIX369						HIX365-HIP369					
			HID365-HIX369			HIE365-HIX369			HIX365-HID369			HIX365-HIE369		
			HID	HIE	HIP	HID	HIE	HIP	HID	HIE	HIP	HID	HIE	HIP
3CVU: T6-4T	FAD	PDIE=6	4,78	0,22	-3,20	1,53	-0,55	-3,13	5,20	3,14	-1,45	3,38	3,02	0,39
	$FADH^-$	I=0.15M	6,88	2,71	-1,05	5,05	3,21	0,32	6,46	4,14	-0,17	4,61	3,84	1,27
	FAD	PDIE=20	9,21	7,68	5,87	7,94	7,23	5,78	9,46	8,49	6,46	8,91	8,55	7,17
	$FADH^-$	I=0 M	10,31	8,85	7,00	9,44	8,77	7,28	10,35	9,30	7,34	9,69	9,26	7,89
	FAD	PDIE=20	6,58	5,07	3,54	5,53	4,82	3,62	6,63	5,75	3,96	6,02	5,72	4,58
	$FADH^-$	I=0.15M	7,44	6,04	4,47	6,81	6,16	4,92	7,25	6,35	4,64	6,58	6,23	5,09

On this basis, one needs to treat the conclusions derived from such results with some apprehension. It is because of this uncertainty that we have also probed an alternative route to assigning the correct protonation state for the photolyase enzymes. Namely, we have systematically investigated the structural stability of various protonation states with classical molecular dynamics simulations. The associated results are presented in the subsequent section.

4.3.3 Molecular Dynamics Study

In addition to the spectroscopic (EPR) and energetic (pK_a) criteria discussed above, we have employed a structural criterion to address the question of the most likely protonation states in the (6-4) photolyases. Specifically, we have run classical MD simulations of all 9 possible protonation states and monitored their respective deviations relative to the relevant crystal structures. The rationale behind choosing such a criterion derives from the observation that the (6-4) photolyase enzymes can repair the DNA lesion in their crystalline form,[16] meaning that the conformation of the active site present therein is a productive one. Protonation states

that maintain the crystal structure arrangement throughout unrestrained MD simulations are therefore more likely to be relevant for repair than those that deviate strongly from it. Because the crystal structures were resolved with oxidized FAD, while the reduced form ($FADH^-$) is required for activity, we have performed simulations in the presence of both cofactor states. For completeness, we have also simulated the radical form ($FADH^•$).

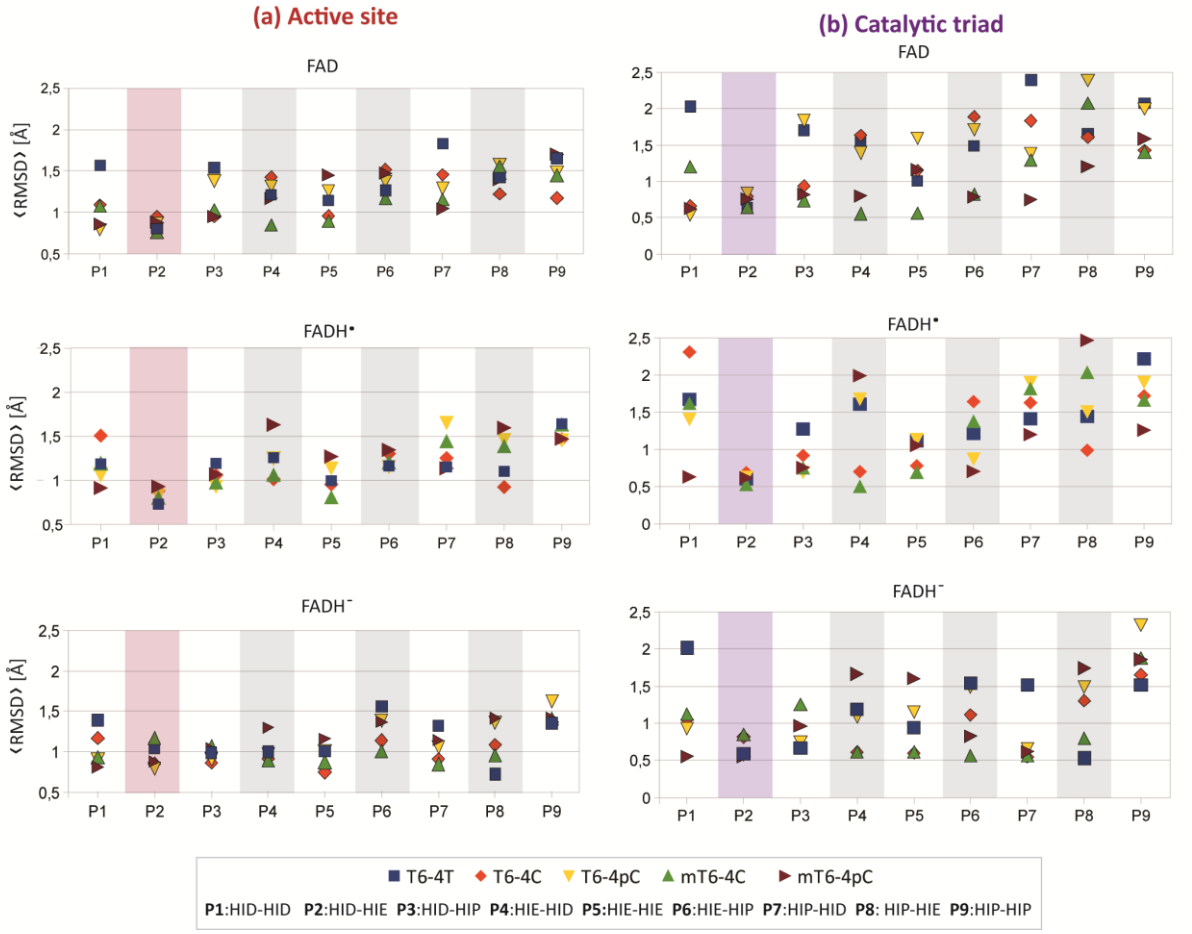


Figure 4.5 RMS deviations from the crystal structure positions during 2 ns simulations for: (a) active site for keto and enol forms of the repaired T-T dimer; (b) catalytic triad (His365, His369, Tyr423); (c) the lesion or the repaired T-T dimer.

Figure 4.5 shows a summary of the MD simulations that we carried out for systems containing the T6-4T lesion (3CVU), the T6-4C lesion (2WB2), and the N4-methyl T6-4C lesion (2WQ7). For the T6-4C and N4-methyl T6-4C lesions, we performed simulations in which the primary and secondary amino groups were both protonated (T6-4pC, mT6-4pC) and deprotonated (T6-4C, mT6-4C), respectively. Each depicted point in Figure 4.5 represents an average of the RMSD deviation, from the relevant crystal structure, over 2 ns of production dynamics. For selected examples, we elected to extend the simulations to a length of up to 5 ns. The corresponding results show the behaviour over the longer simulation parallels that

observed in the shorter one. This is especially true for those protonation states that remain close to the crystal structure and supports the hypothesis that the results are not artificially biased by short simulation times.

For the systems presented in Figure 4.5 it can be seen that the average RMSD values for the active site (Figure 4.5a, defined as His365, His369, Tyr423, FAD and the lesion) parallel the analogous values for the catalytic triad (Figure 4.5b, His365, His369 and Tyr423) rather well. This implies that the motion of the catalytic residues comprises a major component of the dynamics in the active-site and thus serves to highlight the importance of protonation states of the two histidine residues.

In the presence of the neutral form of the cofactor (FAD), the neutral states (P1, P2, P4, and P5) generally exhibit lower deviations from the crystal structures than states involving a protonated histidine (P3, P6-9). Particularly consistent is the state P2. This is in agreement with the pK_a calculations presented above, which indicated a clear preference for neutral over protonated states in the presence of FAD (Figure 4.5a). The introduction of the reduced cofactor ($FADH^-$) results in an increased stability of the protonated states, which is manifested in slightly less dispersed average RMSD values for the triad (Figure 4.5b). Specifically, the state P3 appears as particularly stable. This observation is also consistent with the calculated pK_a values, including the observed shift towards higher values when FAD is reduced to $FADH^-$.

Even though the P2 state and, to a lesser extent, the P3 state exhibit the lowest RMSD deviations in general, it is obvious from Figure 4.5 that several other states show low deviations for specific structures. Despite the low RMSD values, however, not all of these states correspond to the preservation of the initial hydrogen bond network. Namely, even though the residues may stay close to their initial crystal-structure positions, they frequently experienced a modified orientation, with a corresponding rearrangement of the hydrogen bonds. This situation occurs, for example, for P7 in combination with $FADH^-$ (Figure 4.5b). On the other hand, the largest RMSD deviations generally (but not exclusively) correspond to the complete fragmentation of the H-bond network in the active site. It has been observed that the hydrogen bond between His369 and Tyr423 is disrupted most frequently. This causes an increased mobility of His369, which, in many cases, results in the loss of the H-bond to His365 and a consequent divergence from the initial geometry. In this context, it is worth mentioning that the P2 state is the only arrangement capable of restoring the original hydrogen bonding once it has been perturbed. This feature was not observed for any other state.

4.3.3.1 Simulations with the Repaired DNA

In addition to the results on the lesions presented in Figure 4.5, we have also carried out a series of simulations on the repaired T-T dimer (3CVY, Figure 4.6) in the presence of $FADH^-$. Since keto-enol tautomerization of 3' base is likely to be the last step of the repair (see, e.g., Scheme 4-1), we have performed simulations with both the enol and keto form of the repaired 3' thymine. The upper panels of Figure 4.6 show the corresponding average RMSD deviations of the active site and catalytic triad, respectively.

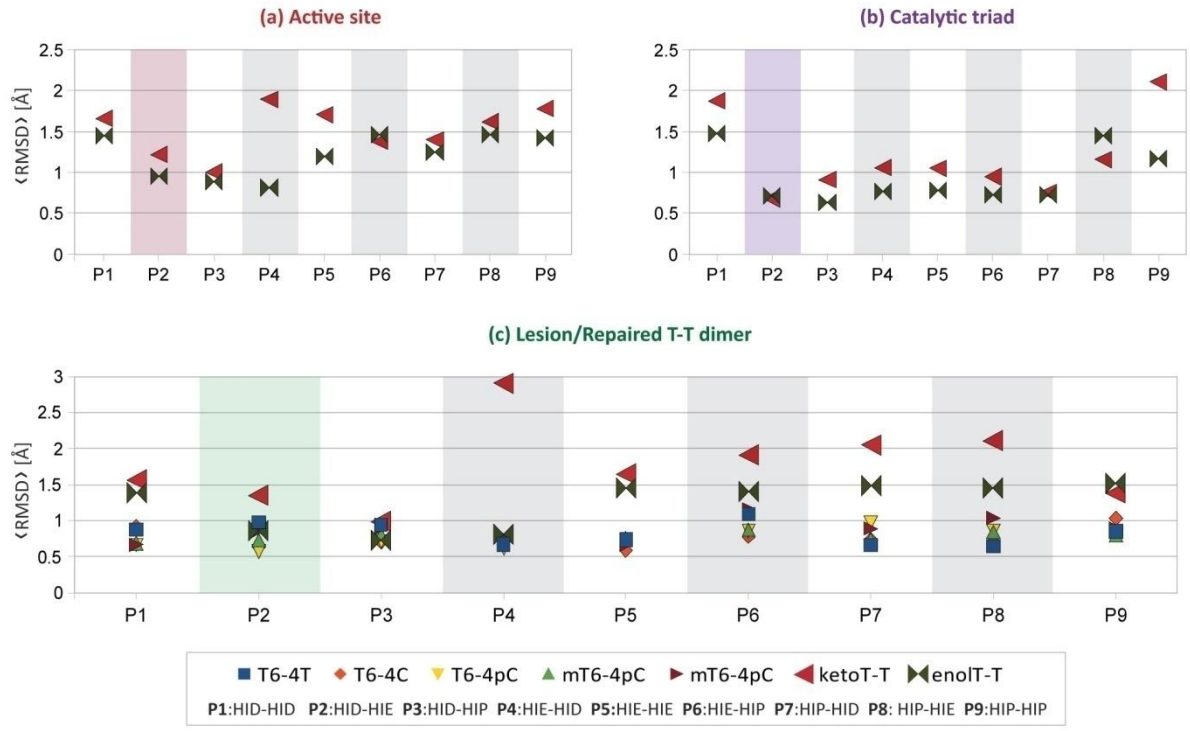


Figure 4.6 RMS deviations from the crystal structure positions during 2 ns simulations for: (a) active site for keto and enol forms of the repaired T-T dimer; (b) catalytic triad (His365, His369, Tyr423); (c) the lesion or the repaired T-T dimer.

In contrast to the results for the lesions in Figure 4.5, it is clear that the variations in the catalytic triad do not correlate exceedingly well with the overall active-site behaviour. Instead, the overall active-site deviations are much more closely related to the variations of the T-T dimer (lower panel Figure 4.6). Indeed, those variations can be seen to be stronger for the repaired T-T bases than for any of the lesions shown in Figure 4.5.

The deviation of the catalytic triad from its crystal structure position is seen to be relatively insensitive to the protonation state of the histidine residues (upper left panel in Figure 4.6),

with all but three states (P1, P8 and P9) remaining comparatively close to their original positions. The DNA bases show a stronger dependence on the protonation states (lower panel Figure 4.6). Once again, the P2 and P3 states stand out as being able to maintain the closest resemblance to the crystal structure conformations.

Even though both the keto and enol forms of the repaired thymine dimer show stronger deviations than the lesions, the effect is definitely more pronounced for the keto form (lower panel Figure 4.6). The variations of the enol form, on the other hand, are of the same order as those of the lesions. The reduced mobility of the enol form, relative to the keto, appears to be due to its ability to engage in hydrogen bonds, through the hydroxyl group, in a similar manner as the lesions. The larger amplitude of the DNA motion, especially in the case of the keto form, leads towards the loss of the hydrogen bonds between the 3'-thymine and the histidine residues, allowing for a greater flexibility of the active site.

The greater flexibility observed for the repaired DNA is of interest from the perspective of the product release. Generally, the binding of the DNA strands to the repair enzyme is achieved by flipping the lesion out of the double helix and into the active site. Thus there is a need to restore the helical structure after the repair is complete. The results presented in Figure 4.6 indicate that the final enol-keto tautomerization, and the resulting hydrogen bond losses, could provide the trigger that reduces the strength of the binding to the enzyme and allows this restoration to take place. We see evidence of this in the behaviour of the DNA backbone in the simulations. The bend in the backbone, which is caused by the presence of the lesion, tends to flatten out in the repaired DNA. This effect is considerably more pronounced in the keto form, pointing to an increased freedom for the DNA strand to return to its preferred helical form.

4.3.3.2 A Free Energy Cycle

To better understand the meaning of the pK_a calculations, and their connection to the molecular dynamics simulations, it is instructive to consider the cyclical connectivity of the protonation states via single titration events. Such a cycle is shown for the four neutral and four singly-charged protonation states in Figure 4.7. Using the APBS results for T6-4T ($FADH^-$) shown in Table 4-2, and the expression $\Delta G°=-RT\ln K_a$, one can assign relative free energies separately to the neutral and the charged states. Setting the respective free energies of P2 and

P3 to zero, each four-membered cycle closes to within less than 1 pK_a unit. The actual free energy values are therefore given as an average of the forward and reverse paths to a given state from P2 or P3, as appropriate.

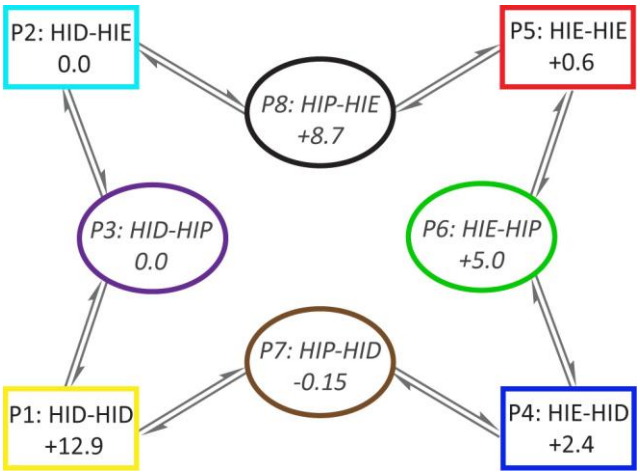


Figure 4.7.Cyclical representation of the four neutral (rectangles) and four singly-charged (ovals) protonation states connected by single titration events. The APBS free energies (in kJ mol^{-1}), which are relative to P2 and P3, respectively, are shown for the case of T6-4T in the presence of $FADH^-$. See also Table S3.

The final calculated pK_a values of the histidine residues are crucially dependent on the selection of the appropriate neutral reference state. In the analysis of Table 4-2, the neutral reference state was assigned as P2 on the basis of the *reduce* algorithm. This assignment is strongly supported by the molecular dynamics simulations as well as the neutral free energy cycle (Figure 4.7). Direct protonation of the P2 state can yield only P3 or P8, the former of which is shown to be favored by the pK_a calculations (Table 4-2), the free energy cycle (Figure 4.7), and the molecular dynamics simulations (Figure 4.5). On this basis, the P2 to P3 titration event, which is documented in Table 4-2, would indeed appear to be the most relevant.

Interestingly, the P7 state is found to be slightly lower in energy than P3 by the free-energy cycle (Figure 4.7). P7 cannot arise directly from P2 and the two neutral states that could result in its direct formation (P1 and P4) are not likely to be competitive with P2 (Figure 4.5 and Figure 4.7). However, P7 could be formed from P3 through a proton transfer between the two Nε atoms of His369 and His 365 (see Figure 4.2). While our molecular dynamics calculations indicate that P7 would not be structurally stable on the ns time scale (especially for T6-4T - Figure 4.5), this result opens the intriguing possibility that the P7 state could be a viable intermediate in the repair mechanism,[19] if protonation indeed plays a role.

4.4 Conclusion

It has been established that the active site of the (6-4) photolyases, capable of repairing UV-induced lesions in DNA strands, contains two histidines (His365 and His 369 in *D. melanogaster*) and a tyrosine, which are key residues in catalysis. Based on the EPR/ENDOR study of the *Xenopus laevis* (6-4) photolyase, several proposed mechanisms assume that the two histidines act as an acid-base pair with His365 being the protonated residue. However, because the two histidines are relatively closely positioned in the active site, it may be expected to be difficult to resolve their protonation states by the use of the EPR-derived hyperfine couplings (hfcs). To investigate this phenomenon, we used the ONIOM QM/MM approach to calculate the relevant EPR hyperfine couplings for all 9 possible combinations (P1-P9) that two histidines with various protonation states can adopt. Surprisingly, all combinations resulted in hfcs in good agreement with the experimentally measured values. Under such circumstances, it seems difficult to make use of the EPR data as a definitive criterion for assigning the protonation states of the histidines.

In order to resolve the problem, we proceeded to explore the active site from both the energetic and structural points of view, for various oxidation states of the FAD cofactor and several available lesions. The energetics aspects of the system are reflected in calculations of the pK_a values for the two histidines. The results based on the solution of the Poisson-Boltzmann equation, combined with the *reduce* algorithm, tend to favour an overall neutral state in the presence of the oxidized FAD cofactor, identifying the P2 state (HID-HIE) as the most relevant one. Introduction of the reduced cofactor ($FADH^-$) increases the pK_a values of both histidines and, in selected cases, causes the pK_a of His369 to exceed the reference pH of 7. In these cases, the state P3 (HID-HIP) emerges as the most likely combination. In this context, it is worth noting that the H++ procedure results in pK_a values that are systematically higher than those from APBS, which are uniformly below 7. At the same time it is pertinent to recall that (6-4) photolyase exhibits strong pH dependence and reaches maximum activity around pH ~ 8. It is reasonable to expect that, under these conditions, the histidines would prefer a neutral state, even according to the higher pK_a values obtained from H++.

Interestingly, the neutral P2 state (HID-HIE) also stands out from the structural point of view, as reflected in the results from molecular dynamics simulations. Of the neutral states, P2 exhibits the lowest structural deviations from all the experimentally-derived structures. The P2 state also exhibits generally low deviations from the crystal structures in the presence of the

reduced cofactor ($FADH^-$). This same HID-HIE combination is found to be unique in its ability to maintain the hydrogen bond network between the catalytic residues for all types of lesion, resulting in the lowest average triad deviations for all simulated systems (Figure 4.8). This is true irrespective of the oxidation state of the cofactor, including the $FADH^\bullet$ variant.

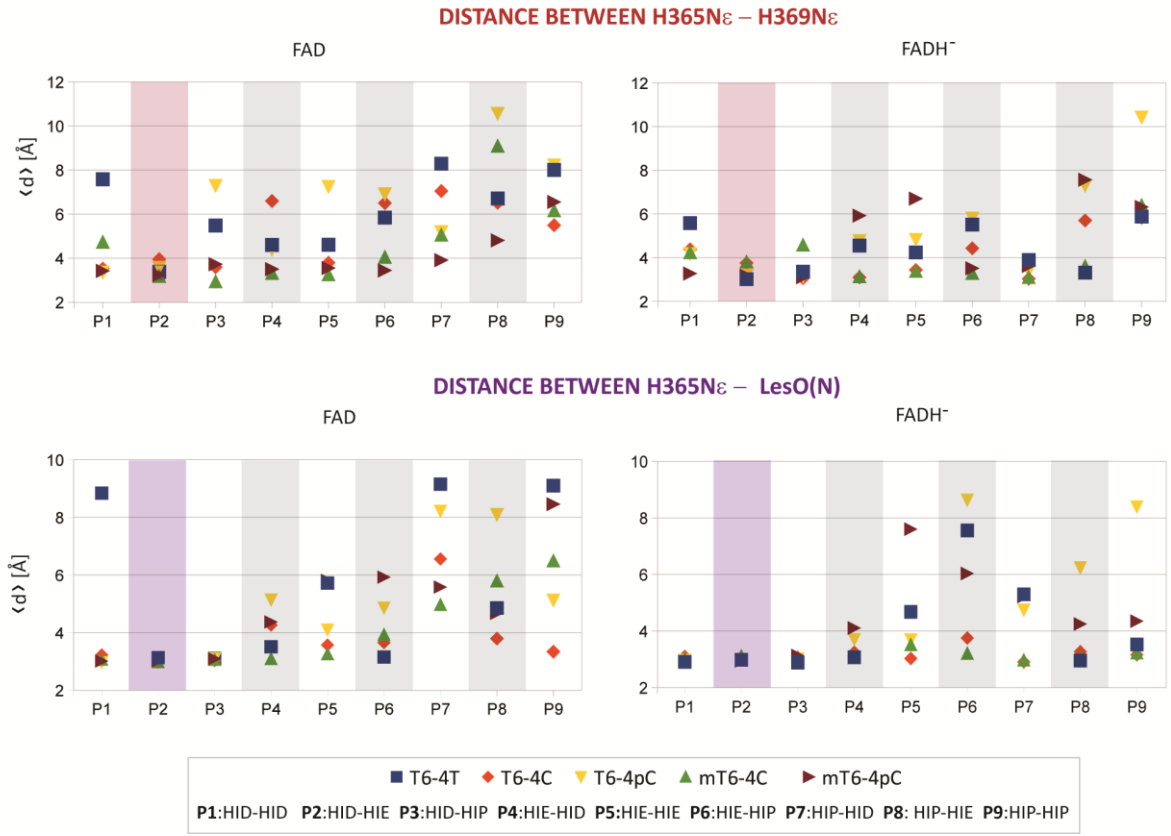


Figure 4.8 Average interatomic separations for H365Nε–H369Nε and H365Nε–LesO(N) (where LesO(N) represents the heavy atom of the hydroxyl or amino group of the T6-4T or T6-4C lesion, respectively) during 2 ns MD simulations of systems containing oxidized (FAD) and reduced (FADH-) cofactors. The corresponding values in the T6-4T crystal structure are both equal to 2.8 Ångstroms. Results are presented for simulations containing T6-4T lesion (3CVU), T6-4C lesion (2WB2), and the N4-methyl T6-4C lesion (2WQ7). The T6-4C and N4-methyl T6-4C lesions come in their deprotonated (T6-4C, mT6-4C) and protonated (T6-4pC, mT6-4pC) forms.

According to the structural criterion, the best performing of the protonated states is P3 (HID-HIP). Although this state fails to keep the crystal structure arrangement in the presence of oxidized FAD, the inclusion of $FADH^-$ significantly improves its ability to maintain the initial structure. Indeed, under these conditions, the P3 state shows a performance comparable to that of P2. Once again, there is an interesting correlation between the structural and energetic criteria, in the sense that the pK_a/*reduce* results also identified P3 to be the most likely of the protonated combinations. Analysis of a free energy cycle connecting eight of the protonation states further supports the appropriateness of P2 as the neutral reference state and

P3 as its relevant protonated counterpart. Interestingly, P7 is found to be associated with a free energy marginally lower than P3 and could thus participate in the repair mechanism, in so far as protonation does take place,

We find that the stability of the active site is intimately related to the stability of the resident hydrogen bond network. Once again, the P2 state is found to be the only combination capable of restoring the network after a disruptive fluctuation. It appears that the stability of the network may also be significantly affected by water molecules present in the active site. This is relevant in the context of a recent suggestion that a water molecule, hydrogen bonded to the OH group of the lesion and to Nε of His369, plays a significant role in the mechanism of the repair.[19] This water has a high crystallographic B-factor.[16]

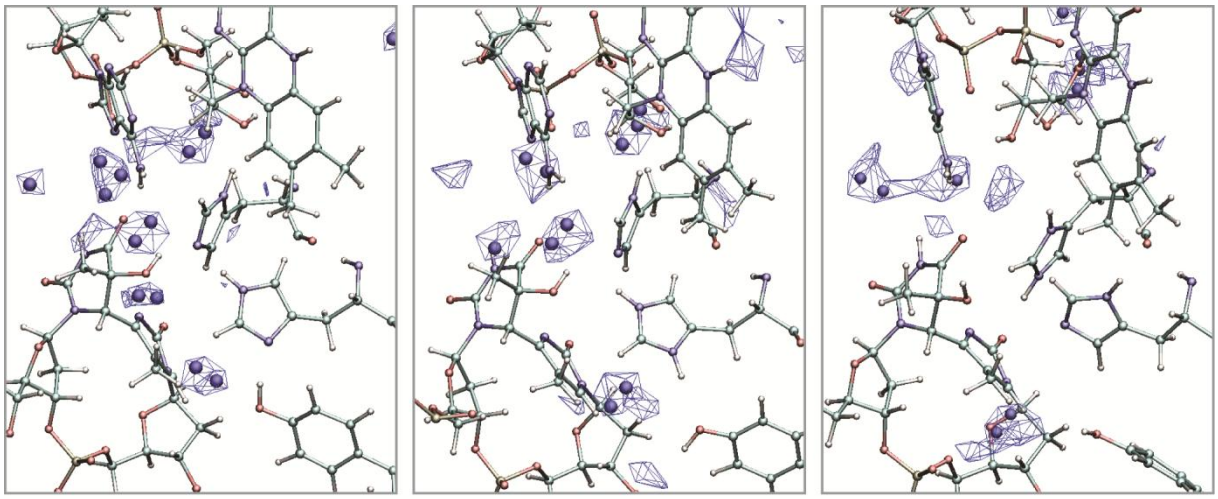


Figure 4.9 Graphical representation of the most probable placements of water molecules surrounding the active site for selected protonation states from simulations of the T6-4T lesion (3CVU) with $FADH^-$, obtained over the 2 ns production period. The probability is given in the form of a histogram of water molecules on a 3D grid, calculated with the ptraj module of the AMBER package and depicted in VMD using XPLOR density formatted files (wire) and pseudoatoms (spheres) as points defining most probable grid entries (> 50% of maximum value for the given system). The placements close to the hydroxyl group of the lesion in P2 and P3 are consistent with the pH-dependent water occupancy shown in Figure S4 of reference 16.

Its position is thus not well resolved and it might be quite dynamic. In our simulations, a number of water molecules surrounding the active site are found, often making a hydrogen bond with the OH group of the lesion. The "bridging" conformation, however, is rarely present. We have also observed that preservation of the crystal structure conformation tends to prevent the entrance of additional water molecules to the active site, while the disruption of the initial hydrogen bonding network is much more permissive in this respect. This effect can be seen in

the radial distribution of water around key atoms involved in the active-site network and in plots of the water density in the active site (Figure 4.9). The states that remain closest to the crystal structures (P2 and P3) can also be classified as the "driest", in that they are associated with less water in and around the active site than those states that diverge from the experimental reference (Figure 4.10).

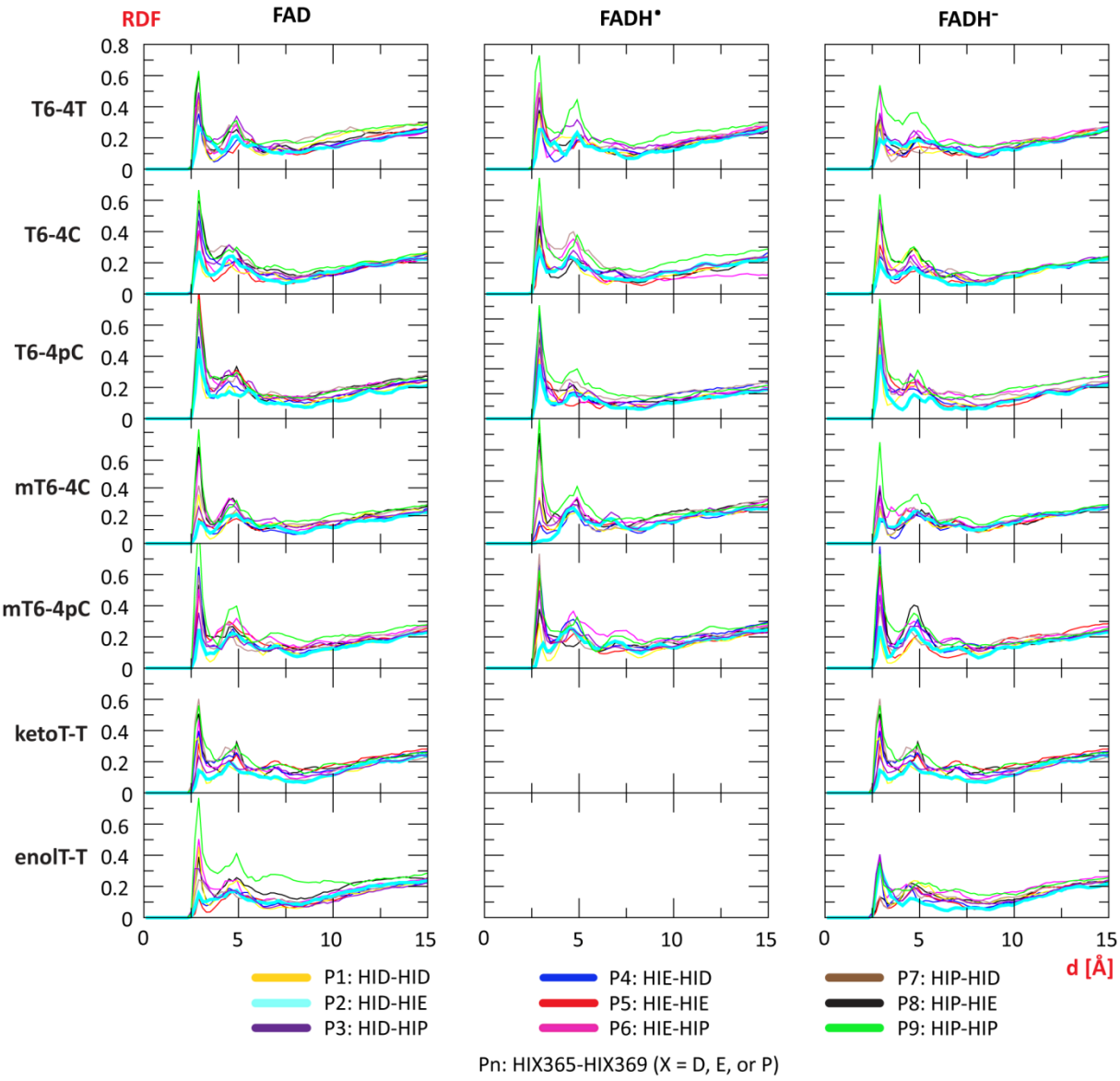


Figure 4.10 Radial distribution function (RDF) of water molecules around the active site. RDF is calculated in respect to the selected atoms (His365: Nε and Nδ, His369: Nε and Nδ; T6-4T: OH) for period of 2 ns of free NVT dynamics in MD simulations of the (6-4) photolyase systems containing T6-4T lesion (3CVU), T6-4C lesion (2WB2), N4-methyl T6-4C lesion (2WQ7) and the repaired T-T dimer (3CVY). T6-4C and N4-methyl T6-4C lesions come in their deprotonated (T6-4C, mT6-4C) and protonated (T6-4pC, mT6-4pC) form, while the repaired T-T dimer base comes in the form of keto and enol tautomer (3'thymine). State P2 is represented with a thicker line in cyan.

The success of state P2 has been further confirmed in the set of simulations performed with the repaired DNA bound to the active site of the enzyme. We observe that the repaired DNA undergoes larger fluctuations and drifts away from the active site, especially when the keto tautomer is present. This is consistent with a recent study of product release,[56] which suggested that restoration of the original bases occurs on a picosecond time scale, while the conformational changes leading towards product dissociation take place on the observed time scale of 50 μs.

In conclusion, it is useful to recall that the traditional view of the (6-4) photolyase enzymes has His365 as the protonated residue in the active site. In contrast, our systematic analysis of spectroscopic, structural and energetic aspects provides strong evidence that the P2 (HID-HIE) combination is the dominant protonation state for the His365 and His369 residues. We find that the presence of the reduced cofactor ($FADH^-$) endows His369 with a higher propensity to become protonated (HIP), although subsequent proton transfer to His365 is possible. However, considering the pH associated with maximum activity, it is quite likely that protonation does not play a role in the mechanism. With the proton distribution in the active site of (6-4) photolyase thus established, we may confidently focus our attention towards the characterization of a consistent mechanism for the lesion repair.

4.5 REFERENCES

1 Sancar, A. *Chem. Rev*. **2005**, 103, 2203.

2 Chang, C.-W.; Guo, L.; Kao, Y.-T.; Li, J.; Tan, C.; Li, T.; Saxena, C.; Liu, Z.; Wang, L.; Sancar, A.; Zhong D. *Proc. Natl. Acad. Sci. USA* **2010**, 107, 2914.

3 Harrison, C. B.; O'Neill, L. L.; Wiest, O. *J. Phys. Chem.* **2005**, 109, 7001.

4 Masson, F.; Laino, T.; Röthlisberger, U.; Hutter, J. *ChemPhysChem* **2009**, 10, 400.

5 Cichon, M. K.; Arnold, S.; Carell, T. *Angew. Chem., Int. Ed*. **2002**, 41, 767.

6 Sancar, A. J. *Biol. Chem*. **2008**, 283, 32153.

7 Rahn, R. O.; Hosszu, J. L. *Photochem. Photobiol*. **1969**, 10, 131.

8 Clivio, P.; Fourrey, J.-L.; Gasche, J.; Favre, A. *J. Am. Chem. Soc*. **1991**, 113, 5481.

9 Kim, S. T.; Malhotra, K.; Smith, C. A.; Taylor, J. S.; Sancar, A. *J. Biol. Chem*. **1994**, 269, 8535.

10 Zhao, X.; Liu, J.; Hsu, D. S.; Zhao, S.; Taylor, J. S.; Sancar, A. *J. Biol. Chem*. **1997**, 272, 32850.

11 Hitomi, K.; Nakamura, H.; Kim, S.-T.; Mizukoshi, T.; Ishikawa, T.; Iwai, S.; Todo, T. *J. Biol. Chem*. **2001**, 13, 10103.

12 Joseph, A.; Prakash, G.; Falvey, D. E. *J. Am. Chem. Soc*. **2000**, 122, 11219.

13 Asgatay, S.; Petermann, C.; Harakat, D.; Guillaume, D.; Taylor, J. S.; Clivio, P. J. *J. Am. Chem. Soc*. **2008**, 130, 12618.

14 Borg, O. A.; Eriksson, L. A.; Durbeej, B. *J. Phys. Chem*. **2007**, 111, 2351.

15 Heelis, P. F.; Liu, S. *J. Am. Chem. Soc*. **1997**, 119, 2936.

16 Maul, M. J.; Barends, T. R. M.; Glas, F. A.; Cryle, M. J.; Domratcheva, T.; Schneider, S.; Schlichting, I.; Carell, T. *Angew. Chem. Int. Ed*. **2008**, 47, 10076.

17 Domratcheva, T.; Schlichting, I. *J. Am. Chem. Soc*. **2009**, 131, 17793.

18 Yamamoto, J.; Tanaka, Y.; Iwai, S. *Org. Biomol. Chem*. **2009**, 7, 161.

19 Sadeghian, K.; Bocola, M.; Merz, T., Schütz, M. *J. Am. Chem. Soc*. **2010**, 132, 16285.

20 Li, J.; Liu, Z.; Tan, C.; Guo, X.; Wang, L.; Sancar, A.; Zhong, D. *Nature* **2010**, 466, 887.

21 Harbach, P. H. P.; Borowka, J.; Bohnwagner, M.-V.; Dreuw, A. J. *Phys. Chem. Lett*. **2010**, 1, 2556.

22 Humphrey, W.; Dalke, A; Schulten, K. *J. Mol. Graphics* **1996**, 14, 33.

23 Schleicher, E.; Hitomi, K.; Kay, C. W. M.; Getzoff, E. D.; Todo, T.; Weber, S. *J. Biol. Chem*. **2007**, 282, 4738.

24 Glas, A. F.; Schneider, S.; Maul, M.; Hennecke, U.; Carell, T. *Chem.-Eur. J*. **2009**, 15, 10387.

25 Glas, A. F.; Kaya, E.; Schneider, S.; Heil, K.; Fazio, D.; Maul, M.; Carell, T. *J. Am. Chem. Soc*. **2010**, 132, 3254.

26 Wang, J. M.; Cieplak, P; Kollman, P. A. *J. Comput. Chem*. **2000**, 21, 1049.

27 Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham III, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J*. **2007**, 92, 3817.

28 Svozil, D.; Šponer, J. E.; Marchan, I.; Pérez, A.; Cheatham, T. E.; Forti, F.; Luque, F. J.; Orozco, M.; Šponer, J. *J. Phys. Chem. B* **2008**, 112, 8188.

29 Hornak, V; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, 65, 712.

30 Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, 25, 247.

31 Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem*. **2004**, 25, 1157.

32 Case, D.A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. E.; Wang, J.; Duke, R. J.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R.C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, C. S.; Kollman, P. A. *AMBER 9*, University of California, San Francisco, CA, **2006**.

33 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W. ; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision D.02; Gaussian, Inc.: Wallingford, CT, **2004**.

34 Bayly, C.I; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem*. **1993**, 97, 10269.

35 Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. *J. Comput. Chem*. **1995**, 16, 1357.

36 See, for example: Beierlein, F. R.; Kneale, G. G.; Clark, T. *Biophys. J*. **2011**, 101, 1130.

37 Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem*. **2003**, 24, 1999.

38 Zgarbová, M.; Otyepka, M.; Sponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. *J. Chem. Theory Comput.***2011**, 7, 2886.

39 Zipse, H.; Artin, E.; Wnuk, S.; Lohman, G. J. S.; Martino, D.; Griffin, R.G.; Kacprzak, S.; Kaupp, M.; Hoffman, B.M.; Bennati, M.; Stubbe, J.; Lees, N. *J. Am. Chem. Soc*. **2009**, 131, 200.

40 Vreven, T.; Morokuma, K.; Farkas, Ö.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem*. **2003**, 24, 760.

41 Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, Jr., J.A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput*. **2006**, 2, 815.

42 Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins* **2005**, 61, 704.

43 Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins* **2008**, 73, 765.

44 *H++: web-based computational prediction of protonation states and pK of ionizable residues in macromolecules*: http://biophysics.cs.vt.edu/H++/index.php

45 Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic Acids Res*. **2005**, 33, W368.

46 Bashford, D.; Karplus, M. *Biochemistry* **1990**, 29, 10219.

47 Anandakrishnan, R.; Onufriev, A. *J. Comp. Biol*. **2008**, 15, 165.

48 Word, M. J.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. *J. Mol. Biol*. **1999**,285, 1735.

49 Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. USA* **2001**, 98, 10037.

50 Bank, R.; Holst, M. *SIAM Rev*. **2003**, 45, 291.

51 Holst, M. *Adv. Comput.Math*. **2001**, 15, 139.

52 Nielsen, J. E.; Vriend, G. *Proteins* **2001**, 43, 403.

53 Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. *Nucleic Acids Res*. **2004**, 32, W665.

54 Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys*. 1984, 81, 3684.

55 Pauwels, E.; Declerck, R.; Verstraelen, T.; De Sterck, B.; Kay, C. W. M.; Van Speybroeck, V.; Waroquier, M. *J. Phys. Chem. B* **2010**, 114, 16655.

56 Kondoh, M.; Hitomi, K.; Yamamoto, J.; Todo, T.; Iwai, S.; Getzoff, E. D.; Terazima, M. *J. Am. Chem. Soc*. **2011**, 133, 2183.

5. SUMMARY

In this thesis, different computational methods were applied to selected radical enzyme systems in order to describe their structure and functionality. Specifically, the systems under investigations were pyruvate formate-lyase (PFL), a glycyl radical enzyme that plays a key role in glucose metabolism in many microorganisms under anaerobic conditions, and (6-4) photolyase, a light-dependent enzyme capable of repairing the DNA lesions formed upon exposure to UV radiation.

PFL uses a two-step radical mechanism to catalyse the reversible CoA-dependent conversion of pyruvate into formate and acetyl-CoA. The radical is introduced into PFL by abstracting the hydrogen atom from a glycine residue by the corresponding activating enzyme. The radical is then shuttled to catalytic cysteine residues. The cysteinyl radical undergoes addition reaction with the pyruvate to produce formate and the acetylated enzyme at the cysteine site. The electron density changes that take place as the reaction occurs require the use of QM methods, which are only applicable to small systems (~100 atoms) and not entire enzymes. One of the strategies devised to address this issue is building of tractable models that represent the active site of chosen enzyme. In this study, we used three different small model systems and highly accurate QM techniques to describe the mechanism of PFL and demonstrate the influence of model design on the calculated potential energy surfaces. An additional goal was to provide a proper description of salt-bridge interactions in small-model treatments and a sound recommendation for truncation of arginine-bound carboxylate motifs. In this respect, the usage of neutral model with proton in *syn* position is the optimal choice.

An alternative to the small-model approach of enzyme catalyzed reactions involves a multiscale approach, in which the system is partitioned into regions that can be treated with methods of varying sophistication. Hence, it is possible to define a subsystem that contains the active site and use expensive QM methods to describe chemical reactions, while the rest of the system is treated with a lower level of theory. In the case of enzyme catalyzed reactions, it is common to describe the non-reacting part of the system with classical molecular mechanics (MM), which employs an empirical force fields to describe the interactions between atoms. The hybrid or multi-scale combination of these two approaches is known as QM/MM and it allows the incorporation of the effect of the environment on the chemical reaction, which is of

crucial importance in catalysis. To verify the reliability of this approach before applying it to large systems, we performed a validation study using small models relevant to the PFL-catalyzed reactions. Specifically, we carefully compared the QM/MM results with those obtained with pure QM calculations and the agreement between two approaches was impressive, confirming the accuracy of ONIOM[G3(MP2)RAD] method and justifying its use in further research involving large enzyme models. In addition, alternative mechanisms of PFL were investigated, as well as the inhibition that is triggered by the binding of oxamate, an isosteric and chemically inert analogue of substrate pyruvate, to the active site. A high energy pathway of the reaction involving oxamate was found to be a likely explanation for the inhibition, and alternative mechanisms of action for PFL were discarded on the same basis.

In parallel with the small model studies, a series of molecular dynamics simulations was carried out to investigate potential entrance pathways of CoA from the protein surface to the buried active site, in order for the second half-reaction to take place. This technique uses the aforementioned empirical MM force fields and propagates the system forward in time using classical equations of motion. Trajectories obtained in this way can provide statistical data to compute various structural, dynamic and thermodynamic properties. We used extensive MD simulations in combination with variety of state-of-the-art techniques for free energy calculation to identify crucial conformational changes that could lead to channel opening. To examine the potential coupling of the channel opening with the chemical reaction in the active site, monomeric and dimeric models of the PFL system before and after pyruvate cleaveage were used. The free dynamics simulations resulted with the channel opening in a dimeric model representing the system after the first half-reaction. The performed free energy calculations provide additional support to the statement that the approach of CoA to the active site is mitigated by the pyruvate cleavage. A successful acetylation of PFL at Cys418 might serve as a signal that the system is ready for the subsequent reaction with the second substrate, CoA, and set the conformational changes into motion. However, additional studies are required to fully resolve this issue.

The free energy methods applied on the PFL system included umbrella sampling technique on the monomeric systems and a series of steered MD simulations on all the models used. The free energy estimators used to extract PMF from the data collected during the umbrella sampling (WHAM, UI, MBAR) all give almost identical result. The results obtained with the steered MD simulations exhibit a high dependency on the initial conformations and the switching rate. A large number of the pulling experiments is required to collect sufficient

amount of data, but it is also important to define an appropriate switching speed that would allow system to sample the optimal pathways.

The other system investigated in this thesis is (6-4) photolyase, a light-dependent enzyme capable of repairing the pyrimidine-pyrimidone photoproduct or (6-4) lesions in DNA. (6-4) lesions are formed by dimerization of two adjacent pyrimidine bases in DNA upon the exposure to UV light. The repair requires reduced $FADH^-$ and visible light for catalysis and the mechanism of (6-4) photolyase is still under debate. What is clear is that two highly conserved active site histidines (His365 and His369) have been identified as key residues in catalysis, most likely acting as an acid-base pair. Therefore, the unravelling of (6-4) photolyase mechanism is tightly connected to the knowledge of the correct protonation states of these two histidines. It is even more important when it comes to molecular modelling, where protonation states are generally predefined in the given model. Due to the fact that neutral histidine can exist in two tautomeric forms, or it can be fully protonated, there are three possible states for each histidine, resulting with nine combinations in which the protons can be distributed among two histidine residues.

On the basis of an EPR/ENDOR study it has been widely accepted that His365 is protonated (acid), while His369 is neutral (base). Our own investigations revealed that the complex hydrogen bonding network in the active site could allow other protonation combinations, none of which could be easily discarded on the basis of structural or intuitive criteria. To systematically address this issue, we used a combination of various computational tools to explore all nine possibilities. We began with QM/MM calculations of the EPR coupling constants. While we obtained good agreement with the experimental values, we found that the spectroscopic coupling constants were not overly sensitive to the protonation states of the histidines and thus cannot be used reliably to assign them. Subsequently, we evaluated the pK_a values of the histidine residues, using a continuum electrostatic solvation model (Poisson-Boltzmann equation), and performed MD simulations to examine the structural stability of different protonation states under varying conditions. The combination of the complementary approaches was found to strongly support the statement that both histidines are most likely to be neutral, with a higher probability of protonation attributed to His369.

All the results presented in this thesis provide a powerful illustration to the versatility of issues that can be addressed by molecular modelling. Of course, the limitations exist in every approach, but the example of (6-4) photolyase shows nicely how different techniques can act in a complementary fashion, each providing a piece to successfully solve the puzzle.

Considering the fact that computers still follow Moore's law and become more powerful every year, together with the amazing advances in computational techniques that have been made by employing GPU units, it is reasonable to assume that molecular modelling will become irreplaceable part of many scientific endeavours. After all, it is a rather young discipline – and its best is yet to come.

6. List of Publications

1. Brkljača, Zlatko; Čondić-Jurkić, K.; Smith, A.-S.; Smith, D. M.: Calculation of the CD spectrum of a peptide from its conformational phase space: The case of Met-enkephalin and its unnatural analogue, *J. Chem. Theory Comput.,* **2012**, 1694-1705.

2. Čondić-Jurkić, K. Smith, A.-S.; Zipse, H.; Smith, D. M.: The Protonation States of the Active-Site Histidines in (6-4) Photolyase, *J. Chem. Theory Comput.* **2012**, 8, 1078-1091.

3. Čondić-Jurkić, K.; Zipse, H.; Smith, D. M.: A Compound QM/MM Procedure: Comparative Performance on a Pyruvate Formate-Lyase Model System, *J. Comput. Chem.* **2010**, 31 (5), 1024-1035.

4. Čondić-Jurkić, K.; Zipse, H.; Perchyonok, V.T.; Smith, D. M.: On the modeling of arginine-bound carboxylates: A case study with Pyruvate Formate-Lyase, *J. Comput. Chem.* **2008**, 29 (14), 2425–2433.