Karlsruhe Institute of Technology

Improvements of the Density-Functional Tight-Binding Parameters and Their Application in Histidine Kinase

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)
von der KIT-Fakultät für Chemie und Biowissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

Mayukh Kansari, M.Sc.

aus Kalkutta (Indien)

Erster Gutachter: Prof. Dr. Marcus Elstner
Zweiter Gutachter: Prof. Dr. Alexander Schug
Tag der mündlichen Prüfung: 18. Juli 2022

Abstract

Two component systems (TCS) are one of the main signal transduction pathways in bacteria. Through these systems bacterias sense their environment and regulate their cellular activities. Histidine kinases (HK) are a vital component of TCS. Though TCS are employed by some eukaryotes, they are especially missing from the animal kingdom. For this particular reason and because of their importance in the bacterial life cycle, histidine kinases are a promising target for developing drugs to combat bacterial activities. They are also structurally conserved and exhibit some common inter-molecular events in every TCS, which are autokinase, phosphotransfer, and phosphatase activities. These events are highly fascinating for bio-chemists to understand bacterial activities.

The main goal of this thesis is to analyze histidine-phosphorylation, a vital event in TCS. We have studied the chemical steps of auto-phosphorylation in an extensive QM/MM hybrid enhanced sampling simulation and unveiled the detailed mechanism. The subsequent auto-phosphorylation inside the DHp domain proceeds via a penta-coordinated transition state to a protonated phosphohistidine intermediate. Then, this intermediate is consequently deprotonated by a suitable nearby base. The reaction energetics are controlled by the final proton acceptor and presence of a magnesium cation.

We re-parameterised the DFTB3 parameters for the phosphorus-nitrogen interaction and benchmarked it on a cancer drug hydrolysis reaction. Afterwards, we applied these parameters to our main goal, studying the reaction in histidine kinase in a QM/MM simulation. We further used an artificial neural network on the same drug hydrolysis reaction to improve DFTB energies quantitatively as an alternative method.

In addition, we also re-parameterised the sulfur-sulfur repulsive parameters to improve the disulfide-thiol exchange reaction and reproduced the DFT(B3LYP) level potential energy using DFTB. This was then applied to the disulfide-thiol exchange reaction in QM/MM simulations of proteins.

Zusammenfassung

Zweikomponentensysteme (Two Component Systems, TCS) sind einer der wichtigsten Signaltransduktionswege in Bakterien. Durch dieses System nehmen Bakterien ihre Umgebung wahr und regulieren ihre zellulären Aktivitäten. Histidinkinasen (HK) sind ein wichtiger Bestandteil von TCS. Obwohl TCS von einigen Eukaryoten genutzt werden, fehlen sie insbesondere im Tierreich. Aus diesem Grund und wegen ihrer Bedeutung für den bakteriellen Lebenszyklus sind Histidinkinasen ein vielversprechendes Ziel für die Entwicklung von Medikamenten zur Bekämpfung bakterieller Aktivitäten. Sie sind außerdem strukturell konserviert und weisen einige gemeinsame intermolekulare Funktionalitäten in jedem TCS auf, nämlich Autokinase-, Phosphotransfer- und Phosphatase-Aktivitäten. Diese Vorgänge sind für Biochemiker äußerst faszinierend, um bakterielle Aktivitäten zu verstehen.

Das Hauptziel dieser Arbeit ist die Analyse der Histidin-Phosphorylierung, einem wichtigen Vorgang in TCS. Wir haben die chemischen Schritte der Autophosphorylierung in einer umfassenden QM/MM Hybrid Enhanced Sampling Simulation untersucht und den detaillierten Mechanismus aufgedeckt. Die darauffolgende Autophosphorylierung innerhalb der DHp-Domäne verläuft über einen pentakoordinierten Übergangszustand zu einem protonierten Phosphohistidin-Intermediat. Dieses wird anschließend durch eine geeignete benachbarte Base deprotoniert. Die Energetik der Reaktion wird durch den endgültigen Protonenakzeptor und die Anwesenheit eines Magnesiumkations gesteuert.

Wir haben die DFTB3-Parameter für die Phosphor-Stickstoff-Wechselwirkung neu parametrisiert und an einer Krebsmedikamenten-Hydrolysereaktion gemessen. Anschließend wendeten wir diese Parameter auf unser Hauptziel, die Untersuchung der Reaktion in der Histidin-Kinase in einer QM/MM-Simulation, an. Des Weiteren haben wir ein künstliches neuronales Netz auf dieselbe Arzneimittelhydrolysereaktion angewendet, um die DFTB-Energien als alternative Methode quantitativ zu verbessern.

Darüber hinaus haben wir auch die Schwefel-Schwefel-Abstoßungsparameter neu parametrisiert, um die Disulfid-Thiol-Austauschreaktion zu verbessern, und die potenzielle Energie des DFT(B3LYP)-Niveaus mit DFTB reproduziert. Dies wurde anschließend auf die Disulfid-Thiol-Austauschreaktion in QM/MM-Simulationen von Proteinen angewandt.

Acknowledgement

I would like to express my sincere gratitude to my supervisor Tomáš Kubař for his support, scientific advice and contribution in my thesis. I also appreciate his help and support throughout my time in Karlsruhe. I would like to thank Prof. Marcus Elstner for giving me the opportunity to work in his group and encouraging. This thesis will be incomplete without his guidance.

I would like to extend my deepest gratitude to collaborators in KIT-Campus North, Fathia Idiris and Alexander Schug for their scientific input and contribution in P5 project. I also would like to thank my collaborators in Boston Prof. Qiang Cui, Tanmoy Pal and Darren Demapan for important scientific discussions and their inputs.

I would like to thank all the members of TCB group for their support, especially Denis Maag for his contribution in chapter 11 and Lena Eichinger for helping me in trans Histidine Kinase. I would like to thank all members in GRK 2450 for their support.

I am extremely thankful to Graduiertenkolleg 2450 (GRK 2450) for research funding and financial support.

Finally, I also want to thank my friends in Bremen especially Abhishek Acharya and in Bochum Chandan Kumar Das for valuable scientific discussions.

Contents

List of Figures

List of Tables

Part I.

Introduction

1. Introduction

Histidine Kinases are part of two component systems [31], which are involved in bacterial signal transduction [30]. Signal transduction is an information-processing pathway by which a chemical or physical signal is transmitted through a cell through a series of molecular events, most commonly protein phosphorylation, assisted by various kinases, which ultimately results in a cellular response [6, 172].

The two-component system (TCS) is one of the most abundant mechanisms used by bacteria to adapt to their environment. They are involved in regulation of responses to a variety of environmental factors or cellular signals [174]. The individual components are consists of several parts, periplasmic domain sits in the transmembrane location receives the signal, passes through to histidine kinase (HK) and the response regulator (RR) protein that coordinates the response, most commonly by acting as a transcription factor (see Fig. 1.1). The two proteins communicate via histidine to aspartate phosphoryl-group transfer. Based on domain architectures, evolutionary origin and activities there are numerous variations of TCS [94, 189].

Due to their prevalence and the associated wealth of genomic data, TCS are also a common target of bioinformatics studies to, e.g., investigate TCS complex formation [146], predict [37] or investigate [127] conformational transitions, or redesign protein signalling [31]. Other extensive studies and reviews highlight the range of TCSs and their activities [85, 166, 23, 133, 41].

Most of these signalling pathways proceeds through protein phosphorylation, It is involved in all of signal transduction system assisted by protein kinases, phosphorylate themselves or other protein substrates at specific Ser, Thr, Tyr, His residues, thereby regulating cellular activities[183]. While other kinases were known for long time histidine kinase activity and histidine autophosphorylation were discovered 1980s [163].

Many HKs are bifunctional, acting as both the kinase and phosphatase for their RR; the ratio of kinase to phosphatase activity, and thus the phosphorylation state of the RR, is controlled by the input [162, 136, 11, 79, 107, 89]. For an example, two Bacteriophytochromes DrBphP and Agp1 both possess HK effector domains with Agp1 acting as a histidine kinase whereas DrBphP as a light-activated phosphatase [119].

Signalling networks in eukaryotes often exhibits extensive “crosstalk” with individual kinases acting on large numbers of targets for example the kinase Cdk1, has hundreds of substrates in yeast [173, 73]. Bacterial TCS networks show a remarkably different mechanism, bacterial HKs usually act on a single target [124, 53, 155, 153, 101, 52]. Intensive experimental studies over the past 10 years have showed that bio-chemical and biophysical basis for this lack of promiscuity. In general, HKs demonstrate a strong “kinetic preference” for their cognate substrates, preferentially phosphorylating them on short timescales. [156, 68, 154, 27, 52].

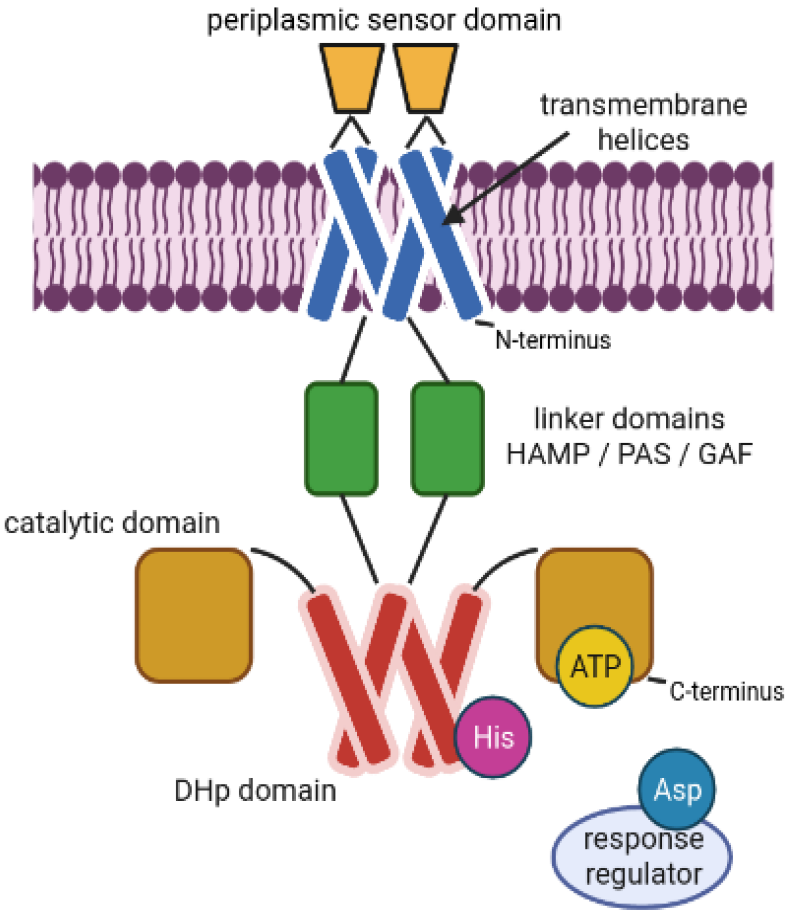

Figure 1.1.: Schematic diagram of a typical two component system in bacteria featuring domains for signal recognition, transmission, and catalysis. A stimulus is first detected at the periplasmic sensor domain (yellow) and this signal transmits along the transmembrane helices (purple) and the linker domains (green) before reaching the catalytic core at the C-terminus. The catalytic core comprises the dimerization and histidine phosphotransfer (DHp, red) and catalytic ATP-binding (CA, yellow) domains. Signal detection results in a phosphoryl transfer reaction from ATP in the CA domain to a conserved histidine in the DHp domain. This phosphoryl group is then transferred to an aspartic acid in the response regulator (blue) protein, resulting in an appropriate cellular response (Image created using BioRender.com)

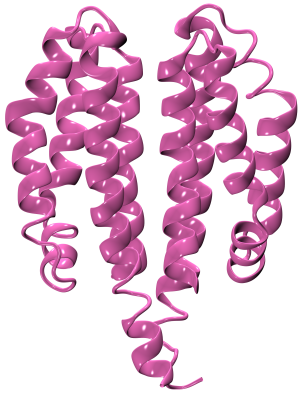

Figure 1.2.: Sensor domain (pdb id:3EZH) of histidine kinases illustrating two common structural folds

While TCS are employed by some eukaryotes, they are notably absent from the animal kingdom. That, together with their importance to bacteria makes these enzymes promising targets for developing novel compounds [16] that selectively inhibit the growth of bacteria or suppress virulence. For instance, waldiomycin [44, 137, 47, 88], an angucycline antibiotic, inhibits the HK activity of WalK [83, 126] in *Staphylococcus aureus*, a human pathogen responsible for a variety of acute and chronic diseases [182, 171, 139]. The molecular signal of this system is still unknown but emanates from the bacterial cell wall [20]. In general, the WalRK system has garnered significant experimental attention since it is conserved across Gram-positive bacteria of the order Firmicutes where it has been shown to be essential for viability in a variety of different species of bacteria.

There are a vast variety of TCS systems available in bacterial genome and even highly similar TCS do not necessarily have the same function. Based on detection of different stimuli at their sensor domain, there are several types of TCS, for example DesK detects cold temperature [2], EnvZ senses and responds to osmotic stress [179], and CheA mediates bacterial chemotaxis [67]. Even though sensor domains of these systems are functionally different they possess structural similarity, which suggests that the signalling mechanism is conserved [32]. These common structural similarities include two parallel alpha helices.

After the sensor domain, next conserved feature in TCS is one or more linker domains involves in transmitting the N-terminal signal to the catalytic domains at the C-terminal. The most widely studied of these linker domains is known as HAMP (histidine kinase, adenylyl cyclase, methyl-accepting chemotaxis protein, and phosphatase) and is found in 30% of HK [5, 49, 189]. Other known linker domains include PAS (Per-Arnt-Sim) [116], and GAF (GMP-specific phosphodiesterases, adenylyl cyclases and FhlA). Structural overview of these linker domains are shown in fig 1.3

After the linker domain, the part of Histidine Kinase core starts. The kinase core consists of the dimerisation and histidine phosphotransfer (DHp) and catalytic ATP–binding (CA) domains (see Fig. 1.4). These two domains are highly conserved across all HKs [43]. Similar

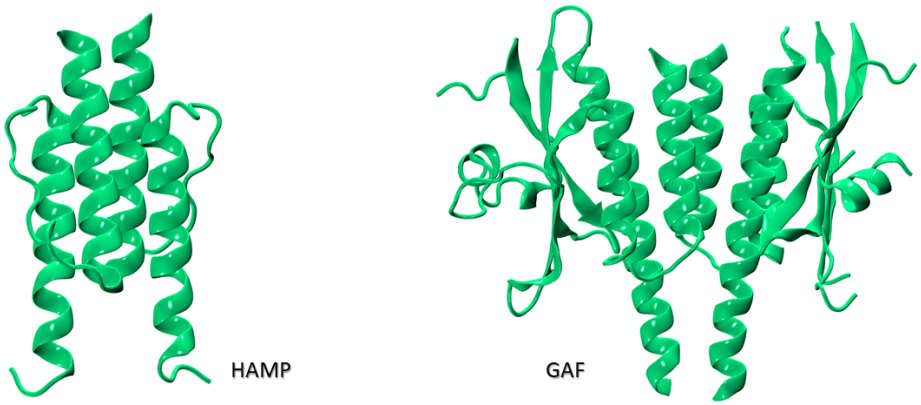

Figure 1.3.: Structures of HAMP (PDB ID: 2L7H) and PAS (PDB ID: 4I5S) domains

to the HAMP domain, the DHp domain is typically homodimeric with each protomer comprising two α-helices that form an antiparallel coiled-coil. The CA domain, on the other hand, has an α/β sandwich fold made up of a five-stranded β-sheet flanked by three α-helices. The nucleotide binds between two α-helices and is held by a highly mobile loop known as the ATP lid. Well conserved nucleotide-binding sequence motifs known as the N, G1, F, and G2 boxes comprise the binding site [91, 159, 128, 112].

HK exhibits kinase activity through a interplay of conformational change and reaction in a cyclic manner 1.5). Once signal detected at the extra-cellular part in the sensor domain, The signal transmits through a series of allosteric changes within the transmembrane domains to the conserved kinase core. It triggers a change in the kinase conformation to a typical asymmetric structure in such a way that one of the two subunits of the homodimer (Dhp and CA) comes closer to each other (Kinase Active conformation) while the other remains inactive (cf. Fig. 1.5). In the kinase inactive conformation, ATP can enter to the CA domain and the binding site of the DHp domain. Here, the gamma-phosphoryl group of the bound ATP of one CA is positioned in close proximity to a specific conserved phosphorylatable histidine of DHp.

Through this activation, Kinase triggers a series of reactions known as biochemical cascade, a very common feature of for Signal transduction pathways[147, 103, 21, 18, 40]. A very popular example will be mitogen-activated protein (MAP) kinase and cyclic nucleotide cascades in mammal signal transduction systems.

In WALK TCS it starts with acceptance of stimuli, which brings the driving force (energy) of the whole cascade. This begins with association of ATP with histidine kinase core, ATP known as highly energetic compound initiates series of reaction, first with the auto-phosphorylation to conserved histidine, forming unstable phosphohistidine as a final product. Phosphohistidine, still carries residual energy from ATP, serves as only intermediate in the cascade and initiates the next reaction to response regulator (RR), which is phosphoryl transfer reaction from histidine to aspartate residue of RR. Thus this cascade continues until it forms a sable product. This particular cascade reaction is known as His-Asp phosphorelay. RRs typically have autophosphatase activity that limits the

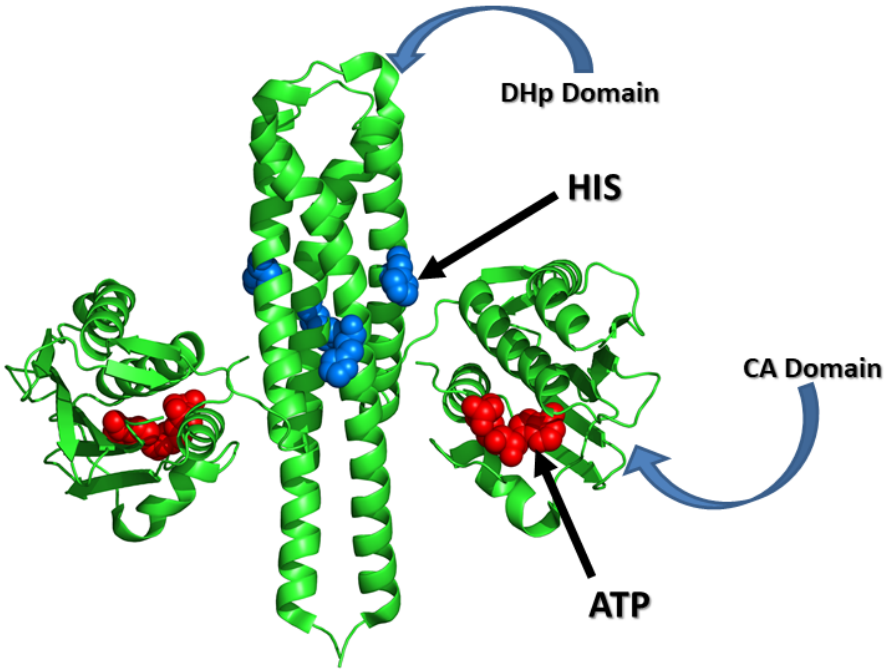

Figure 1.4.: A typical structure of histidine kinase (pdb id: 2c2a) featuring conserved histidine phosphotransfer (DHp) and catalytic ATP-binding (CA) domains, ATP is shown in red spheres, Histidine is shown blue spheres.

lifetime of the phosphorylated state, yielding half-lives in the range of seconds to hours. Both phosphotransfer and phosphatase activities of HKs are found to be regulated either directly by stimuli or indirectly through interaction with auxiliary proteins (in the case of cytoplasmic HKs). There are over 100 examples of such His-Asp phosphorelay systems in bacteria and 17 of them has been identified in E. coli [176].

Two different auto-phosphorylation mechanisms are exhibited in individual HKs, cis- and trans-phosphorylation 1.6. In cis-phosphorylation, the ATP from the CA domain phosphorylates its own DHp domain within the homodimer, while in trans-phosphorylation the DHp domain on the other monomer within the homodimer is phosphorylated [7]. It appears that the difference in phosphorylation mechanism is merely structural, based on a left handed versus right handed orientation of the dimeric four-helix bundle that forms the DHp domain. As soon as the histidine is phosphorylated, transfer of this phosphoryl group to an aspartate of a bound RR for communication between the two proteins is possible. Mechanism of cis-phosphorylation in WALK histidine kinase has been studied and discussed in **chapter 9** in detail.

Observing enzymatic reactions in its native biological environment (in vivo) are extreamly difficult. The only way to study these reactions is possible through computer simulations. QM/MM methods are very powerful tool to study biochemical reactions [108, 106, 160, 148, 161]. There are many previous successful studies has been carried out using QM/MM method to investigate phosphoryl transfer reaction in different enzymes and proteins. [98, 76, 142, 158, 117, 144]

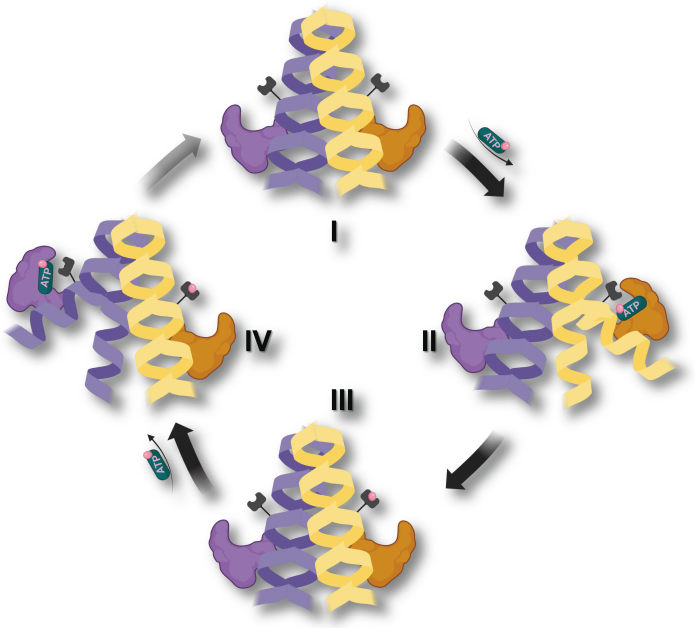

Figure 1.5.: Schematic diagram of cyclic interplay of Inactive (symmetrical) to active (asymmetrical) conformations of histidine kinase, Red dot on green ATP represents γ phosphate (image created using biorender.com)

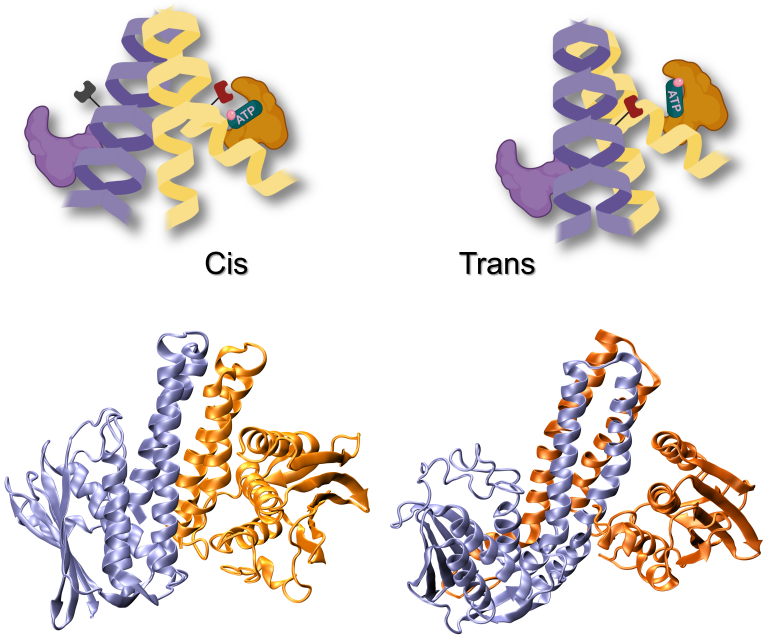

Figure 1.6.: Typical representation of Cis and Trans Phosphorylation in Histidine Kinase. Cis structure is based on pdb id: 4u7o, Trans structure is based on pdb id: 5lfk. (image is created using biorender.com)

The goal of this thesis is to study autophosphorylation reaction in Histidine Kinase applying QM/MM simulation using DFTB3 as QM level of theory. From the initial observation of the study, it turns out that product minimum and penta-valent phosphate transition state is not detectable in our simulation. After looking it closer it appeared that parameters (3OB) of DFTB3 is not good enough to study the reaction. So for part of achieve our goal, as part of the thesis we also re-parametrised required parameters of DFTB3 and applied it to study autophosphorylation of histidine kinase.

With DFTB3, the 3OB set of parameters are most commonly used for organic and biological systems. However, a few transferability problems were found for some complex chemical reactions, which led to incorrect reaction energetics and structures. Histidine autophosphorylation is one of those cases, which we studied in this thesis. Another such case is found in disulfide-thiol exchange reaction, one of the important steps in the folding process of many proteins that has to form structural disulfide bonds. In such nontrivial cases Semi-empirical (SE) methods come up with special reaction parameter (SRP) [122] as a quick solution. In this thesis we re-parametrised both phosphorus-nitrogen, sulfur-sulfur interaction to make new SRPs in DFTB3 specifically for studying these mentioned reaction.

In **chapter 8** and **chapter 10**, improvements of DFTB3 energies are discussed in detail. In **chapter 8** repulsive potential for P-N interaction has been reparameterised and benchmarked on different molecules and as an example of the benchmark reaction we have simulated hydrolysis of a Cancer drug called TEPA and compared with previous studies. Also these parameters are used in chapter 9 for simulating the reaction in Histidine Kinase.

In **chapter 10** we used Neural Network to improve DFTB3 energies. Specifically, we took the example of durg hydrolysis reaction of TEPA and compared energies of re-parametrised DFTB3 and Neural-Network corrected energies of the same reaction.

In **chapter 11** the accuracy of the density-functional tight binding method (DFTB3) regarding thiol-disulfide exchange reactions is evaluated. In DFTB, the S–S bonds in the transition states are too long and the transition state is a local minimum on the free energy surface instead of a saddle point. We have corrected these errors by reparameterising the S–S repulsive potentials and tested them in QMMM Simulation of disulfide-thiol exchange reaction.

Part II.

Methods

2. Overview

The field "computational chemistry" has got its roots from theoretical chemistry, defined as the mathematical description of chemistry. Computational chemistry takes the role when a mathematical method is sufficiently well developed that it can be automated for implementation on a computer. Though it is still considered as approximation of the real chemical properties, it can give useful insight into chemistry and can potentially solve chemical problems. Understanding mechanism of a reaction and predicting a new mechanism is domain where computational chemistry is highly regarded. Nowadays It is vastly used in pharmaceutical industry, for predicting potential drug candidate, predicting protein structures. From Biochemistry to Material Chemistry this field has got its branches spread in improving predicting power of chemistry[50].

Computer simulations act as a bridge between microscopic length and time scales and the macroscopic world of the laboratory [3, 55]. It provides a guess of the interactions between molecules and obtain an 'exact' predictions of bulk properties. The predictions are 'exact' in the sense that they can be made as accurate as we like, subject to the limitations imposed by our computational cost 2.1. Simulations act as a bridge in different sense: between theory and experiment. We may test a theory by conducting a simulation using the same model. We may test the model by comparing with experimental results also we can carry out simulations on the computer that are difficult or impossible in the laboratory (for example, working at extremes of temperature or pressure or in vivo biochemical reactions).

Ultimately, we may want to make direct comparisons with experimental measurements made on specific materials, in which case a good model of molecular interactions is essential. For example, the aim of so-called ab-initio molecular dynamics is to reduce the amount of fitting and guessing in this process to a optimum level. On the other side, this is also interesting that we may want to differentiate between good and better theories. When it comes to aims of this kind, it is not necessary to have a perfectly realistic molecular model; one that contains the essential physics may be quite suitable.

All the methods used in computational chemistry can be divided based on their accuracy and computational speed. Ab-initio and DFT are most accurate methods but larger systems and processes greater than pico-seconds are difficult to treat in these methods. On the other hand, with MM and CG see 2.1 methods one can study bigger systems and longer timescale processes but accuracy goes down.

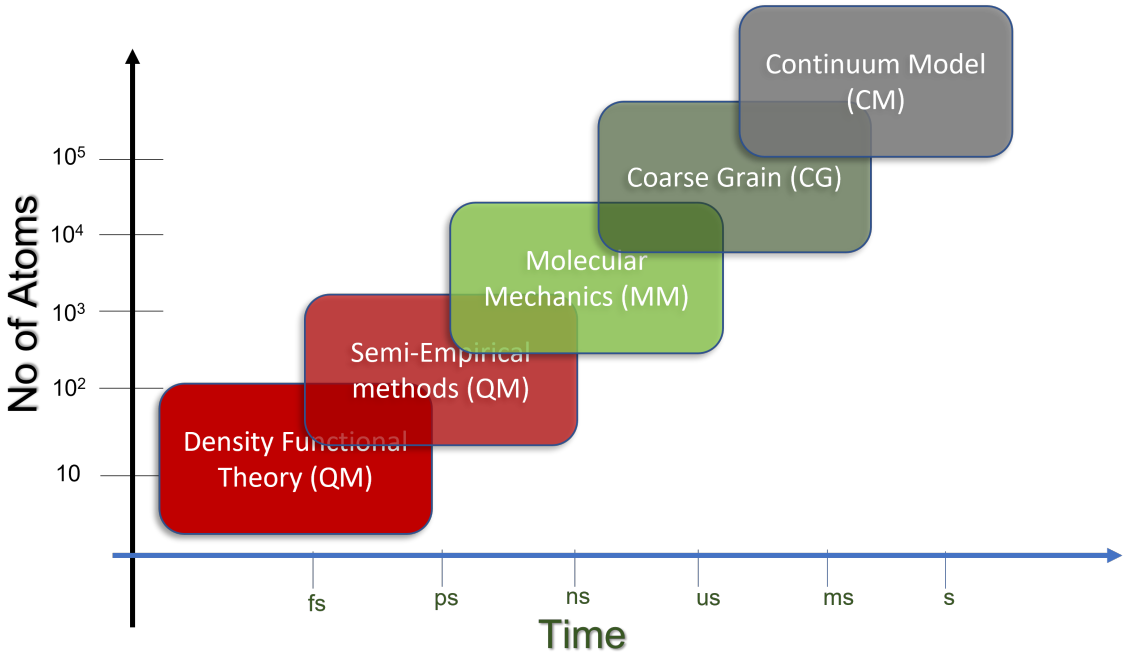

Figure 2.1.: Hierarchy of computational chemistry methods computational chemistry based on simulation time-scale and size of system (number of atoms)

3. Electronic Structure Methods

3.1. Hartree Fock

3.1.1. The Born–Oppenheimer Approximation

After the solution of Hydrogen Atom problem. The next puzzle in the quantum chemistry appears to be solving the Schrodinger wave equation for Many electron system. The Hamiltonian of the system for 'N' no of electrons is the following (equation is in the atomic units).

$$H_{full} = \sum_{i=1}^{N} -\frac{1}{2}\nabla_i^2 + \sum_{A=1}^{M} -\frac{1}{2}\nabla_A^2 + \sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B=1}^{M}\frac{Z_A Z_B}{r_{AB}} \tag{3.1}$$

The first term represents kinetic energy of single electron, second term is the kinetic energy of the nuclei, third term represents the electron-neutron interaction, fourth term is two electron repulsion and the last term is repulsion between the nuclei.

Now due to the fact that the Nuclei are much heavier than electron, wave functions of nuclei and electron in a molecule can be treated separately. This assumption was first proposed by Max Born and J. Robert Oppenheimer in 1927 known as 'The Born–Oppenheimer Approximation'. This approximation is widely used in quantum chemistry to accelerate computational speed of molecular wave function optimisation and calculation of other properties.

Now under the boundary of Born–Oppenheimer Approximation we can treat the electronic hamiltonian separately. After neglecting the nuclear interactions which is the second and the last term in equation 3.1, we get solely electron dependent hamiltionian in equation 3.2.

$$\begin{aligned} H_{elec} &= \sum_{i=1}^{N}\left(-\frac{1}{2}\nabla_i^2 + \sum_{A=1}^{M}\frac{Z_A}{r_{iA}}\right) + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} \\ &= \sum_{i=1}^{N} h_1(i) + \sum_{i=1}^{N}\sum_{j>i}^{N} h_2(i,j) \end{aligned} \tag{3.2}$$

3.1.2. Wave Function

From the solution of hydrogen atom problem we get set of exponential functions, known as orbitals, which can be considered as wave functions to describe a single electron. An

electronic orbital, denoted as ψ(r) which is a function of position vector describes spatial distribution of an electron and the probability of finding the electron inside a volume dr will be $\int \psi)^2\, dr$.

Hartree Product:
After learning about single electron wave function, we now need to construct the wave function of N number of electrons. Under the Hartree product assumption total wave function of N electron is simply the product of N independent single electron wavefunction given in equation 3.3. Such a wavefunction is termed as Hartree product, where each electron being described in distiguishable orbitals. Hartree product is an independent wave function, which means position of electron-one is independent of electron-two. But in reality they will repel each other [165, 51].

$$\Psi^{HP}(r1, r2, r3...rN) \quad = \quad \psi_1(r1)\psi_2(r2)\psi_3(r3).....\psi_N(rN) \tag{3.3}$$

Slater Determinant:
Hartree product doesn't obey antisymmetry principle but however such wave functions can be antisymmetrised using slater determinant. The purpose is the wave function must have opposite sign when two electrons are interchange their coordinate [Ψ(r1,r2) = − Ψ(r2,r1)], which means if two electron occupy same orbital they must have opposite sign in other meaning Pauli exclusion principle maintained. Slater determinant for N electrons can be written as:

$$\Psi^{SD}(r_1, r_2, ..., r_N) \quad = \quad \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(r_1) & \psi_2(r_1) & \dots & \psi_N(r_1) \\ \psi_1(r_2) & \psi_2(r_2) & \dots & \psi_N(r_2) \\ \vdots & \vdots & & \vdots \\ \psi_1(r_N) & \psi_2(r_N) & \dots & \psi_N(r_N) \end{vmatrix} \tag{3.4}$$

3.1.3. Variation principle

The variation theorem allows us to calculate an upper bound for the system's ground-state energy.

3.1.4. Fock Operator

Now we have ortho-normal wave function using Slater determinant. The total energy of the N electron wave function using Slater determinant using Hamiltonian of equation from 3.2 is given in equation 3.5.

$$E = \Psi^{SD} H \Psi^{SD} \tag{3.5}$$

Now according to variation principle best orbitals are those which minimise the electronic energy. Which means if we apply variation principle on equation 3.5 we get minimum energy E_0 and best possible orbitals Ψ_0

$$E_0 = \Psi_0 H \Psi_0 \tag{3.6}$$

Now we apply variation principle on two electron system and obtained E_0 and Ψ_0 given in equation 3.5 for single electron. After solving equation 3.6 we get the following eigen value equation for ith electron, known as Hartree-Fock equation, where E(i) is the energy of the ith electron.

$$f(i)\psi(r_i) = E(i)\psi(r_i) \tag{3.7}$$

$$f(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + v^{HF}(i) \tag{3.8}$$

Looking at equation 3.8 first two terms represent simply one-electron Schrodinger wave equation for orbital state and s single electron in the field of nuclei. where v^{HF}(i) represents average potential experience by ith electron [109, 157].

$$h(i) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} \tag{3.9}$$

$$v^{HF}(i) = \mathcal{J}(i) + \mathcal{K}(i) \tag{3.10}$$

$$f(i) = h(i) + \mathcal{J}(i) + \mathcal{K}(i) \tag{3.11}$$

Rearranging equation 3.8 using equation 3.9 and 3.10, we get equation 3.11, where h(i) represents one electron operator $\mathcal{J}(i)$ and $\mathcal{K}(i)$ are respectively coulomb and exchange operators. The Coulomb operator $\mathcal{J}(i)$ represents the electron-electron interaction while exchange operator $\mathcal{K}(i)$ has no classical interpretation, arising out from the antisymmetric nature of the Slater determinant.

3.1.5. Self Consistent Field

Once we have seen the fock operator, by making simple guess of the orbitals we can calculate average field (v^{HF}) seen by each electron (ith) and solve the eigen value equation of equation 3.7 using equation 3.8 for new set of orbitals. Now with new set of orbitals we can again construct average field (v^{HF}) and a new fock operator (eqn 3.8) repeat the procedure until the average field (v^{HF}) no longer changes to construct new fock operator thus we can say self-consistency is achieved [80]. In practice if the field attains a particular tolerance value set by the user, then we can say SCF is converged.

3.1.6. Basis Set

Electronic wave functions or atomic orbitals are expressed in terms of set of functions in HF and as well as in all electronic structure methods. These functions are called basis

functions[39]. Generally these functions are atom centered gaussian type function or exponential function.

ψ, Molecular orbitals is expressed as a linear combination of n basis functions ϕ_μ. Coefficients c_μ are called molecular orbital expansion coefficients.

$$\psi(i) = \sum_{\mu=1}^{n} c_\mu \phi_\mu \tag{3.12}$$

Slater type orbitals (STOs)

STOs are the direct solution coming from Hydrogen atom problem.STOs are constructed from a radial part describing the radial extend of the orbital and an angular part describing the shape of the orbital.

$$\phi_\mu^{STO} = N r^{n-1} exp(-\zeta r) Y_{lm} \tag{3.13}$$

r is the distance from the origin of the basis function (usually the location of the nucleus), the orbital exponent ζ, n is the principal quantum number. The spherical part Y_{lm} depends on the angular quantum number l and the magnetic quantum number m.

Gaussian type orbitals (GTOs)

GTOs are also constructed from a radial and a spherical part, but the radial part is now a Gaussian type function [81].

$$\phi_\mu^{GTO} = N exp(-\alpha r^2) X^a Y^b Z^c \tag{3.14}$$

The radial part is proportional to $exp(-ar^2)$, α is the gaussian exponent. The normalization factor N serves the same purpose as in STOs. The spherical part is now expressed through the Cartesian coordinates x,y, and z in powers of a, b, and c, respectively.

The accuracy of the HF method improves when the basis set size is increased. Even though with very large basis sets the HF method can only account for 99% due to the mean field approximation where every electron is treated independently, i.e., moving under the influence of an averaged electrostatic field induced by all other electrons. This leads to the neglect of electron correlation which often is very important for the description of chemical phenomena accounting for the remaining 1% of the total energy.

There are a choice of a large variety of basis sets including Poples' basis sets [39], dunning type basis sets [42], Karlsruhe basis sets. [188]

3.2. Density Functional Theory

3.2.1. Electron Density

Density Functional theory [129, 74] is developed over density of electrons rather than wave-function of electron in ab-initio methods (Hartree Fock). Electron density ρ(r) is defined as a multiple integral over the spin coordinates of all electrons and the spatial variables of all electrons. N electrons within a volume element r1 is N times the probability for one particular electron. So at given volume element total number of electron, N can be defined as follows:

$$\rho(r) = N \int \dots \int \Psi(r_1, r_2, r_3 \dots r_N)^2 dr_1 dr_2 \dots dr_N \tag{3.15}$$

$$N = \int \rho(r) dr_1 \tag{3.16}$$

3.2.2. Thomas–Fermi model

This model is the predecessor to density functional theory [167], one of the first few approaches to derive energy of the electronic system using electron density. In this approach energy of an atom is approximated by a kinetic-energy functional combined with the classical expressions for the nucleus–electron and electron–electron interactions (density). This model's accuracy was limited because it did not estimate exchange energy (Pauli's Principle). After that Paul Dirac reformulated this model using exchange-energy functional. However this model remain inaccurate for most chemical systems.

$$T_{\mathrm{W}}[n] = \frac{\hbar^2}{8m} \int \frac{|\nabla n(\mathbf{r})|^2}{n(\mathbf{r})} \, \mathrm{d}^3\mathbf{r} \tag{3.17}$$

3.2.3. Hohenberg-Kohn Theorems

Hohenberg-Kohn theorem [64] layed the founding stone of density functional theory. Theorems formulates energy of a system of electrons moving under the influence of an external potential. Theorem states:

1. The external potential $V_{ext}(r)$ is a unique functional of density ρ (r), i.e., there cannot be two different $V_{ext}(r)$ that yield the same ground state electron density ρ^0 (r). Since $V_{ext}(r)$ fixes the Hamiltonian, the ground state energy (and all other properties) are a functional of the ground state electron density ρ^0 (r)

$$E_0[\rho(r)] = T_S[\rho(r)] + E_{ne}[\rho(r)] + E_{ee}[\rho(r)] \tag{3.18}$$

2. The functional that delivers the ground-state energy of the system, also gives the lowest energy if and only if the input density is the true ground-state density ρ^0 (r). To rephrase it, the energy content of the Hamiltonian reaches its absolute minimum (ground state) when the charge density is that of the ground state.

$$E[\rho(r)] \geq E[\rho_0(r)] \tag{3.19}$$

3.2.4. Kohn-Sham Approach

Kohn and Sham (KS) suggested to split the kinetic energy functional, $E_{ee}[\rho(r)]$ into two parts: (i) the kinetic energy T_S of a reference system of non-interacting electrons with the same electron density as the real system, for which orbitals have to be re-introduced; (ii) the exchange-correlation energy E_{XC} which is the remainder of the exact kinetic energy that has to be treated approximately [92]. The essential idea of exchange correlation is an artifact of Hartree-Fock method. The general DFT energy expression can be re-written as follows:

$$E_{DFT}[\rho] = T_S[\rho] + E_{ne}[\rho] + J[\rho] + E_{xc}[\rho] \tag{3.20}$$

$$E_{ne}[\rho] = \sum_{a}^{N_{nuclei}} \int \frac{Z_a(R_a)\rho(r)}{\left|R_a - r\right|} dr \tag{3.21}$$

$$J[\rho] = \frac{1}{2} \int \int \frac{\rho(r)\rho(r')}{\left|r - r'\right|} dr dr' \tag{3.22}$$

$$T_s[\rho] = \sum_{i=1}^{N} \int d\mathbf{r}\, \psi_i^*(\mathbf{r}) \left(-\frac{\hbar^2}{2m}\nabla^2\right) \psi_i(\mathbf{r}) \tag{3.23}$$

where $T_S[\rho]$ is the kinetic energy of non-interacting electrons, $J[r(r)]$ the Coulomb interaction, $E_{XC}[r(r)]$ the exchange-correlation energy, and $E_{ne}[r(r)]$ the attractive nuclei–electron energy. $E_{XC}[r(r)]$ is the only term without an explicit form and physical interpretation, it depends on an external field exerted by the electric field by the rest of the electron, v_{ext}.

$$E_{xc}[\rho] = \int dr\, v_{\text{ext}}(r)\rho(r) \tag{3.24}$$

$$v_{\text{eff}}(r) = v_{\text{ext}}(r) + e^2 \int \frac{\rho(r')}{|r - r'|}\, dr' + \frac{\delta E_{\text{xc}}[\rho]}{\delta\rho(r)} \tag{3.25}$$

$$v_{\text{xc}}(r) \equiv \frac{\delta E_{\text{xc}}[\rho]}{\delta\rho(r)} \tag{3.26}$$

If both exchange-correlation terms were known, the Kohn-Sham approach would provide the exact energy. However, it is not possible, they have to be approximated in different ways, which is the key development in DFT. Many functionals have been proposed, such as the local density approximation (LDA), the generalized gradient approximation (GGA) and hybrid functionals.

Local Density Approximation

Local-density approximations (LDA) [130] are a class of approximations to the exchange–correlation (XC) energy functional in density functional theory (DFT) only depends on the value of the electronic density at each point in space. E_{XC} is the exchange-correlation energy per

particle that can be split into exchange and correlation contributions. The exchange part E_X of an electron is considered in a uniform (homogeneous electron gas model) distribution of electrons.

$$E_{xc}^{LDA}[\rho] = \int \rho(r) E_{xc}(\rho(r)) dr \tag{3.27}$$

$$E_{xc} = E_x + E_c \tag{3.28}$$

Generalized Gradient Approximation

Generalized Gradient Approximation (GGA) [100] includes the first derivative of the density $\nabla\rho(r)$ as a variable in v_{ext}.

$$E_{XC}^{GGA} = \int \rho(r) E_{XC}[\rho(r), \nabla\rho(r)] dr \tag{3.29}$$

GGA functionals add correction terms on top of the LDA functional, such as the **B88** functional by Becke [138] or the **LYP** functional [65] by Lee, Yang and Parr. For this, parameters have to be determined by fitting to reference data. Another possibility is to derive the parameters from certain conditions, which has be done for the popular Perdew-Burke-Ernzerhof (**PBE**)[131] functional.

Hybrid Functionals

The GGA functionals are further improved by including HF exchange. In the famous **B3LYP** [65] hybrid functional, the exchange-correlation energy is given as a combination of density-functional exchange and correlation and HF exchange.

$$E_{\mathrm{xc}}^{\mathrm{B3LYP}} = (1-a)E_{\mathrm{x}}^{\mathrm{LSDA}} + aE_{\mathrm{x}}^{\mathrm{HF}} + b \triangle E_{\mathrm{x}}^{\mathrm{B}} + (1-c)E_{\mathrm{c}}^{\mathrm{LSDA}} + cE_{\mathrm{c}}^{\mathrm{LYP}} \tag{3.30}$$

with a = 0.20, b = 0.72 and c = 0.81, which are determined by fitting to experimental data; LSDA stands for Linear Spin Density Approximation.

3.3. Density Functional Tight Binding

Density functional tight binding method (DFTB) [45, 58] is a semi-empirical method, obtained from the DFT total energy functional expanding the exchange correlation energy in a Taylor series. The starting point is the use of a reference density ρ0 , which is calculated from a superposition of precalculated neutral atomic densities. In DFTB only valence electrons are considered using a minimal atomic basis set explicitly; chemical cores are treated in an effective manner via additive two-center potentials. To further reduce computational cost crystal field and three-center integrals are neglected. The remaining two-center Hamilton and overlap matrix elements are precalculated for a dense mesh of interatomic distances in an atomic orbital (AO) basis. The remaining contributions to the total energy are then approximated and thus, no further computational cost arises beyond the dominant step, which is the diagonalisation of the Hamilton matrix. This and the use of the minimal valence basis set leads to huge computational savings (2–3 orders of magnitude) compared to full DFT.

Energy equations of DFTB can be derived from Taylor series expansion of the Kohn–Sham total energy with respect to charge density fluctuations $\Delta\rho = \rho - \rho^0$, here follows:

$$E[\rho^0 + \Delta\rho] = \sum_i n_i \int \psi_i^* \left(-\frac{1}{2}\nabla^2 + V^{ne} + \int \frac{\rho^0{}' + \Delta\rho'}{|r - r'|} + V^{xc}[\rho^0 + \Delta\rho] \right) \psi_i$$
$$-\frac{1}{2} \int \int' \frac{(\rho^0{}' + \Delta\rho)(\rho^0 + \Delta\rho)}{|r - r'|} - \int V^{xc}[\rho^0 + \Delta\rho](\rho^0 + \Delta\rho)$$
$$E^{xc}[\rho^0 + \Delta\rho] + E^{nn} \ldots\ldots\ldots\ldots\ldots$$

$$E[\rho] = E^0[\rho^0] + E^1[\rho^0, \Delta\rho] + E^2[\rho^0, (\Delta\rho)^2] + E^3[\rho^0, (\Delta\rho)^3] + \ldots \tag{3.31}$$

The above equation is a shortened version of the taylor series expansion of DFT Kohn–Sham equation. The series is considered up to third order. First two term represents DFTB1, taking another term will introduce DFTB2 and taking the third order term will give us the equation of DFTB3.

Using of the LCAO representation, using a minimal basis leads to an approximated function for the total energy:

$$E = \sum_i^{\mathrm{MO}} n_i \sum_{A,B}^{\mathrm{atoms}} \sum_{\mu\in A} \sum_{\nu\in B} c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{A,B}^{\mathrm{atoms}} \Delta q_A \Delta q_B \gamma_{AB} + \frac{1}{3} \sum_{A,B}^{\mathrm{atoms}} \Delta q_A^2 \Delta q_B \Gamma_{AB} + \frac{1}{2} \sum_{A,B}^{\mathrm{atoms}} V_{AB}^{rep} \tag{3.32}$$

where n_i is the occupation of the i-th molecular orbital (MO), $c_{\mu i}$ is the expansion coefficient of atomic orbital (AO) μ in MO i, H^0 is the charge-independent Hamiltonian matrix in the AO basis, γ_{AB} is an analytical function that describes the interaction of charge monopoles, and Γ_{AB} is its derivative with respect to Δq_A. V_{AB}^{rep} corresponds to the repulsive energy parameters, an approximation of short-term pairwise interactions [59].

DFTB1

$$E(DFTB1) = E^0[\rho^0] + E^1[\rho^0, \Delta\rho] = \sum_i^{\mathrm{MO}} n_i \sum_{A,B}^{\mathrm{atoms}} \sum_{\mu\in A} \sum_{\nu\in B} c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{A,B}^{\mathrm{atoms}} V^{rep}_{AB}$$

DFTB1 [134] is only consist of first and zeroth order term. The diagonal matrix elements $H^0_{\mu\mu}$ are approximated as the orbital energies ϵ_μ of individual atoms, which are calculated from PBE functional. For off-diagonal elements a two-center approximation is applied, i.e., three and four center integrals are neglected. All matrix elements are precomputed and tabulated for each pair of orbitals and interpolated for a given geometry during a DFTB calculation [93]. DFTB1 performs well for systems with no charge transfer or a complete charge transfer. For systems that are sensitive to charge fluctuations higher order terms have to be included.

DFTB2

$$E(DFTB2) = E^0[\rho^0] + E^1[\rho^0, \Delta\rho] + E^2[\rho^0, (\Delta\rho)^2] = \sum_i^{\mathrm{MO}} n_i \sum_{A,B}^{\mathrm{atoms}} \sum_{\mu\in A} \sum_{\nu\in B} c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{A,B}^{\mathrm{atoms}} \Delta q_A \Delta q_B \gamma_{AB} + \frac{1}{2} \sum_{A,B}^{\mathrm{atoms}} V^{rep}_{AB}$$

Considering the second order term gives the expression of DFTB2. where γ_{AB} is an analytical function. For large distances between atoms A and B the γ_{AB}-function serve as Coulombic interaction, for short distances it describes an on-site electron-electron interaction of same atom A as $\gamma_{AA} = U_A$. U_A is called the Hubbard parameter, obtained as second derivative of the total energy with respect to the charge density of an isolated atom from DFT level. U_A is also an indicator of chemical hardness which defines how the energy of an atom changes when an electron is added or removed. The γ-function assumes that the width of the atomic charge density is proportional to the chemical hardness which works well for many elements except hydrogen, hence a modified γ-function for hydrogen was introduced.

DFTB3

$$E(DFTB3) = E^0[\rho^0] + E^1[\rho^0, \Delta\rho] + E^2[\rho^0, (\Delta\rho)^2] + E^3[\rho^0, (\Delta\rho)^3] = \sum_i^{\mathrm{MO}} n_i \sum_{A,B}^{\mathrm{atoms}} \sum_{\mu\in A} \sum_{\nu\in B} c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{A,B}^{\mathrm{atoms}} \Delta q_A \Delta q_B \gamma_{AB} + \frac{1}{3} \sum_{A,B}^{\mathrm{atoms}} \Delta q_A^2 \Delta q_B \Gamma_{AB} + \frac{1}{2} \sum_{A,B}^{\mathrm{atoms}} V^{rep}_{AB}$$

Considering the last third order term gives the complete energy expression of DFTB3 [59, 57]. The third-order term introduces the $\Gamma_A B$-function consisting derivative of Hubbard derivative U_A with respect to charge. In other words, chemical hardness of atom is now dependent on charge which is a vital energy contribution for charged species.

Repulsive Potential

$$E_{rep} = \frac{1}{2} \sum_{A,B}^{\text{atoms}} V_{AB}^{rep} \tag{3.33}$$

Zeroth order term only depend on reference density and therefore independent of electronic contribution, termed as repulsive potential. It is a function of interatomic distances and doesn't depend on atomic charges.

4. Molecular Mechanics

4.1. Forcefields

If we want to study the dynamics for large biological System like Protein. cell-membrane, Nucleic Acids, quantum mechanical effects are too negligible for such big biological processes rather classical Newtonian mechanics plays a vital role in those processes. In order to study those processes we need another less expensive computational way. Chemical force fields gave us the promising approach. In this approach the whole system is treated as "ball and spring" model and every tiny motion such as stretching, bending, rotation of such "ball and spring" model is parametrised. Therefore we can say the system is guided through a empirical potential energy surface (also called force field) where tiny motions of the system is well defined.

A force field is defined in mathematical expression describing the dependence of the energy of a system based on the coordinates of its particles. It is composed of an analytical form of the interatomic potential energy, and a set of parameters entering into this form. The parameters are typically obtained from ab initio, DFT or semi-empirical quantum mechanical calculations or by fitting to experimental data such as neutron, X-ray and electron diffraction, NMR, infrared, Raman and neutron spectroscopy. Molecules are simply defined as a set of spheres that is held together by simple elastic (harmonic) forces. Generally it must be simple enough to be evaluated quickly, but sufficiently detailed to reproduce the properties of interest of the system. There are many force fields available in the literature, having different degrees of complexity, and oriented to treat different kinds of systems.

However a typical energy expression for a force field may look like this:

$$E_{FF} = \overbrace{E_{stretch} + E_{bend} + E_{torsion}}^{bonded} + \overbrace{E_{LJ} + E_{coul}}^{non-bonded} \tag{4.1}$$

The whole equation can be classified into nonded and nonbonded terms based on connectivity.

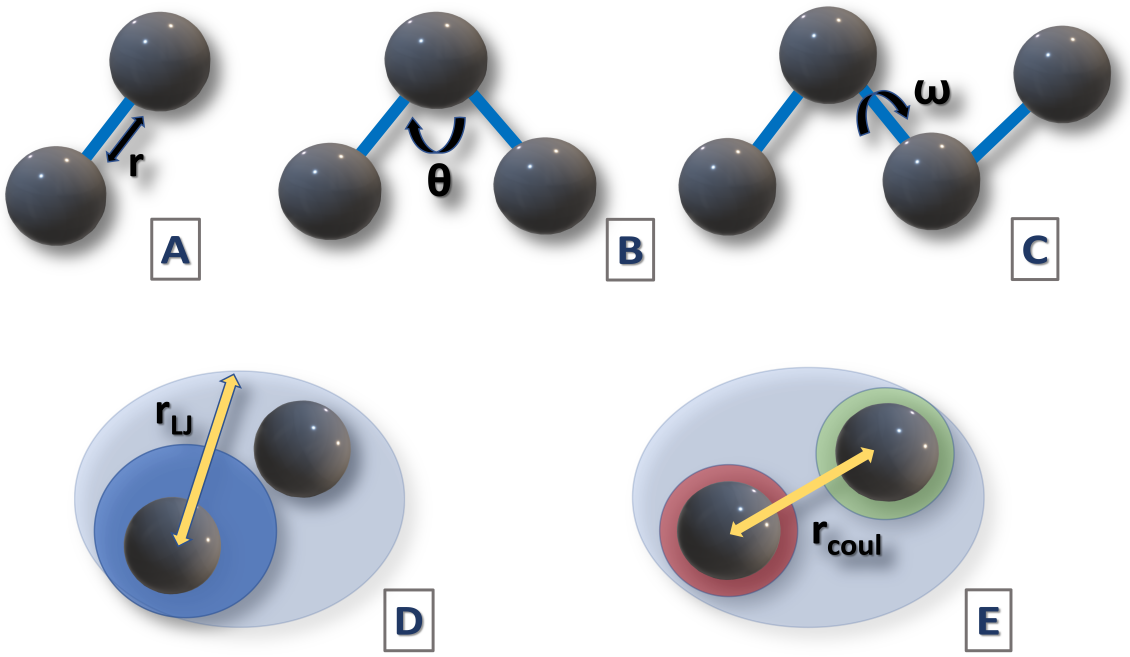

Figure 4.1.: Schematic diagram of forcefield parameters

$$E_{FF} = \underbrace{\frac{1}{2}\sum_i k_i(r_i - r_i^0)^2}_{E_{stretch}} + \underbrace{\frac{1}{2}\sum_j k_j^\theta(\theta_j - \theta_j^0)^2}_{E_{bend}} + \underbrace{\frac{1}{2}\sum_n V_n cos[n\omega - \gamma_n]}_{E_{torsion}} + \sum_i^N \sum_{j>1}^N \left[\underbrace{4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right)}_{E_{LJ}} + \underbrace{\frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}}}_{E_{coul}} \right] \tag{4.2}$$

4.1.1. Bonded interactions

$$E_{bonded} = E_{stretch} + E_{bend} + E_{torsion} \tag{4.3}$$

The bonding terms help to define the covalent energy of a molecule. Here all parameters are defined in a form of harmonic potential. This included bond-stretching 4.1(A), bond-bending 4.1(B), bond-rotation 4.1(C).

Bond Stretching

$$E_{stretch} = \frac{1}{2}\sum_i k_i(r_i - r_i^0)^2$$

Bond stretching is very often represented with a simple harmonic function that controls the length of covalent bonds. Reasonable values for r0 can be obtained from X-ray diffraction experiments or optimised geometry from ab-initio calculations while the spring constant may be estimated from infrared or Raman spectra. The harmonic potential is a poor approximation for bond displacements. Additionally the use of the harmonic function implies that the bond cannot be broken, so no chemical processes can be studied.

Bond Bending

$$E_{bend} = \frac{1}{2} \sum_{j} k_j \theta (\theta_j - \theta_j^0)^2$$

Bending energy potentials are usually treated very similar to stretching potentials; the energy is assumed to behave quadratically with displacement of the bond angle from equilibrium. Only unusual thing happens when θ becomes 180°: the derivative of the potential is enforced to go to zero.

Bond Rotation

$$E_{torsion} = \frac{1}{2} \sum_{n} V_n cos[n\omega\text{-}\gamma_n]$$

The third type of bonding term is the term that describes how the energy of a molecule changes as it undergoes a rotation about one of its bonds, i.e. the dihedral or torsion energy for the system. In contrast to the bond and angle terms a harmonic form for the dihedral energy is not usually appropriate. This is because, for many dihedral angles in molecules, the whole range of angles from 0 to 360 can be accessible with not too large differences in energy.

4.1.2. Nonbonded interactions

The non-bonding terms describe the interactions between the atoms of different molecules or between atoms that are not directly bonded together in the same molecule. These interactions help to determine the overall conformation of a molecular system. The non-bonding interactions arise from the interactions between the electronic distributions surrounding different atoms. The theory of intermolecular interactions is well established, at short range the interactions are primarily repulsive due to the interactions between the electron clouds attributed to quantum mechanical effect of exchange repulsion, which arises when the two clouds are pushed together. At long ranges there are several important classes of interactions. The first are the electrostatic interactions that arise from the interaction of the charge distributions about each molecule or portion of a molecule. Second are the dispersion interactions that are produced by correlated fluctuations in the charge distributions of the two groups. Finally, there are induced or polarization interactions that are caused by the distortion of the charge distribution of a molecule as it interacts with neighbouring groups.

$$E_{non-bonded} = E_{LJ} + E_{coul} \tag{4.4}$$

Lennard-Jones potential

$E_{\mathrm{LJ}}(r) = 4\epsilon_{ij}[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6}]$

Lennard-Jones term estimates long-range dispersion interactions and the short-range repulsive interactions. The energy expression has two terms, first term $(\frac{\sigma_{ij}}{r_{ij}})^{12}$ represents repulsive interaction between two atoms (can be realised as repulsion of electron clouds of two atoms) this interaction is only effective within r0 distance. On the other hand $(\frac{\sigma_{ij}}{r_{ij}})^{6}$ represents attractive interaction in between them (can be considered as dipolar interaction between atoms), it is only effective beyond r0 distance cut-off. Magnitude of the attractive force decreases in the order of 6 with distance which means at longer distance the effect of this potential becomes negligible.

Electrostatic Interactions

$E_{\mathrm{coul}} = \frac{1}{4\pi\varepsilon_0}\frac{q_i q_j}{r_{ij}}$

The electrostatic energy is defined as coulomb interaction between two atoms with point charges q_i and q_j is described in the above equation, where r_{ij} is the distance between them. Charge distributions and the electrostatic energies are essentially arising out from quantum chemical methods. The goal of the force fields though is slightly different. Charge models (q_i, q_j) for the charge distribution in the atoms are simple enough to allow fast calculation of the electrostatic energy but sufficiently accurate enough that the effects due to these interactions can be reproduced. The simplest representation of a charge distribution (charge models) is one in which a fractional charge is assigned to each atom. This is the total net charge of the atom obtained as the sum of the nuclear charge and the charge in the part of the electron cloud that surrounds it.

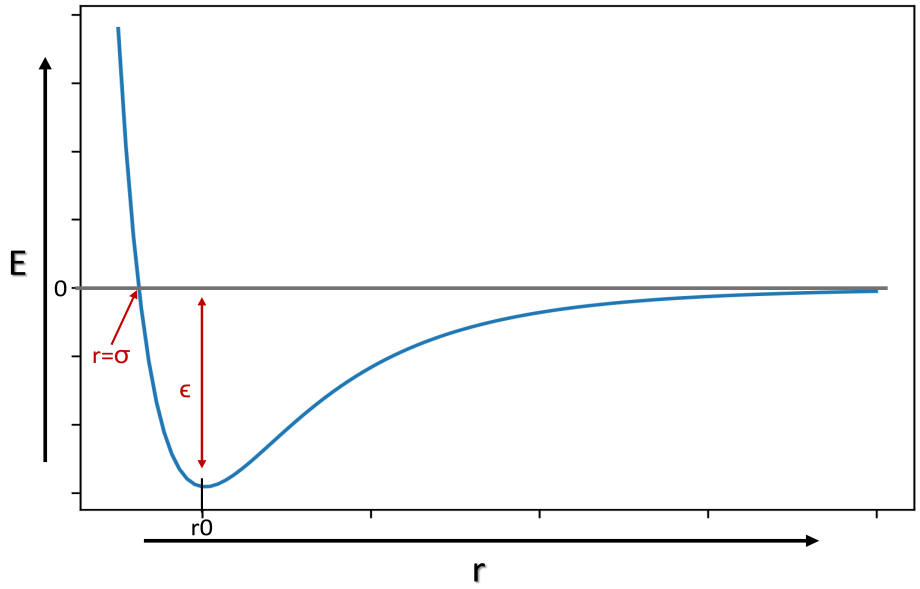

Figure 4.2.: Schematic diagram of Lennard-Jones Potential

Particle Mesh Ewald

Now it is clear that non-bonded interactions are vital part of force fields and at the same time most computationally expensive part of the whole calculation, especially the long range interactions. To reduce the computational cost, these longer-range interactions are typically approximated by using a scheme with more favourable scaling properties, such as Particle Mesh Ewald algorithm. Goal of this algorithm is simple, it just re-scales the long range interactions (which are the function of N^2) to $N \log N$ order. Once these calculations can be done in a linear format, they can be easily parallelized and hence MD will be much faster.

4.2. Molecular Dynamics

Molecular dynamics (MD) is a method in which we can analyse the physical movements of atoms and molecules by solving Newton's equations of motion for a system of interacting particles, where forces these particles are computed from their potential energies designed by the force fields, discussed above. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic evolution of the system. For such systems under the ergodic hypothesis, the evolution of one molecular dynamics simulation is used to determine macroscopic thermodynamic properties of the system.

The essential elements for a molecular dynamics simulation are the interaction potential (potential energy function) for the particles, from which the forces can be calculated, and the initial coordinates of the particles, consisting the system. With these two, we can solve the equation of motion using Newton's law.

$$F = -\frac{d(V)}{dr}$$
$$a = \frac{F}{m} = \frac{d^2r}{dt^2}$$

The above equation is second order differential equation, solving that we get

$$r(t) = r_0 + v_0 t + \tfrac{1}{2}at^2$$

The above equation is the simplest form of the integrator, where r(t) is the displacement at time t, r_0 is the initial coordinate, v_0 is the initial velocity, a is initial acceleration. v_0 comes from Maxwell-Boltzmann's velocity distribution of that temperature, a comes from the forcefield.

Verlet Method

The standard Verlet method is derived from the above equation. If, at a time t, the positions of the atoms in the system are at R(t), then the positions of the atoms at a time t + Δt can be obtained from a Taylor expansion in terms of the timestep, Δt and the positions and their derivatives at time t. After expansion and rearrangements we get the simplest form of numerical integrator, Verlet equation. The final expression looks like this:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + a(t)\Delta t^2 + O(\Delta t^4)$$

Velocity verlet

Velocity verlet is another computationally less expensive approach than verlet equation, the advantage is, it requires less computer memory, because only one set of positions, forces and velocities are needed to be carried at any point of time. Final expressions are given below

$$r(t + \Delta t) = r(t) + v(t)\,\Delta t + \frac{1}{2}\,a(t)\Delta t^2$$
$$v(t + \Delta t) = v(t) + \frac{a(t) + a(t + \Delta t)}{2}\Delta t$$

leapfrog integrator

Another approach is leapfrog, it is very similar to the velocity Verlet method. Here velocity, position, acceleration gets update in every $t - \frac{1}{2}\Delta t$ time.

$\mathrm{v(t + 1/2\ \Delta t)} = v(t - \frac{1}{2}\Delta t) + a\Delta t$
$\mathrm{r(t + \Delta t)} = r(t) + v\Delta t(t + \frac{1}{2}\Delta t)$

Thermostat

The described MD scheme only generates microcanonical ensemble (NVE ensemble). But to mimic the experimental conditions for the simulation we need to add the information of temperature into the system. The easy answer to that is to generate NVT ensemble for the simulation. The way to do it is, the system is kept weakly coupling with a heat bath with some temperature. Then thermostat suppresses the fluctuations of the kinetic energy of the system and therefore cannot produce trajectories consistent with the canonical ensemble (NVT). The temperature of the system is corrected such that the deviation exponentially decays with some time constant τ . Simplest example of thermostat, known as Berendsen's thermostat [145], expression looks like the following:

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}$$

The velocity rescaling (v-rescaling) thermostat [24] is an extension of the Berendsen thermostat for producing a correct ensemble. This is done by adding a random force to ensure the correct distribution of the kinetic energy. In this approach, the velocities are multiplied by a factor K0/K, for forcing the total kinetic energy K towards the average kinetic energy at the target temperature, K0. The rescaling is eventually done by using an auxiliary dynamics as in the following equation.

$$\mathrm{d}K = (K_0 - K)\frac{\mathrm{d}t}{\tau_T} + 2\sqrt{\frac{KK_0}{N_f}}\frac{\mathrm{d}W}{\sqrt{\tau_T}}$$

Although this particular thermostat does not generate a correct canonical ensemble, there are better choice available for us, namely Nosé–Hoover thermostat, V-rescaling thermostat. We have used V-rescaling for our simulations.

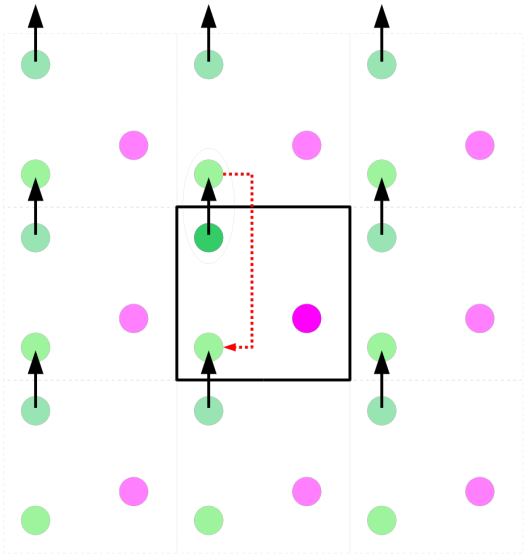

Figure 4.3.: Schematic representation of Periodic Boundary Condition

Barostat

For mimicking the experimental condition for pressure, similar to thermostat, simulation system is coupled with a suitable pressure coupling to generate NPT ensemble. Simplest example is known as Berendsen's barostat.

$$\frac{dP}{dt} = \frac{P_0 - P}{\tau_P}$$

Peridic Boundary Condition

Periodic boundary conditions (PBCs) is a technique in which a complete condensed phase system is modelled as an infinitely and periodically repeated series of copies of a small, but representative part of the full system (see figure 4.3). The assumption of periodicity immediately makes the simulation of such a system tractable because equivalent atoms in each of the copies behave identically and so do not need to be treated distinctly during a simulation.

5. QM/MM

Though forcefields are capable enough to simulate large systems and calculate properties, a major drawback for forcefields is its inability to simulate bond breaking/making. In a biochemical reaction where reaction has to be described in QM level to achieve full description of bond breaking/making/charge transfer. To solve this problem Hybrid QMMM method was introduced by A. Warshel and M. Levitt who were studying the mechanism of the chemical reaction catalysed by the enzyme lysozyme. In hybrid QMMM scheme, the reactive region where the chemical bond formation and breaking events are occurring is treated by QM potential, while the remaining part of the system which is not actively participating in the chemical reaction, is taken care by MM potential; see Figure 5.1. Total energy of the system is divided in the following equation.

$$E_{total} = E_{QM} + E_{MM} + E_{QM/MM} \tag{5.1}$$

E_{QMMM} can be further devided into three contributions.

$$E_{QM/MM} = E^{bonding}_{QM/MM} + E^{vdW}_{QM/MM} + E^{el}_{QM/MM} \tag{5.2}$$

The bonding interactions between QM and MM subsystem, $E^{bonding}_{QM/MM}$ is computed from the MM level of theory when the QM/MM partition cuts across a covalent bond. $E^{vdW}_{QM/MM}$ is the van der Waals dispersion interactions between QM and MM atoms, which is also computed using the force–field. The last term $E^{el}_{QM/MM}$ represents the electrostatic interactions between QM and MM. The calculation of $E^{el}_{QM/MM}$ is technically non-trivial. Based on the interactions in $E_{QM/MM}$, QM/MM scheme can be further divided into Three subclasses: namely mechanical embedding, electrostatic embedding, polarisable embedding.

Mechanical Embedding

In mechanical embedding scheme both QM charge densities and MM charges are considered as point charges and evaluated simply by coulomb law. QM charges doesn't get polarised by MM charges, so the effect of the MM environment to the is only a little. For a charged QM region the results could be misleading.

$$E_{QM/MM} = \sum_{i}^{QM-atoms} \sum_{m}^{MM-atoms} \left(\frac{q_i q_m}{r_{im}} + 4\varepsilon_{im} \left(\frac{\sigma^{12}_{im}}{r^{12}_{im}} - \frac{\sigma^{6}_{im}}{r^{6}_{im}} \right) \right) \tag{5.3}$$

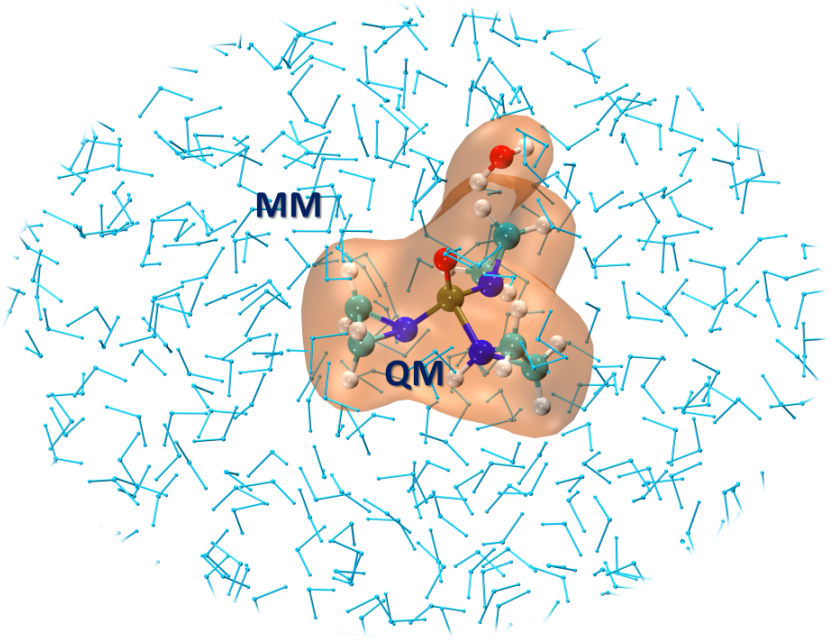

Figure 5.1.: A typical QMMM steup: Participating reagents are considered in QM (orange cloud) rest of the system is treated in classical mechanics(MM)

Electrostatic Embedding

The drawbacks of mechanical embedding scheme is rectified in electrostatic embedding scheme, where QM charge density gets polarised by MM point charges. In this scheme effect of MM environment is well considered into the QM/MM calculations. With DFTB as QM method, the electrostatic potential induced by all MM atoms enters the DFTB3 Hamiltonian matrix elements and affects the QM charge distribution. From the equation the new rescaled QM charge is $Z_i q_i$, Z_i is the rescaling factor.

$$E'_{QM/MM} = \sum_{i}^{QM-atoms} \sum_{m}^{MM-atoms} \left(\frac{Z_i q_i q_m}{r_{im}} + 4\varepsilon_{im} \left(\frac{\sigma_{im}^{12}}{r_{im}^{12}} - \frac{\sigma_{im}^{6}}{r_{im}^{6}} \right) \right) \tag{5.4}$$

Polarisable Embedding

In this scheme both MM and QM charge get polarised by each other and the new rescaled charges are $Z_i q_i$ for QM charge density and $Z_m q_m$ for MM charges. Polarisable embedding is important to consider when both MM environment and QM region is highly charged. Simulations of systems like photoswitches, fluorescent proteins, chromophores especially in excited state dynamics this embedding scheme becomes important.

$$E'_{QM/MM} = \sum_{i}^{QM-atoms} \sum_{m}^{MM-atoms} \left(\frac{Z_i q_i Z_m q_m}{r_{im}} + 4\varepsilon_{im} \left(\frac{\sigma_{im}^{12}}{r_{im}^{12}} - \frac{\sigma_{im}^{6}}{r_{im}^{6}} \right) \right) \tag{5.5}$$

6. Enhanced Sampling Techniques and free energy computation

6.1. Enhanced Sampling Methods

Free Energy

To calculate the free energy of the system, we first need to understand the concept of probability in simulation. According to ergodic hypothesis ensemble average property is equivalent to time average property of the system. Under this circumstance partition function can be regard as probability and according to Boltzmann distribution, probability can be written as:

$$P(r,p) \propto exp(\frac{-E(r,p)}{k_B T})$$

Where E is energy of the system, k_B is Boltzmann constant and T the absolute temperature. Now using this probability we can calculate various properties of the system including free energy (Gibbs' and Helmholtz free energy):

$$F = -k_B T \ln(P(r,p))$$

Collective Variable

Collective variables (CVs) are predefined reaction coordinates which are used to describe pathway of a physical process or mechanism of a reaction or It can be any function S(r) of atomic coordinates such as a distance between two atoms, an angle between three atoms or a dihedral angle between four atoms. An ideal CV should distinguish each of the important states in the mechanism of the process of interest in a identifiable manner. For complex biological processes even more complex CVs are used which include many or even all atoms, such as a normal mode from a harmonic vibrational analysis or a RMSD to a reference structure.

Potential of Mean Force

Free energy along a specific reaction coordinate S is referred to as the potential of Mean force (PMF). Probability of finding state A ($P(S_A)$) and state B ($P(S_B)$ along reaction coordinate S can be evaluated using histogram method. Then relative free energy difference could be obtained using the following equation

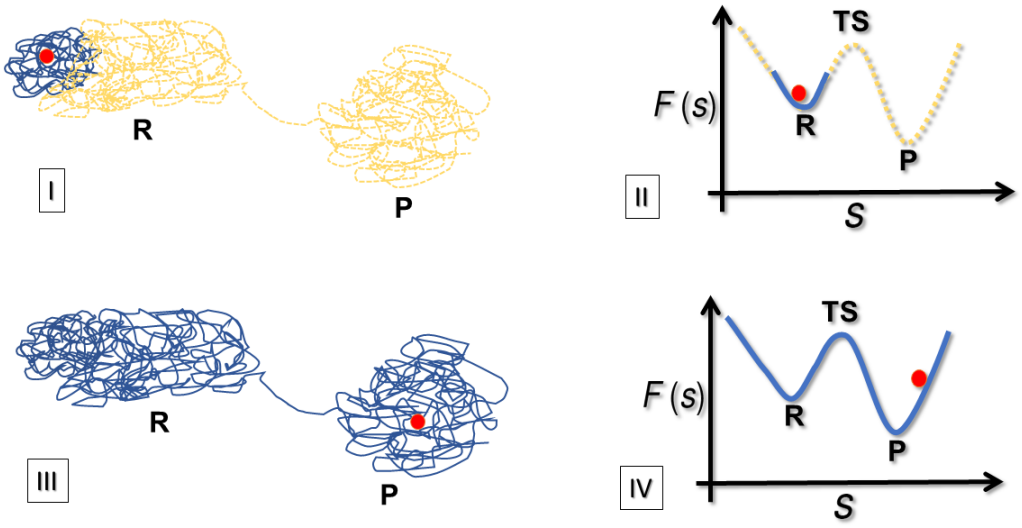

Figure 6.1.: Schematic diagram of two minima seperated by a bariier: sampling bottleneck

$$\Delta F = F(S_A) - F(S_B) = -k_B T \ln \frac{P(S_B)}{P(S_A)}$$

Sampling Bottleneck

In real world chemical or biochemical problems are often associated with very slow kinetics, indicates such processes have high free energy barrier. The typical timescale that can be accessed by MD simulation is restricted to few hundred of nanoseconds, whereas the enzymatic reactions occur in the order of milliseconds to seconds. These processes are termed as "Rare events". In our context examples of rare events are typically chemical reaction, transition state searching, protein folding etc. To simulate such processes associated with large free energy barrier in a phase space (S) we face a sampling problem, where ergodicity is hindered by the form of the system's energy landscape of the system.

To understand this problem in detail lets consider the schematic diagram in 6.1(I) and (II), where a complex phase space (S) is shown. In normal molecular dynamics the sampler (red dot) can only sample a tiny part of the phase space thus the probability and the free energy obtained, could not capture the whole process. This phenomenon is called "Sampling Bottleneck". Now to solve this problem either one need a high powerful computing power or the other way to solve this problem is enhanced sampling method.

In simple words enhance sampling methods are the ways to increase the probability of visiting a configurational state by modifying the potential energy by adding biased potential or increasing the simulation temperature. To understand this, lets go back to the 6.1(III) and (IV). If enhanced sampling method is applied in the direction of S (R to P)

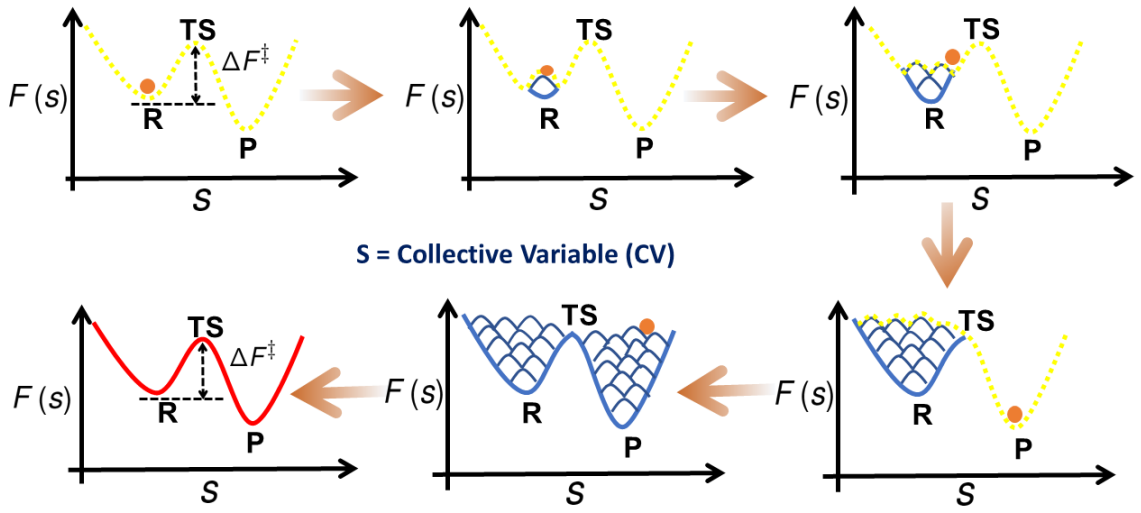

Figure 6.2.: Schematic diagram of Metadynamics: showing how "Gaussian potential" bias is getting filled in a certain CV space (S) with the information of previously deposited bias

the sampler (red dot) can visit all the important configurations and thus the probability of state R and P could be estimates and thus the free energy could be obtained.

There are various enhanced sampling methods available, largely classified into two catagories: 1) CV based methods (Metadynamics, Umbrella Sampling, Blue Moon Sampling) 2) Non-CV methods (REMD,REST2).

6.2. Metadynamics

Metadynamics proposed by Laio and Parrinello (in 2002) [99] , is one of the CV based enhanced sampling methods that relies on modifying the potential energy by supplementing biased potential. For more than a decade, metadynamics has been proven to be successful method in the fields of chemistry, biology, physics and material science citations Herein this work all the chemical reactions and bio-chemical reactions are employed with metadynamics within framework of NVT QM/MM MD simulations.

In this approach, one or a set of collective variables (S) the system are chosen and accelerated by slowly augmenting the bias potentials along the CV–trajectory; see Figure 6.2. The trail of bias potential essentially prevents the system to revisit the previously explored region of CV–space, thus the bias is history dependent and thereby accelerating the sampling.

Equation of the bias potential ($V^b(s,t)$) is given below, shaped like a Gaussian curve. W_0 is the initial Gaussian height, σ is the Gaussian width, τ is the interval of deposition each Gaussian bias. These parameters could be tuned to optimise for a suitable bias potential

for a specific system. When the time–dependent bias potential completely compensates the underlying free energy surface, the system will escape from the current free energy basin to the next in self–guided manner 6.2 .

$$V^b(s,t) = \sum_{\tau<t} W_0(\tau) \exp\left(-\frac{[s - s^{(0)}(\tau)]^2}{2\sigma^2}\right)$$

After getting the information of the complete bias deposition one can obtain the free energy of the system using reweighing the total bias potential, in this case the negative sum of augmented bias potentials simply provides the estimation of the underlying free energy surface.

$$F(S) = -\lim_{t\to\infty} V^b(S) = -\sum_{i<n(t)} W_0 e^{-\frac{[S-S_i]^2}{2\sigma_i^2}}$$

6.2.1. Well-tempered Metadynamics

In standard metadynamics, Gaussian bias with constant heights are added for the entire course of a simulation. As a result, when the simulation reaches eventually at high free-energy regions and the estimation of the free energy calculated from the bias potential will give a higher bound of the real value. This error is an artifact introduced by Gaussian bias height. As a result, diffusive free energy surface appears to be rough. In other words, it is difficult to get a converged free energy surface (free energy doesn't change after that) in case of standard metadynamics. As an alternative to resolve this requirement computationally more expensive Well-tempered metadynamics was proposed.

In well-tempered metadynamics, in order to get smooth free energy surface it is necessary to re-scale Gaussian bias height with respect to progression of sampling. Gaussian heights are expected to decrease slowly with time and increase of the energy hill. For this purpose one extra term is added in the equation called "bias factor" which determines the rescaling of the bias heights.

$$W(k\tau) = W_0 \exp\left(-\frac{V(\vec{s}(q(k\tau)), k\tau)}{k_B \Delta T}\right) \gamma = \frac{T + \Delta T}{T}$$

Looking at the above equation, gaussian bias height term is replaced by height rescaling term W(kt) which depends on initial gaussian height W_0 and bias-factor γ. Bias-factor γ is the rescaling factor. when γ is zero the whole simulation becomes free molecular dynamics independent of bias, when γ is infinity the simulation turns back to standard metadynamics. So γ factor should be optimised for each system for a better convergence.

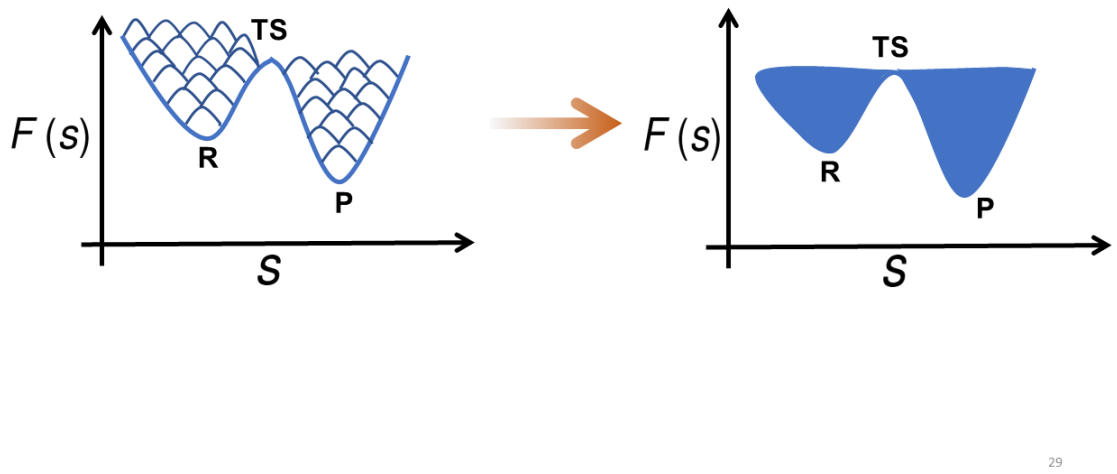

Figure 6.3.: Schematic diagram of well-tempered metadynamics: showing scaled down biases could produce much smoother energy surface (right side) than the normal metadynamics (left side) and thus introduce less error in the calculation

7. Machine Learning and Neural Networks

7.1. Artificial Neural Network

Artificial neural networks (ANNs) or simply neural networks (NNs), are statistical computing systems inspired from the biological neural networks that constitute animal brains. First ever artificial neurons were proposed as early as 1943 as mathematical tools to understand signal processing in the human brains. ANN is based on a collection of connected units or nodes called artificial neurons or Perceptron, which loosely model the neurons in animal brain. Each connection, like the synapses in an animal brain, can transmit a signal to other neurons.

The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called edges. Neurons and edges typically have a set of weights that is adjusted, the process is called learning proceeds. The weight increases or decreases the strength of the signal at a connection.

7.1.1. Perceptron

Perceptrons (single neuron) are fundamental building blocks of neural network. It takes values as inputs multiply them with statistical weights and added all together then a bias is added using a suitable nonlinear function. The final value is then the output of the perceptron. Looking at the equation below and fig at 7.1, x_j represents individual inputs, w_j represents corresponding weights, w_0 represents bias, φ is the activation function. The output of the whole function is given by y. tuning w_j values and w_0 we can obtain our desired output value.

$$y = \varphi\left(\sum_{j=0}^{m} w_j x_j + w_0\right)$$

7.1.2. Neural Network Architecture

Now for Big data sets input layers will be highly multidimensional, for such kind of situation only one perceptron will not be enough rather a set of perceptron is needed to maintain statistical stability of the model. Figure 7.3 represents a typical neural network architecture. Hidden layers essentially seperate the data sets nonlinearly for better decision making. More than one hidden layer could be added, as a results we will left with two many weights and biases to tune in order to get correct output.

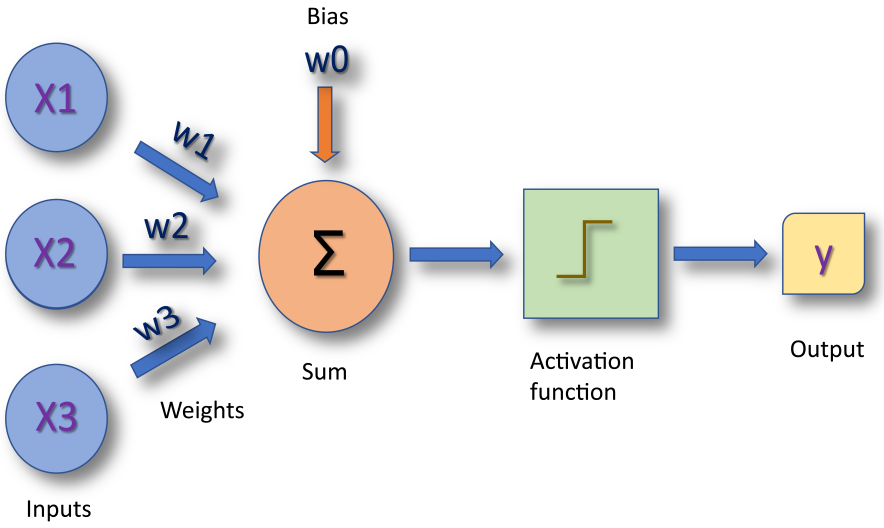

Figure 7.1.: Schematic diagram of single perceptron

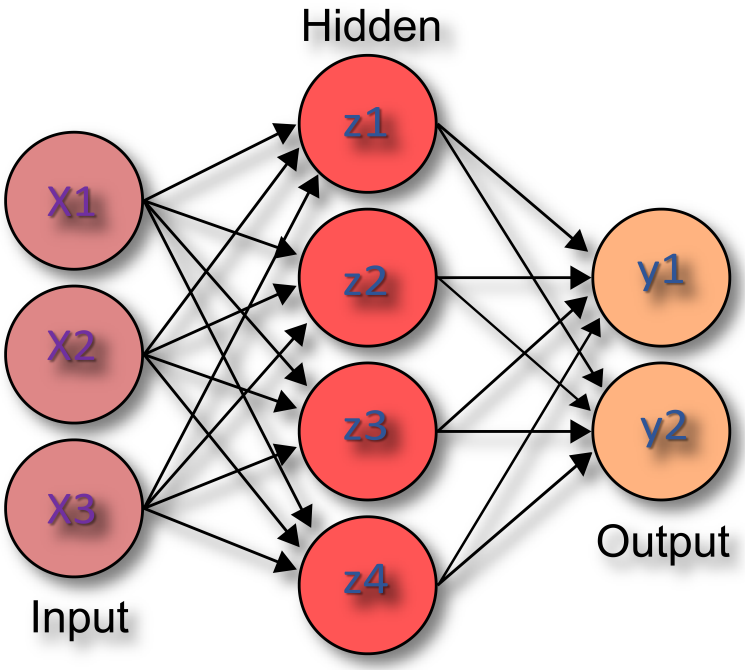

Figure 7.2.: Schematic diagram of Artificial Neural Network: Each circle represents single perceptron except the input layer. A simple connectivity network among perceptrons is shown here.

7.1.3. Activation function

Activation functions are essential part of ANN, weighted sum of its inputs has to pass through an activation function. The primary role of the Activation Function is to transform the summed weighted input from the node into an output value to be fed to the next hidden layer or as output and also it decides whether or not a neuron will be activated or not to the next layer. This simply means that it will decide whether the neuron's input to the network is relevant or not in the process of prediction. For this reason, it is also referred to as threshold or transformation for the neurons which can converge the whole network. Depending on the data set and desired output different types of activation functions are there. Linear function is the most basic activation function most commonly used activation function is sigmoid function and Rectified linear unit function. For each hidden layer different activation function can be employed.

7.1.4. Training

After we constructed our desired neural network model based on our data set, the next step is tuning the weights and biases. The process is called training of Neural Network. The training method is basically a trial and error method to assign new weights and biases, error is estimated by a suitable loss function, the goal is to get weights and biases in such a way that the error is minimum.

Loss function

Simple two examples of loss functions are shown here. First one is "Binary Cross Entropy Loss Function" (BCE) and the second one is called "Mean squared error loss function" (MSE) both of the expressions are given below. Using such function the error between actual(y) and the predicted (x) values are estimated. Minimising this function will generate new sets of weights(W) which will again be applied in the neural network to obtain new predicted value. This process will continue until the loss function attains a minimum threshold value and hence the final weights (W^*).

$$J^{BCE}(W) = -\frac{1}{n}\sum_{i=1}^{n}(\underbrace{y^i}_{actual} \log(\underbrace{\varphi(x^i;W)}_{predicted}) + (1 - \underbrace{y^i}_{actual}) \log(1 - \underbrace{\varphi(x^i;W)}_{predicted})$$

$$J^{MSE}(W) = -\frac{1}{n}\sum_{i=1}^{n}(\underbrace{y^i}_{actual} \log(\underbrace{\varphi(x^i;W)}_{predicted})^2$$

Optimising loss function

The holy grail of this optimisation problem is to get W^* which can give us most accurate prediction from the neural network model. Popular optimiser algorithms include gradient descent, adam optimiser etc.

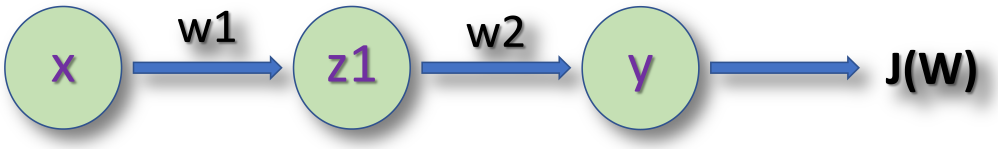

Figure 7.3.: Schematic diagram of backpropagation

$$J(W^*) = argmin(-\frac{1}{n}\sum_{i=1}^{n}(\varphi(x^i; W)), y^i)$$

Gradient descent is the simplest optimiser, basic process is simple, first we need to initial arbitrary random weights for the neural network, then estimate the error using loss function and then compute the gradient of the loss function to get new sets of weights. This process continues till convergence is achieved.

$$\nabla_{J(W)} = \frac{dJ(W)}{dW}$$
$$W_{new} \Leftarrow W_{old} - \eta\frac{dJ(W)}{dW}$$

Backpropagation

The backpropagation is a algorithm that computes the gradient of the loss function with respect to each weight by the chain rule, computing the gradient one layer at a time, iterating backward from the last layer to avoid redundant calculations of intermediate terms in the chain rule.

$$\frac{dJ(W)}{dW2} = \frac{dJ(W)}{dy} * \frac{dy}{dW2}$$
$$\frac{dJ(W)}{dW1} = \frac{dJ(W)}{dy} * \frac{dy}{dz1} * \frac{dz1}{dW1}$$

Learning Rate

Learning rate determines the step size in the optimisation denoted by η in the above equation. Learning rate has to optimised for the system, lower learning steps will make the optimiser for ever to converge and bigger learning rate can skip the actual global minima of the optimisation.

Hyperparameters

Hyperparameters in neural network are the parameters that defines the structure and function of a particular neural network model such as number of hidden layers, number of neurons in hidden layers, activation function, weight initialisation, Number of iterations for training (epochs), learning rate, batch size, choice of optimiser. Hyperparameters can be optimised in a automated manner using different algorithms.

Symmetry Functions

For low-dimensional PESs, e.g., for small molecules, the number of degrees of freedom are typically fixed. For high-dimensional systems we need a NN potential that is applicable to large numbers of atoms. However, the number of neighbouring atoms in the local chemical environments cannot be fixed, since in the course of a MD simulation atoms can enter or leave the cut-off sphere. This represents an additional requirement for the symmetry functions. Their number must be fixed, i.e., it must be independent of the actual number of neighbours in the local environment. Otherwise it would be necessary to train different NNs for each possible number of atoms, which is not practical. This problem can be solved by constructing many-body symmetry functions simultaneously depending on the positions of all atoms inside the cut-off sphere.

Because of the requirement that the atomic NNs have a fixed number of input nodes, the number of symmetry functions in the Gi vectors must not change with the number of atoms in the cut-off sphere, even if this number increases or decreases, for example, in molecular dynamics simulations. This can be achieved by using the "radial" symmetry function defined by equation G^2 below. This is a sum of products of a Gaussian function of the interatomic distance and the cut-off function, which allows for a physical interpretation as the effective coordination number of the central atom. The typical use of 5–6 radial functions with different Gaussian exponents h provides a radial fingerprint of the neighbouring atoms. The parameter R_s can be used to shift the centers of the Gaussians to specific interatomic distances. In the case of multicomponent systems, one set of radial functions is used for every neighbouring element in the system. Since the radial functions alone are unable to distinguish different angular arrangements of neighbours, a set of "angular functions" should be used, depends on the angles qijk centered at atom i and formed with neighbours j and k, which both need to be within Rc. The use of a set of functions with different exponents z allows a fingerprint of the angular distribution to be obtained, while l =: 1 can be used to adjust the positions of the maxima and minima of these functions.

$$G_i^2 = \sum_j exp[-\eta(R_{ij} - R_s)^2] f_c(R_{ij})$$

$$G_i^4 = 2^{1-\zeta} \sum_{j,k\neq i}^{all} (1 + \lambda \cos\theta_{ijk}) exp[-\zeta(R_{ij}^2 + R_{ik}^2 + R_{jk}^2)] f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk})$$

$$f_c(R_{ij}) = \begin{cases} 0.5[\cos(\pi \frac{R_{ij}}{R_c}) + 1] & R_{ij} \leq R_c \\ 0 & R_{ij} > R_c \end{cases}$$

7.2. Behler-Parrinello Neural Network

Neural Network has made its way into computational chemistry when there is a demand of making fast and accurate calculations. DFT is a method that gives very accurate results but computational cost limits the size of the chemical system and even time-scale of the MD simulation of small systems using DFT potential. To bridge this gap there were several attempts by using training a neural network using DFT level energy to make calculations faster by prediction, such kind of trained potential is known as neural network potential (NNP).

A very popular NNP method, which can be applicable into high-dimensional systems containing thousands of atoms was proposed by Behler and Parrinello in 2007 [15]. In this approach a separate NN is used for each atom in the system. Each of these "atomic NNs" or subnets provides the energy contribution each atom as a function of the chemical environment, symmetry function. The total energy of the system is then obtained as the sum over all atomic energies. For a given element, the atomic NNs are constrained to have the same architecture—specifying the number of hidden layers and neurons. The structure of the resulting high dimensional NNP (HDNNP) is shown in figure 7.4. from the figure it can be seen that the input of each atomic NN is a vector of atom-centered symmetry functions [13] describing the local chemical environments of the atoms. These are defined by a cut-off radius Rc, which has values between 6 and 10 c, is a convergence parameter that needs to be increased until all the energetically relevant interactions are included, since atoms outside Rc do not enter the energy contribution of the respective central atom. There are now four generations of this kind of neural networks [14]

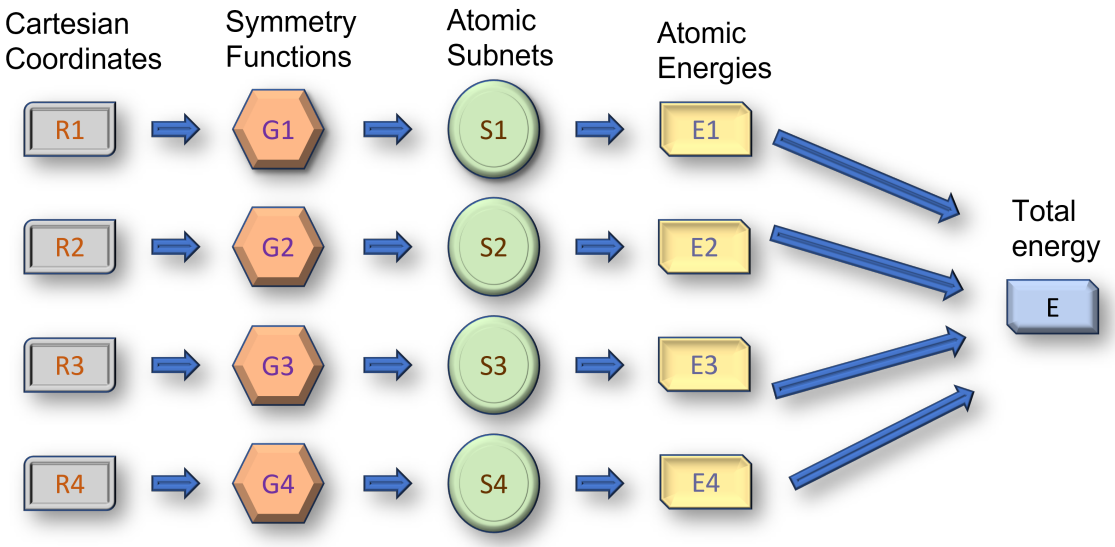

Figure 7.4.: Schematic diagram of Behler–Parrinello Neural Network exhibiting important features

Part III.

Results

8. Re-parameterisation of Phosphorus-Nitrogen Pair Potential in DFTB3

8.1. Introduction

Phosphorus is one of the very important elements in the world of bio-molecules and medicines. From genetic materials like DNA, RNA to other bio-molecules like ATP, phospholipids to cancer drug such as TEPA,ThioTEPA phosphorus can be found in a large spectrum of molecules. Phosphorus containing bio-molecules play key roles in essential biological functions involved in tiny microscopic species to large animals and plants. Such commonly occurring process is phosphoryl transfer reaction, for example, arguably the most important chemical process in biology. It can be found in the process of photosynthesis, Krebs cycle etc. Perturbations in phosphoryl transfer enzymes are involved in many serious human diseases such as cancer. Protein kinases and phosphatases are among the most important drug targets there are 2000 protein kinases and 1000 phosphatases in the human genome, and these enzymes are essential to key cellular processes such as the control of cell cycles and division. Besides human body phosphorus key element in bacterial cell, involves in phosphorylation which is essential for bacterial life cycles. Identifying such phosphorylated intermediate could help us design a potential drug in resistance to bacterial infection. Famous example of such phenomenon is histidine kinase, a key part of in bacterial signal transaction system involves in histidine phosphorylation. Such histidine kinases predominantly present in a large group of bacteria involve in the same process they are notably absent from the human body, and that makes these enzymes are suitable targets for developing drug in bacterial resistance.

This takes us to our main focus of the paper which is simulating and predicting mechanism of histidine phosphorylation. Phosphohistidine (product of histidine phosphorylation) is considered as highly unstable compound mostly serve as intermediate in a long sequence of bio-chemical reactions and therefore it is extremely difficult to detect in the experiments. This experimental bottleneck makes way for computational studies to investigate possible mechanism of this reaction. In this connection, here it is worth mentioning about QM/MM, one of the important methods to study reactions in theoretical chemistry/bio-chemistry. Thanks to the development of powerful computer hardware traditional QM/MM simulations are useful but it remained computationally demanding and only allows us to study 50-100ps time scale which can be sufficient for a simple single step reaction for small molecules. But for large enzymes and multiple step reactions this

methods are still remained challenging. This bottleneck keeps the door open for less expensive semi-empirical QM methods.

DFTB3 brings a promising approach to this bottleneck. As mentioned above DFTB3 is an approximate Density Functional Theory (DFT) and is derived by expanding the DFT total energy functional up to third order around a reference charge density. The resulting perturbative series is further approximated by applying a minimal basis LCAO expansion of the KohnSham orbitals. The resulting approximate total energy terms have to be parametrised, and two classes of parameters can be distinguished: (i) the electronic parameters, which determine the atomic minimal basis set and the atomic reference densities as well as the chemical hardness values of the involved atoms the determination of these parameters is quite straightforward; (ii) the repulsive energy parameters, which are necessary to determine the atomic pair potentials modelling the zero-th order contributions in the density expansion. Although these terms can in principle be computed based on DFT calculations, to achieve good general accuracy and partially compensate for approximations made in the other terms, an empirical fit to larger test sets is necessary, and therefore their determination is usually more involved.

In DFTB3, the 3OB set of parameters is most commonly used for organic and biological systems. However, there are a few cases where a limited transferability was found for some complex chemical reactions. This led to incorrect reaction energetics, e.g. phosphate hydrolysis reactions, thiol-disulfide exchange reaction. In the past Phosphate hydrolysis reaction was fixed with a SRP, in this thesis we also fixed thiol-disulfide reaction with another SRP. Here we have focused on the process of autophophorylation of histidine, a key reaction in bacterial two component signal transaction system where also 3OB failed to reproduce correct geometries and reaction energies. Thus, we reparamterised the phosphorus–nitrogen pair potential as described in detail in Methods section.

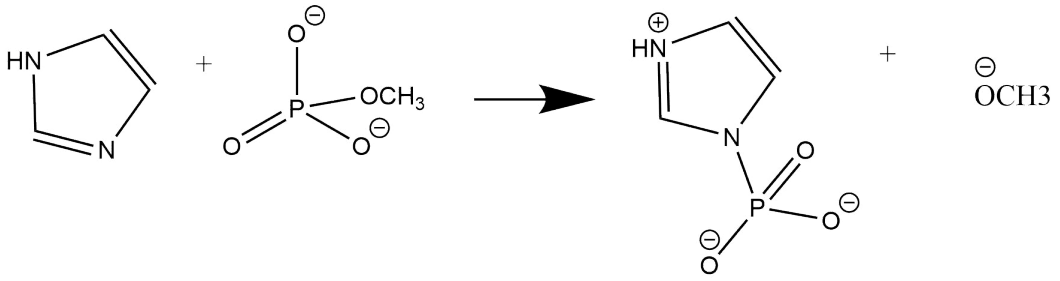

Figure 8.1.: The Model reaction we considered for QM/MM

8.2. Methodology

8.2.1. Reference Free energy calculations for benchmark

Two dimensional QM/MM metadynamics was performed on below mentioned reaction 8.1 (using P-N and P-O distances), where only the reactive molecules were considered in the QM region and explicit water was treated in MM. The PMF obtained from the calculation gave us the estimation of reaction barrier and reaction enthalpy for our parameterisation. We used DFT functional B3LYP and dunning type bass set aug-cc-pVTZ as QM level of theory. QM/MM ran for 372 ps. Free Energy surface obtained from this calculation is discussed in the result section. The details of the metadynamics and Molecular dynamics parameters can be found in appendix.

8.2.2. Reparametarisation

Here we discuss the method we follow for DFTB. To start with we need to again go back to the equation of DFTB3.

The total energy of DFTB3 is given as follows, we discussed in the Method chapter in details:

$$\begin{aligned} E &= E^{(1)} + E^{(2)} + E^{(3)} + E^{\text{rep}} \\ &= \sum_{i}^{\text{MO}} n_i \sum_{a,b}^{\text{atoms}} \sum_{\mu \in a}^{\text{AO}} \sum_{\nu \in b}^{\text{AO}} c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{a,b}^{\text{atoms}} \gamma_{ab} \Delta q_a \Delta q_b + \frac{1}{3} \sum_{a,b}^{\text{atoms}} \Gamma_{ab} \Delta q_a^2 \Delta q_b + \sum_{a \neq b}^{\text{atoms}} V^{\text{rep}}_{ab} \end{aligned}$$

This equation consists of two parts: electronic part $E^{(1)} + E^{(2)} + E^{(3)}$, involving so-called electronic parameters, which in this work are taken from the general-use 3OB parameter set [59, 62]. The repulsive part represents repulsive potential expressed in terms of pair potentials V^{rep}_{ab}, which are specific to respective pairs of chemical elements and depend on interatomic distance but not on atomic charges. Their parametrisation is done by fitting to a selected set of reference atomization energies, molecular geometries, and barrier or reaction energies. Procedure is carried out according to a partially automatized procedure [61], discussed briefly in the next segment.

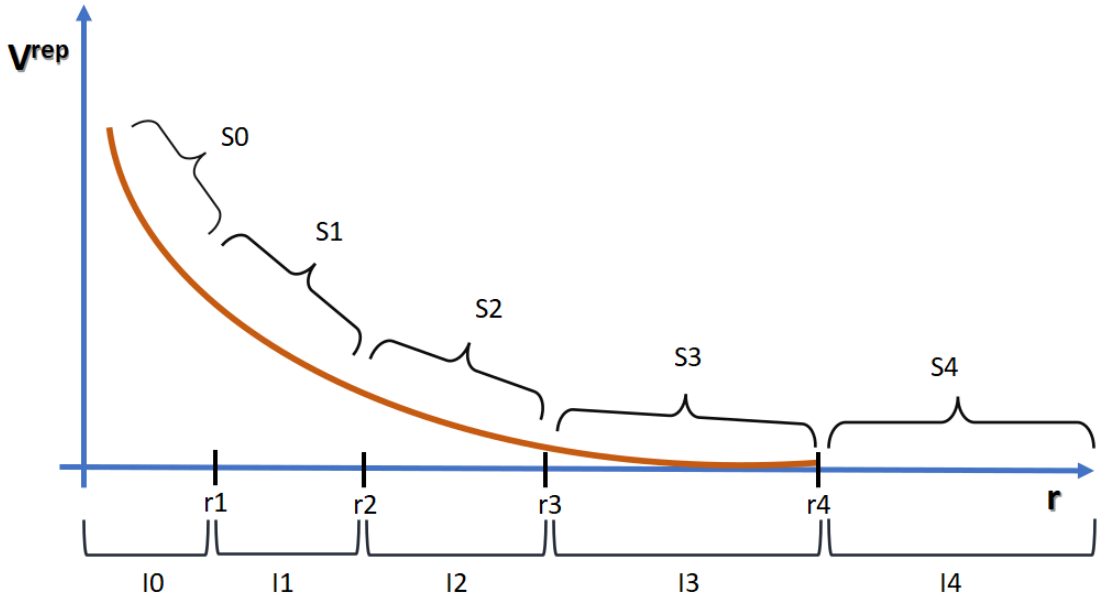

Figure 8.2.: A schemetic diagram of repulsive potential in a form of a spline, where V^{rep} axis represents repulsive potential and r axis represents interatomic distances

Representaion of Repulsive Potential

Fig 8.2 represents a typical repulsive spline with respect to interatomic distance of two atoms. In 3OB parameter scheme repulsive parameter is defined as fourth order spline. At any point of the curve represents energy of the point and slope of that point represents force (second derivative). Now for distance between atom type A and B interatomic distance is divided into several intervals namely I0,I1... using a set of division points (also called grid points) (r1,r2,r3...). Now for each intervals (except for first interval I0) fourth order polynomial is defined as shown below.

$$S_i(r) = \sum_{k=0}^{4} s_{ik}(r - r_i)^k \tag{8.1}$$

In the above equation 8.1 the polynomial is written in terms of interatomic distance, s_{ik} are the coefficients (unknown quantity), in this case (fourth order polynomial) there are 5 coefficients, to be determined. For solving the equation first three derivatives of the equation 8.1 are required to be equal to be same for the next interval ($I(n+1)$) and therefore last division point can be considered a cut-off after which the potential goes to zero. The solution of the equation has to be considered under boundary condition of the continuity equation mentioned later.

However for the first interval (I0) it is essential that the function S0 has to be an exponential function:

$S_0(r)$ =α exp(β r + γ)

Values of three parameters α β γ has to be chosen in such a way that it should match the value of S1 at r=r1 division point.

The requirement of the spline function should be continuously differentiable up to the second derivative in the interval (r(n), r(n+1)). Thus it can be written as:

$$
\begin{aligned}
S_i(r(i+1)) - S_{i+1}(r(i+1)) &= 0 \\
S_i^{'}(r(i+1)) - S_{i+1}^{'}(r(i+1)) &= 0 \\
S_i^{''}(r(i+1)) - S_{i+1}^{''}(r(i+1)) &= 0 \\
S_i^{'''}(r(i+1)) - S_{i+1}^{'''}(r(i+1)) &= 0
\end{aligned}
$$

Atomisation energy is simply the difference between energy of the molecular potential energy and the sum of the individual atomic energies of the same molecule. This quantity is one of the essential reference data for parameterisation. For 3OB sets of parameters, generally the atomisation energies were taken from G3B3 level of theory.

$$
E^{at} = E^{mol} - \sum_{a=atoms} E_a^{el} \tag{8.2}
$$

Overbinding or underbinding is a key concept, used in parameterisation to reproduce correct relative energies (generally to match B3LYP relative energies). All 3OB Special Reaction parameters such as OP-hydrade [62] has some overbinding energy for better reproduction of energy barriers and reaction energies comparable to B3LYP level. Concept is increasing the atomisation energy (overbinding) or decreasing the atomisation energy (underbinding) to make the repulsive curve smoother, which can reproduce our desired relative energetics and structures. We will use this concept later. For instance lets take the case of 8.1 using OP-hydrade SRP, in the reactant there are four P–O bonds and in product there are three P–O bonds, each bond carries 10kcal/mol overbinding energies, which means reactant minimum is shifted down by 40kcal/mol and product minimum 30kcal/mol. In this situation in order to make the synergy in between the reaction energy P–N bond has to have some overbinding energy. Goal of this re-parameterisation is to fix both structure, reaction energy and reaction barrier of the specific reaction. 8.1.

New repulsive potentials for P–N and N–P were created in this work, by means of a fit for Phosphorus–Nitrogen bond containing molecules (including all electronic parameters). In this case there are two specific molecules were taken: namely Imid-Phos-ester-3H, Imid-Phosphate-2H (structures are shown in figure 8.3). Though both of the molecules have no real chemical existence they reproduce very similar chemical environment of the real molecule. Generally charged molecules are protonated to make a neutral species in order to optimise the geometry in gas-phase environment. In parameterisation scheme this is very commonly used technique, because we need to use optimised structures as reference from a better method (e.g. B3LYP) in order to achieve correct geometry of the molecule. In this case molecule Imid-Phos-ester-3H mimics the real transition state of the reaction 8.1 and molecule Imid-Phosphate-2H mimics the real product of the reaction 8.1.

For the solution of the linear equation system set for determining the repulsive pair potential spline, suitable division points are chosen. Further, additional equations were introduced to make the repulsive potential convex at the P–N distance where the spline

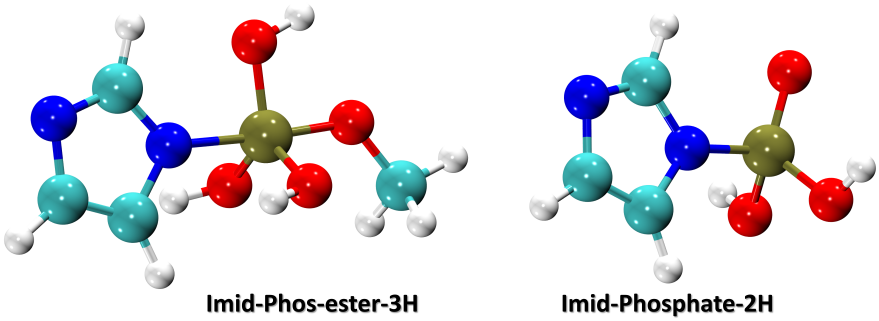

Figure 8.3.: molecules considered for re-parameterisation

starts. The geometries are optimised using B3LYP/aug-cc-pVTZ level of theory, and the atomization energies E^{at} are obtained with G3B3 [8] calculations following standard procedure [60] (already described above in brief). An overview of all reference systems and values that lead to the repulsive potentials related to P–N is provided in Tab. 8.1.

The resulting repulsive potential still could not produce correct energy barrier. This could be due to the fact that other SRP already have some overbinding which makes the relative energies quite obviously show a discrepancy in the reaction energy and barrier. so we decided to overbind the P–N bond in the molecule of Imid-Phosphate-2H by increasing the atomization energy by 25 kcal/mol. The resulting repulsive pair potential reproduces reaction barrier and reaction enthalpy in accordance with the B3LYP/aug-cc-pVTZ QM/MM with an error of ca. 2 kcal/mol 8.5. New repulsive potential and the old 3OB repulsive potential is shown in 8.4.

Molecule	Charge	E^{at} (kcal/mol)
Imid-Phosphate-2H	0	1505
Imid-Phos-ester-3H	0	1947

Potential	Division points (a.u.)	Additional equations
P–N	2.7, 3.5, 4, 4.5, 5	$V''(1.487\,\text{Å}) = 0.84$ a.u.

Table 8.1.: Data used to fit the P-N pair potential spline

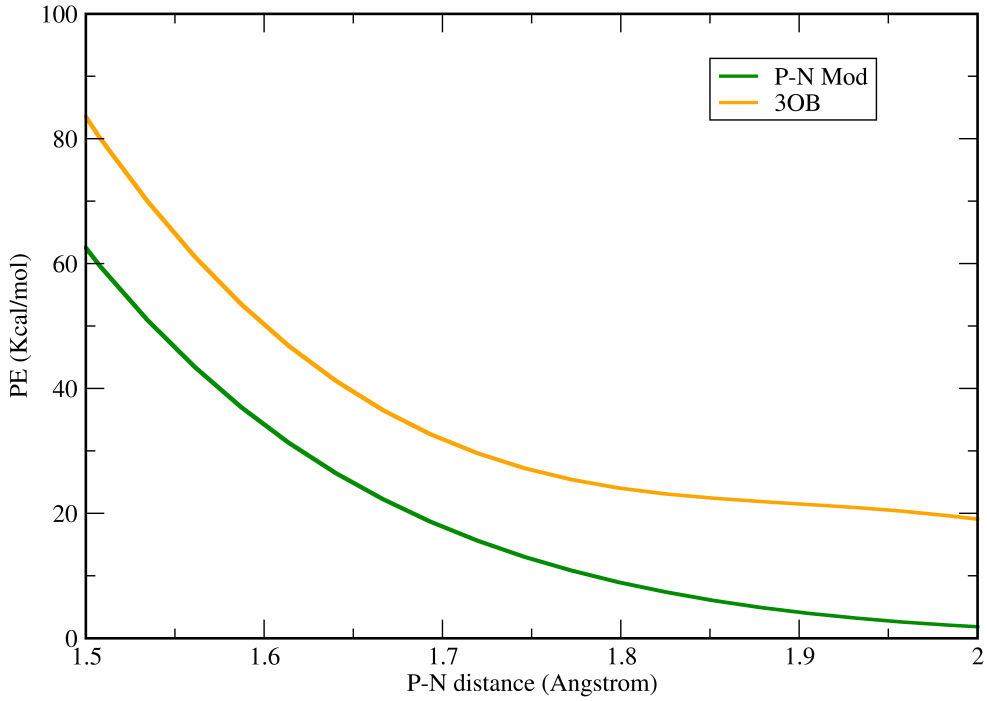

Figure 8.4.: comparison of old 3OB repulsive potential and new modified P-N repulsive potential

From figure 8.4 we can see that the new repulsive potential curve is overbound by 20Kcal/mol. we can also say that in the new repulsive potential P-N interaction is more stronger than previous case. P-N bond is located at 1.8 angstrom and vanishes faster than the old repulsive curve.

8.3. Results

8.3.1. Free Energy plots in QM/MM

Here are the 2D representations of the free energy plots obtained from different QM/MM calculations. From the fig 8.5(A) it is observed that there is no product minimum, only a broad higher energy region appears near 2.3 P-N distance which indicates P-N bond in the final structure is longer than a normal P-N bond, which is expected 1.8 A and the reaction energy therefore is also not trustable. This phenomenon correlates with the behaviour of the DFT-LDA and DFT-GGA approaches. Being based on the PBE functional, 3OB parameters seems to reproduce some DFT-PBE errors. This takes us to reproduce a reference calculation to compare it with, fig 8.5(B) is the QM/MM calculation results from

B3LYP/aug-cc-pVTZ method. It appears product minima appears at P-N distance 1.8 A and P-O distance 3.5. There is also an indication of five-fold phosphorus transition state pathway (SN2 like transition state) in the mechanism which was missing before. From fig 8.5(B) it is estimated that the reaction enthalpy of this particular model reaction is around 40Kcal/mol and the reaction barrier is around 34 Kcal/mol.

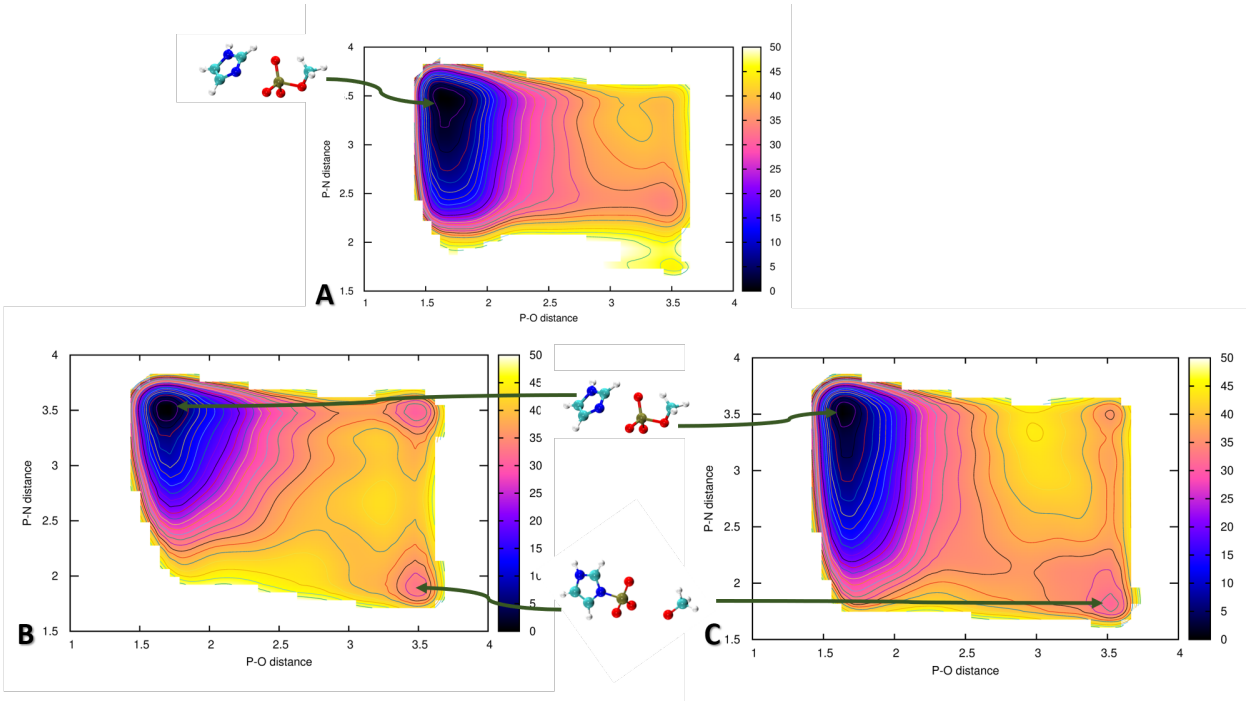

Figure 8.5.: **Free Energy Surfaces obtained from different QM/MM calculations are shown here**:A)FES from QM(DFTB)MM using old 3OB parameter, B)FES from QM(DFT)/MM using B3LYP/aug-cc-pVTZ, C) FES obtained from reparametrised P-N pair potential. Each contour line represents 2Kcal/mol,P-N, P-O distances are given in Angstrom, Energy bar is in Kcal/mol

After the reparameterisation we applied the new parameters to the same reaction. fig 8.5(C) represents the FES plot from the new parameter and it turns out product minima appears in accordance with our reference B3LYP QM/MM reaction enthalpy is now around 32 Kcal/mol and the reaction barrier now becomes 38 Kcal. We can say the re-parameterisation is able to reproduce B3LYP accuracy with an error of 2Kcal/mol.

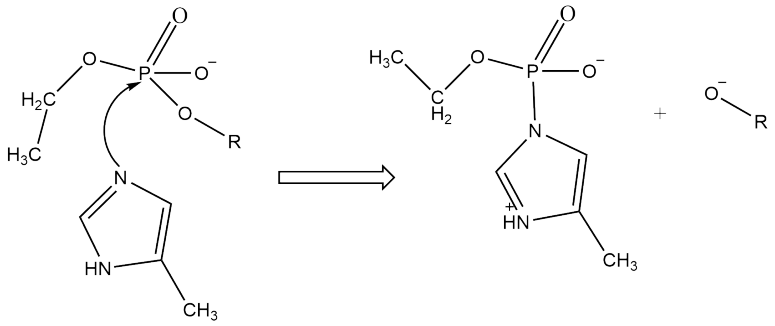

Figure 8.6.: Benchmark reaction with imidazole Nitrogen species as nucleophile, R is different leaving group based on varying electron donating power

8.4. Benchmark

For benchmarking our new SRP, we chose two different reactions 8.6, 8.7, mimics the same chemical environment of Phosphorus. These reactions are also SN2 like and proceeds through penta-coordinated phosphate, hence make suitable choice to assess accuracy of our new parameter. Two reactions has two different nucleophile, 8.7 has SP^3 nitrogen species as nucleophile, thus stronger one and 8.6 imidazole nitrogen as nucleophile, thus weaker one. Now it is fair to declare that, we don't know whether this reaction occurs in actual experimental condition or not, but reactants and products are fairly stable species in vacuum. Thus computing single point energies are easier.

The idea is substituting -R (leaving group) with different electron donating and withdrawing groups, such that it will change the local electron density on phosphorus (electrophilic center) will change for every reaction. Therefore, chemical environment of the reaction will be slightly different every time and thus the reaction energy. We want to compare this reaction energies (reactant energy - product energy) with B3LYP and DFTB (using 3OB + new PN SRP + OP-hyd SRP).

Leaving Group (R-)	RE in DFTB (kcal/mol)	RE in B3LYP (kcal/mol)
R = C(CH3)3	56.90	76.30
R = CH3	61.78	82.51
R = Ph	34.82	44.33
R = COOH	20.15	35.78
R = CN	18.95	41.77

Table 8.2.: Reaction energies (RE) shown in both DFTB and B3LYP with different leaving group for the first reaction 8.6

Overall trends in both of the reaction schemes are same. Reaction energetics are same in both DFTB and B3LYP, which is RE (Reaction Energy) is negative. With stronger electron donating group (EDG) RE is more and with stronger electron withdrawing group (EWG) is less. DFTB seems to favour the products more than B3LYP systematically in both

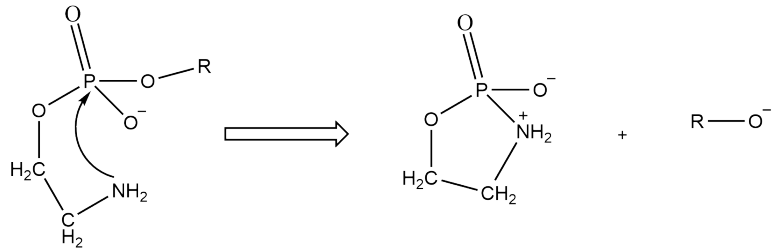

Figure 8.7.: Benchmark reaction with SP^3 Nitrogen species as nucleophile, R is different leaving group based on varying electron donating power

Leaving Group (R-)	RE in DFTB (kcal/mol)	RE in B3LYP (kcal/mol)
R = C(CH3)3	74.59	101.25
R = CH3	79.26	86.64
R = Ph	52.72	76.33
R = COOH	38.45	71.24
R = CN	36.44	51.05

Table 8.3.: Reaction energies (RE) shown in both DFTB and B3LYP with different leaving group for the second reaction 8.7

benchmark system, however the difference ranges from 10 to 23 kcal/mol. This could be a consequence of the introduced overbinding.

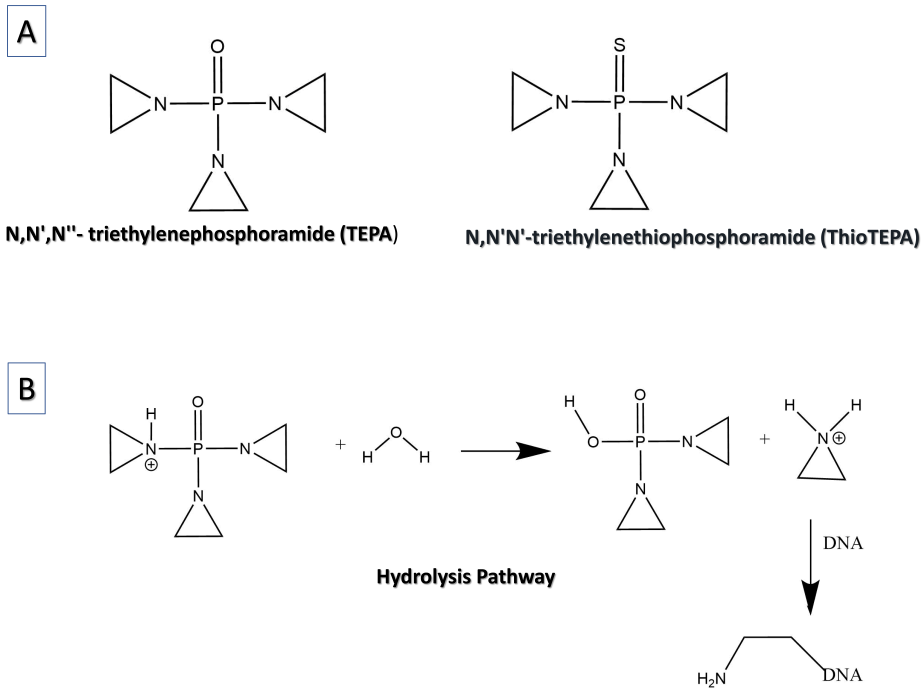

Figure 8.8.: Structure of Thio-TEPA and TEPA, serve as pro-drug of azidirinium ion

Hydrolysis of TEPA

N,N,N-triethylenethiophosphoramide (Thiotepa)[175, 186, 75] and its oxo analogue (Tepa)[86] (its major metabolite) are common drugs to exhibit antitumor activity. It is one of the oldest chemotherapeutic drugs with continuing clinical utility, often used in high dose combination regimens for breast, ovarian, bladder cancers and other solid tumors. Although the fact is known that it alkylates DNA (Guanine) mechanism is still not very clear. Metabolic studies of Thiotepa reveals its oxo analogue (Tepa) as its major metabolite, formed after oxidative desulfuration of Thiotepa in the liver by cytochrome P450[175]. Thiotepa and Tepa have been classified as trifunctional alkylating agents (contains three aziridinyl functionalities) that are proposed to induce cancer cell death by formation of cross-links within DNA [125, 19, 121].

We are interested on the particular hydrolysis because the reaction resembles the very similar chemical environment of reaction (histidine phosphorylation) we want to study in chapter 9 for which we made the special reaction parameter in this work. This makes TEPA hydrolysis a very good case study to estimate the performance of the new P-N repulsive potential. This study will also give us an idea of transferability of the SRP to other similar reactions.

Thiotepa and Tepa are the prodrugs for aziridine or aziridinium ion, which act as actual alkylating agent. In vivo and vitro studies unveils alkylation of DNA (Guanine) by this drug indicates several pathways [111, 110, 175], but it still remains unanswered which pathway is effective. The first step of any alkylating drug activity is regarded to be interaction with DNA, either directly (ThioTEPA) or after metabolic activation (TEPA).

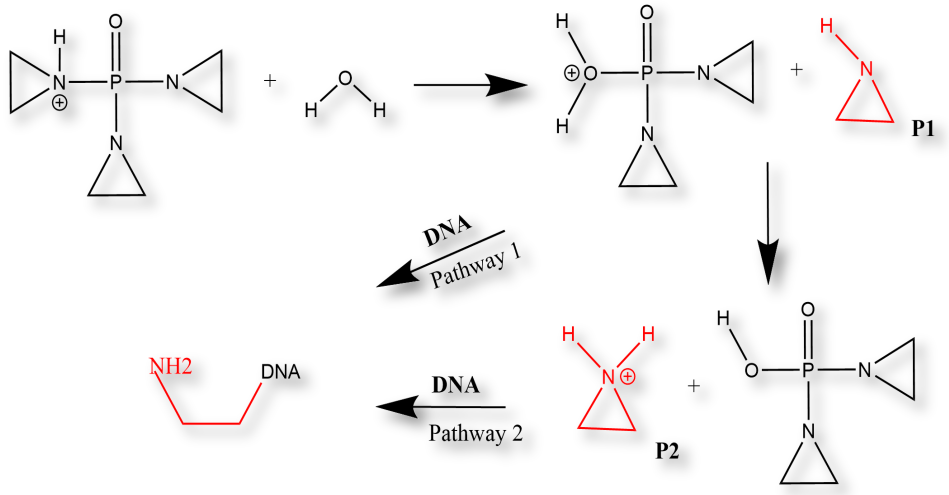

Figure 8.9.: Two different possibility: P1 or P2 which one attacks DNA?

According to with Miller's theory, sites, which potentially could interact with electrophilic species are the DNA nucleophilic centers: nitrogen and oxygen atoms of pyrimidine and purine bases [17]. From these two information we can identify potential electrophile could be azidirine molecule or aziridinium ion extracted after hydrolysis of TEPA/ThioTEPA molecule. Therefore here in this study we focused on the pathway which generates independent azidirine molecule and aziridinium ion, which later alkylate N7 of Guanine of cancer DNA [70].

Experimental studies shows that free aziridine as a weak base. The reactant becomes more stable because of hydrogen bonding between protons on the aziridinuim ion with oxygen and nitrogen atoms in Guanine. We investigated both possibilities 8.9, which could result a stable reaction with Guanine molecule. Earlier computational studies [169, 168, 90] on this particular hydrolysis reaction identified several features of the reaction, structure of the transition state, reaction enthalpy etc. but the complete mechanism of the hydrolysis is still not shown. In this study, we tried to explore the mechanism.

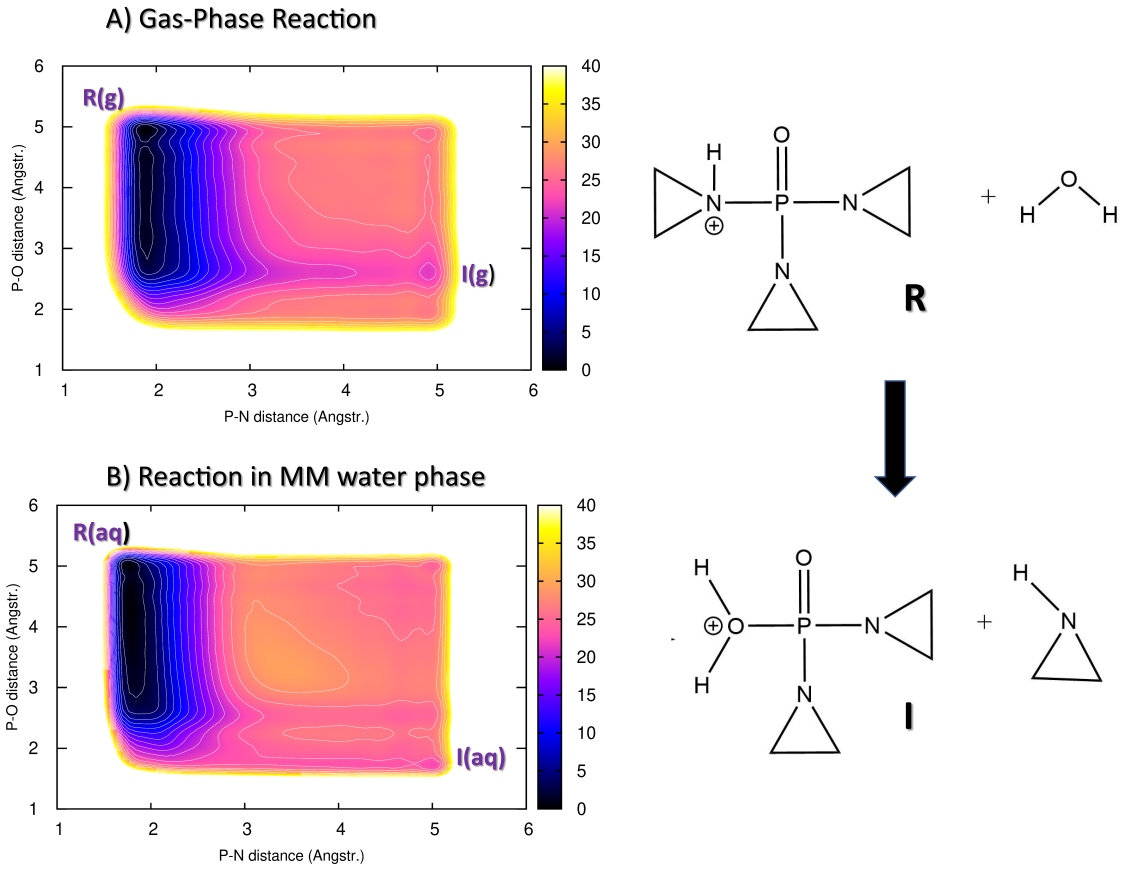

Figure 8.10.: Water attackes on the Phosphorus of TEPA molecule to form an unstable product, reaction is shown both in gas-phase and in explicit water

First we tried to simulate the first step of the reaction 8.9 leading to produce aziridine molecule (P1). We used 2D metdaynamics using P-N distance and P-O distance as collective variable with very small bias height (0.2Kj/mol) in gas-phase (only QM MD) using our new P-N repulsive potential and O-P hydrade SRP along with other 3OB parameters. PMF of the simulation is shown in fig 8.10(A). P-N bond appears to be at 1.8 Angstrom make a broad minimum. The final product (aziridine molecule) comes out to be highly unstable.

After that we repeat the same simulation in explicit water environment (QM/MM). PMF is shown in fig 8.10(B), but the conclusion did not change product region remained in higher energy suggesting the reaction is highly unfavourable.

Now we simulated the whole reaction 8.9 step1 + step2 leading to aziridinium ion (P2) as final product. We performed again 2D Metadynamics using 2CVs 1. P–N bond hydrolysis = P-N distance - P–O distance (which means when this CV is negative P-N bond exist and when its is positive P–N bond is broken P–o bond forms) and 2. N–H distance to describe proton transfer to aziridine ring. Metadynamics is carried out in explicit water environment (QM/MM simulation). PMF is shown in fig 8.11.

From the PMF 8.11 we can observe all the important structures. We can see aziridinium ion (P in 8.11) is now the global minimum of the reaction more stable than the TEPA (R)

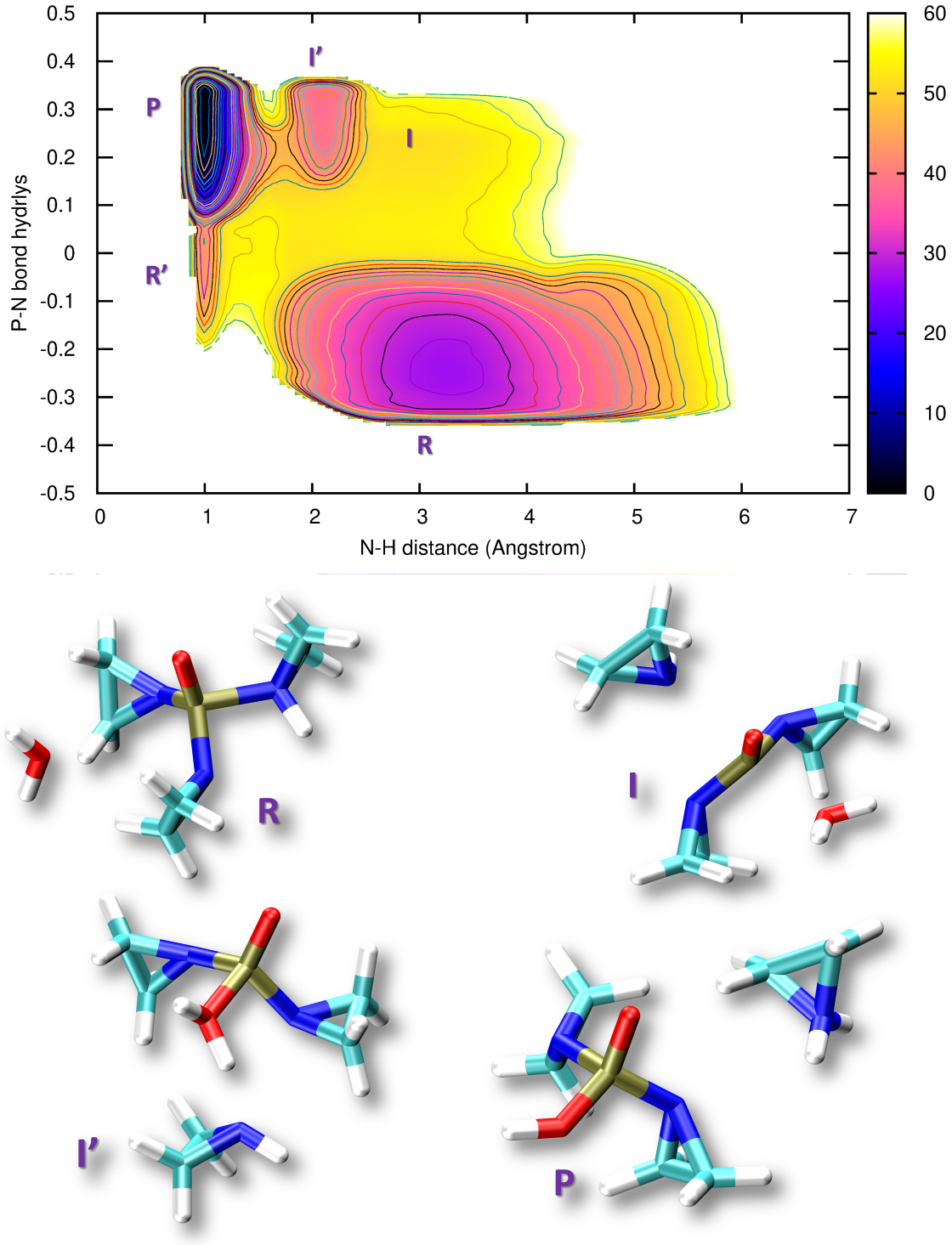

Figure 8.11.: Free energy surface of the complete mechanism of releasing final Aziridinium ion

itself. The previous product aziridine (I) is highly unstable. Reaction energy of (R-P) is -30 Kcal/mol (reaction enthalpy is negative).

8.5. Conclusion

In this chapter we have demonstrated the reparameterisation procedure for repulsive potential in DFTB3. This is the traditional way to improve DFTB3 energies and structures for a specific reaction, called SRP. Current DFTB3 has limited transferability for complex phosphorus chemistry at the level of accuracy in energetics, which is required for detailed mechanistic investigations. The SRP we developed in this chapter to improve P-N interactions in imidazole specific phosphorylation reactions, which is important step in Histidine Kinases (Histidine Phosphorylation). These SRPs are used along with other 3OB parameters in QM/MM simulations. This new re-parametrised P–N parameter is able to reproduce reaction energy and reaction enthalpy pf B3LYP accuracy with an error of 2Kcal/mol. Although these SRPs are clearly not a satisfactory and long term solution, this makes people to develop various automated way to reparametrise specific parameters [104]. Also, it makes room for various Neural-Network algorithms to serve in this problem [66]. Although this parameter is made specifically for the purpose of Histidine phosphorylation, we further tested it on other similar gas-phase reactions to evaluate accuracy and performance. We further extended our study investigate cancer drug (TEPA) hydrolysis using the new parameter and demonstrated the full mechanism of the hydrolysis for the first time.

9. Mechanism of Autophosphorylation in Cis-Activated WALK Histidine Kinase

The work in Chapter 9 was done in collaboration with Fathia Idiris and Alexander Schug [82]

9.1. Introduction

Histidine Kinases (HK) are necessary part of Two component systems (TCS), one of the major signal transduction pathways exist in bacteria. They are involved to regulate the bacterial response to a variety of environmental factors like temperature changes, changes in pH, change in pressure or cellular signals [174]. The individual components are the sensor histidine kinase (HK) that detects the signal and the response regulator (RR) protein that coordinates the response, most commonly by acting as a transcription factor (see Fig. 1.1). These two proteins communicate each other via histidine to aspartate phosphoryl-group transfer. Based on domain architectures, evolutionary origin and activities there are numerous variations of TCS [94, 189]. While TCS are employed by some eukaryotes, they are notably absent from the animal kingdom. That, paired with their importance to bacteria makes these enzymes promising targets for developing novel compounds that selectively inhibit the growth of bacteria or suppress virulence. For instance, waldiomycin, an angucycline antibiotic, inhibits the HK activity of WalK [83, 126] in *Staphylococcus aureus*, a human pathogen responsible for a variety of acute and chronic diseases [182, 171, 139]. The molecular signal of this system is still unknown but emanates from the bacterial cell wall [20]. In general, the WalRK system has garnered significant experimental attention since it is conserved acroos Gram-positive bacteria of the order Firmicutes where it has been shown to be essential for viability in a variety of different species of bacteria.

The structural properties of HKs differ, they all have at the C-terminus as a conserved kinase core (~450 amino acids) consisting of the homodimeric dimerization histidine phosphotransfer (DHp) domain and the ATP-binding catalytic domain (CA).

HK exhibits kinase activity through a interplay of conformational change and reaction in a cyclic manner (discueesd in the Introduction 1.5). Upon signal detection at the trans-membrane part, the conserved core adopts an asymmetric conformation such that one of the two subunits of the homodimer is kinase active while the other remains inactive. In the kinase inactive conformation, ATP can enter the CA domain and the binding site of the DHp domain is accessible to a RR for phosphoryl-group transfer. In the kinase active conformation, the RR cannot bind the DHp domain. Here, the gamma-phosphoryl group of the bound ATP of one CA is positioned in close proximity to a specific conserved phosphorylatable histidine of DHp. Two different auto-phosphorylation mechanisms are observed in individual HKs: cis- and trans-phosphorylation, already discussed in Introduction chapter 1.6.

In this chapter we focused in cis-phosphorylation where the ATP from the CA domain phosphorylates its own DHp domain within the homodimer. As soon as the histidine is phosphorylated, transfer of this phosphoryl group to an aspartate of a bound RR (Response Rehulator) for communication between the two proteins is possible. Bifunctional HK (e.g., EnvZ) also function as phosphatase for the RR and therefore catalyse the hydrolysis of the phosphoryl group [84]. The activation and inactivation mechanisms of the protein are reviewed in detail in Ref. wang2013mechanistic.

Once the protein gets activated it triggers a phosphoryl transfer reaction. Reaction takes place in two steps first the gamma phosphate transfers from ATP to a conserved histidine residue in the DHp domain and followed by proton transfer from histidine to suitable proton acceptor 9.1. Previous computational studies tried to explore the mechanism of the phosphoryl transfer reaction in WalK kinase [127] CpxA kinase [151] [113] reported different barriers (mostly upper bound) and in-depth exploration of the two-step mechanism was not possible due to limited sampling. Moreover most of these studies have been carried out in Physiological pH. Patricia et. al. [28] has shown in their study that there is a conserved Glutamate which could act as a potential proton acceptor in the reaction on the other hand an older experimental study in 2003 [34] had shown the optimum pH of this reaction in 8.5 which raises the question of the mystery of potential proton acceptor.

This brings us to our discussion of phosphorylated-histidine product of the considered phosphoryltransfer reaction. The molecule is also known as phosphohistidine, an unstable compound mostly found as intermediate species in a long cascade of biochemical processes[181, 102, 118, 120]. The phosphate transfer potential of phosphohistidine ΔG° is also quite low, -12 to -14 kcal/mol [164]. First histidine phosphorylation was carried out by Severin and Yudelovich (1947)[149]. First phosphohistidine was extracted from mitochondria involved in ATP synthesis in citric acid cycle [54]. Boyer and coworkers prepared and characterized 3-phosphohistidine (phosphorylation at ϵ nitrogen of histidine) and 1-phosphohistidine (phosphorylation at δ nitrogen of histidine) [78] in different pH conditions and examined the stability; it turned out that both of the compounds are fairly stable in basic medium rather in acidic medium (at higher pH). Moreover it was shown in the same study that 3-phosphohistidine is thermodynamically more stable than 1-phosphohistidine by the estimation of half life.

Goal of this study is to explore the Phosphorylation reaction using state-of-the-art extended-sampling QM/MM simulations on a microsecond scale, provides sufficiently precise energetics of the reaction [36] to distinguish if the phosphoryl transfer takes place first, followed by the proton transfer, or the other way around. A crucial part of this reaction is a magnesium cation as a cofactor, found close to the reaction site in all HK proteins [185, 28, 29]. The simulations also allow us to understand role of Mg^{2+} in this reaction.

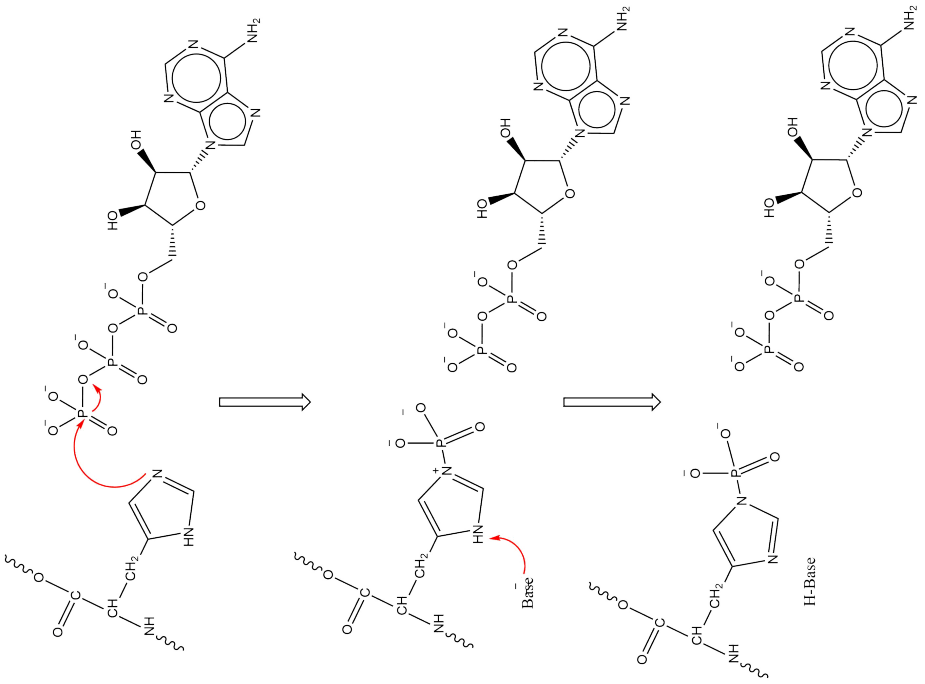

Figure 9.1.: Proposed Reaction occurs in two steps, 1) first step is the phosphoryl transfer from ATP to histidine 2) Second step is the proton transfer from phosphohistidine to suitable base

9.2. Methodology

Preparation and MM Equilibration

We started from crystal structure PDB ID 4U7O[26], which is activated WalK histidine kinase. The structure contains a non-hydrolysable ATP analogue AN2 and no magnesium present. We modelled non-terminal missing loops, and modified AN2 to ATP using UCSF Chimera [132] interfaced with MODELLER [180]. An Mg^{2+} ion was placed carefully in between the γ- and β-phosphate groups. After that, the other ATP-binding domain which, located far away from the DHp domain, was truncated to reduce the size of the system. Finally, the biomolecular complex was enclosed in a periodic box sized ca. $8 \times 8 \times 8$ nm^3, which was filled with water and electro-neutralized by the addition of nine sodium counterions. The density of the system was 1014 kg m^{-3}.

The AMBER99SB-ILDN force field was used to describe the protein [105], while the parametrisation of ATP from Ref. meagher2003development was employed. The solvent was represented with the TIP3P water model [87] and Åqvist's parameters for the counterions [4]. The electrostatic interactions were treated with PME [38, 46], where the short-range contribution was cut-off at 1 nm. The Lennard-Jones interactions were cut-off at 1 nm. All of the QM/MM MD simulations used the leap-frog integrator [35] with a time step of 1 fs, while all bonds involving hydrogen atoms were constrained with LINCS [72].

First, the system in the entirely MM representation was energy minimised with steepest descents. Then, it was equilibrated for 10 ns maintaining the temperature of 300 K by means of the Bussi thermostat [24].

QM/MM Preparation

Two different QM/MM setups were prepared Using the final structure from the MM equilibration. The QM region was introduced, consisting of the reaction center and its nearest neighbourhood:

- **System 1 – Glu392 considered as the final proton acceptor:** The QM region contains the side chains of His391, Glu392 and Asn541, the ATP molecule, the Mg^{2+} ion and 5 water molecules (70 atoms in total).
- **System 2 – a hydroxyl ion considered as the final proton acceptor:** The QM region contains the side chains of His391 and Asn541, the ATP molecule, Mg^{2+} ion, 5 water molecules, and an OH^- ion created by removing a proton from a water molecule (56 atoms in total).

The QM region was treated with the semi-empirical density-functional method DFTB3 [59] employing the 3OB parameter set [60] augmented with a special parametrisations for the pair interactions P–O and P–N [62]. The QM–MM interactions were treated by means of electronic embedding, which involved our PME implementation [97]. The MM region was described with the same force fields as employed in the preceding MM equilibration, as specified above. All of the MD simulation parameters were kept also, and the prepared QM/MM system were equilibrated at 300 K for 1 ns. The QM/MM simulations were performed using a local version of GROMACS [71, 1, 96] interfaced with PLUMED [170] and a local version of DFTB+ [77, 95].

QM/MM Free Energy Calculations

Potentials of the mean force were generated by means of multiple walker [140] two-dimensional metadynamics [25] employing 96 individual simulations (walkers). An initial phase of 47 ns was run with a constant Gaussian height. The second phase involved a well-tempered metadynamics protocol [10].

Collective Variables

Two collective variables were employed in the metadynamics simulations as follows, see also Fig. 6 in the main text.

- **Phosphoryl transfer:** O–P–N antisymmetric stretch, which is the difference of the distances: P(γ-phosphate of ATP)–Nε(His391) – P(γ-phosphate of ATP)–O(β-phosphate of ATP)
- **Proton Transfer:** N–H–O antisymmetric stretch, which is the difference of the distances Nδ(His391)–Hδ(His391) – Hδ(His391)–O(OH^-/Glu392)

Metadynamics

In the normal metadynamics phase, the height of the biasing Gaussians deposited was 1.2 kJ mol^{-1}, and their width was 0.02 nm in both dimensions. In the consective well-tempered metadynamics phase, a bias factor of 80 and 70 was considered in the simulations conseridering Glu392 and a hydroxyl ion, respectively, as the final proton acceptor. In all cases, the period of bias deposition was 500 steps, and the biases were communicated between the individual walkers every 1000 steps.

Forcefield parameterisation of phosphorylated histidine residue

After the reaction take place, we get phosphohistidine and ADP as final products. Though we have forcefield parameters available for ADP [114] but for phosphohistidine there is no parametrs. Further in future if we want to investigate the next chemical step the the phospho-relay cascade, which is phosphoryl transfer from HIS of HK to ASP of RR 1.1, or step II to III in auto-kinase cycle 1.5, forcefield parameters of phosphohistidine will be essential for classical MD and enhanced sampling. The very reason why we parametrised the forcefield parameters of phsphohistidine utilizing antechamber tool [178] using RESP charge fitting method [12].

Results

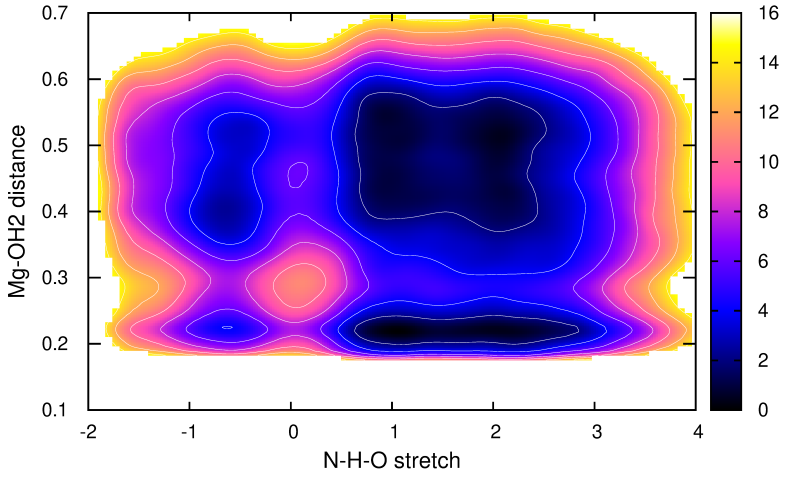

Figure 9.2.: PMF from 2D metadynamics using N-H-O stretch in Angstrom X axis (for proton transfer) and distance of Mg with sixth water molecule in Y axis in Angstrom, other five coordinate bonds of Mg is restrained

.

Phosphoryl Transfer from ATP to His391

As soon as the active state is established, an autophosphorylation reaction takes place, consisting of a transfer of the γ-phosphate group from ATP to the Nε atom of His391 in the DHp domain, followed by a deprotonation of His391. The mechanism of this complex chemical reaction cannot be studied by experimental means in any feasible way, while it poses a serious challenge for a computational investigation, requiring an approach that is both, sufficiently accurate and efficient. The choice taken in this work is a QM/MM multiple-walker metadynamics simulation employing the semi-empirical density-functional approach DFTB as the quantum chemical method. This easily parallelizable protocol makes it possible to reach microsecond sampling, while the accuracy approaches 1 kcal/mol due to a reparametrization of the P–N repulsive potential of DFTB. The QM region consisted of the side chains of His391 and Asn541, the ATP molecule with the coordinated Mg^{2+} ion, five nearby water molecules and a suitable proton acceptor, see Fig. 9.3A & B. Two different simulations were performed, differing in the identity of the proton acceptor: the side chain of Glu392 in system 1, or an OH^- ion as proton acceptor in system 2. The metadynamics simulations involved two collective variables (CV) to describe the progress of the chemical reactions and express the potentials of the mean force (PMF): The O–P–N antisymmetric stretch [i.e., difference of the distances Pγ(ATP)–Oβ(ATP) and Pγ(ATP)–Nε(His391)] describes the transfer of the phosphoryl group, while the N–H–O antisymmetric stretch, [i.e., the difference of distances Nδ(His391)–Nδ(His391) and Hδ(His391)–O(proton acceptor] describes the transfer of the proton to the acceptor, see Fig. 9.3B. For illustration, a negative O–P–N means that the γ-phosphate group has transferred from ATP to His391, and a positive N–H–O denotes a completed proton transfer to the acceptor.

Position of the Magnesium Cation

The action of kinases generally requires the presence of a magnesium cation as a cofactor [187, 185]. Since the crystal structure used here as the initial structure included a non-hydrolyzable ATP analog and no magnesium, it was necessary to proceed with care and find the right position of Mg^{2+}. In order to do so, the PDB was searched for both active and inactive structures of wild-type and mutant kinases that do have a coordinated Mg^{2+} cation. The cation was always found in a very similar position in the ATP binding domain in the structures of different HK proteins, assuming a coordination to an oxygen atom of the γ-phosphate group of the bound ATP. Therefore, to complete the preparation of the initial structure for QM/MM simulations, several structural models were created, featuring an Mg^{2+} cation in slightly different positions close to the γ-phosphate of ATP. Importantly, during an equilibration period of QM/MM simulations, the Mg^{2+} cation was always found coordinating with the same six oxygen atoms (one each in the side chain of Asn541, in the γ-phosphate, β-phosphate and α-phosphate of ATP as well as in two water molecules), see Fig. 9.3C. That eventually provided a suitable initial structure to start the metadynamics simulation.

Both the reactant state and the final product of the reaction feature the Mg^{2+} ion with a stable coordination sphere containing six ligands. In the course of the metadynamics

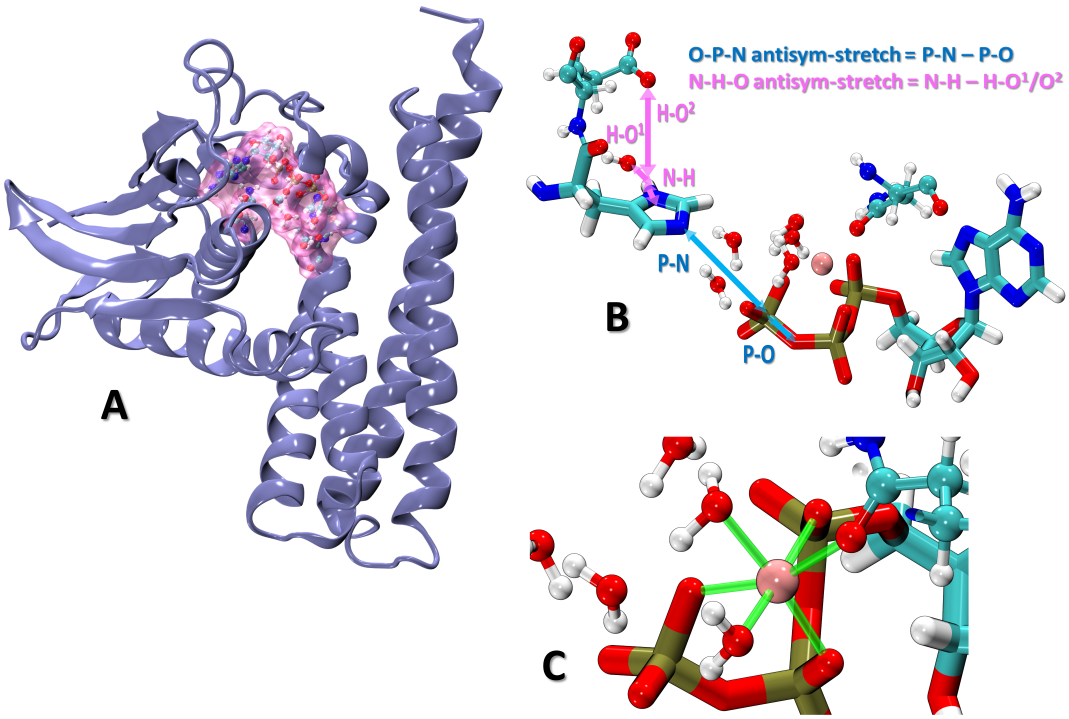

Figure 9.3.: Model of WalK used as the initial structure for QM/MM metadynamics simulations of autophosphorylation. A) The active structure adopted from PDB ID 4U7O with the non-reacting ATP-binding domain truncated; location of the reaction center highlighted in pink. B) The QM region covering the reaction center; the antisymmetric stretch CVs presented in blue and pink. C) Coordination sphere of the magnesium cation in the reactant structure.

simulation, the coordination sphere of Mg^{2+} oscillates between five and six ligands. Water molecules were found to engage in strong hydrogen bonding with the β-phosphate group of ATP and the phosphohistidine, and that is why they showed the propensity to at times decordinate from Mg^{2+}. We further investigated the coordination sphere of Mg^{2+} using an additional QM/MM metadynamics simulation, which included additional water molecules in the QM region to ensure that any nearby water molecule is able to fill up the vacancy in the coordination sphere. The resulting PMF in Fig. S2 shows a negligible energy difference between the coordination numbers of five and six, as well as a very low barrier of 2 kcal/mol to the un- and re-binding of the sixth ligand (a water molecule).

Nature of the Transition State

The nature of the transition state of the phosphoryl transfer reaction was also analysed. To this end, we ran another QM/MM metadynamics simulation using a pair of CVs designed to describe the phosphoryl transfer: the distances Pγ(ATP)–Nε(His391) and Pγ(ATP)–Oβ(ATP). The resulting PMF is shown in Fig. S3. The transition state, which lies 8 kcal/mol above the reactant exhibits a five-fold hypervalent state of the phosphorus atom, shown

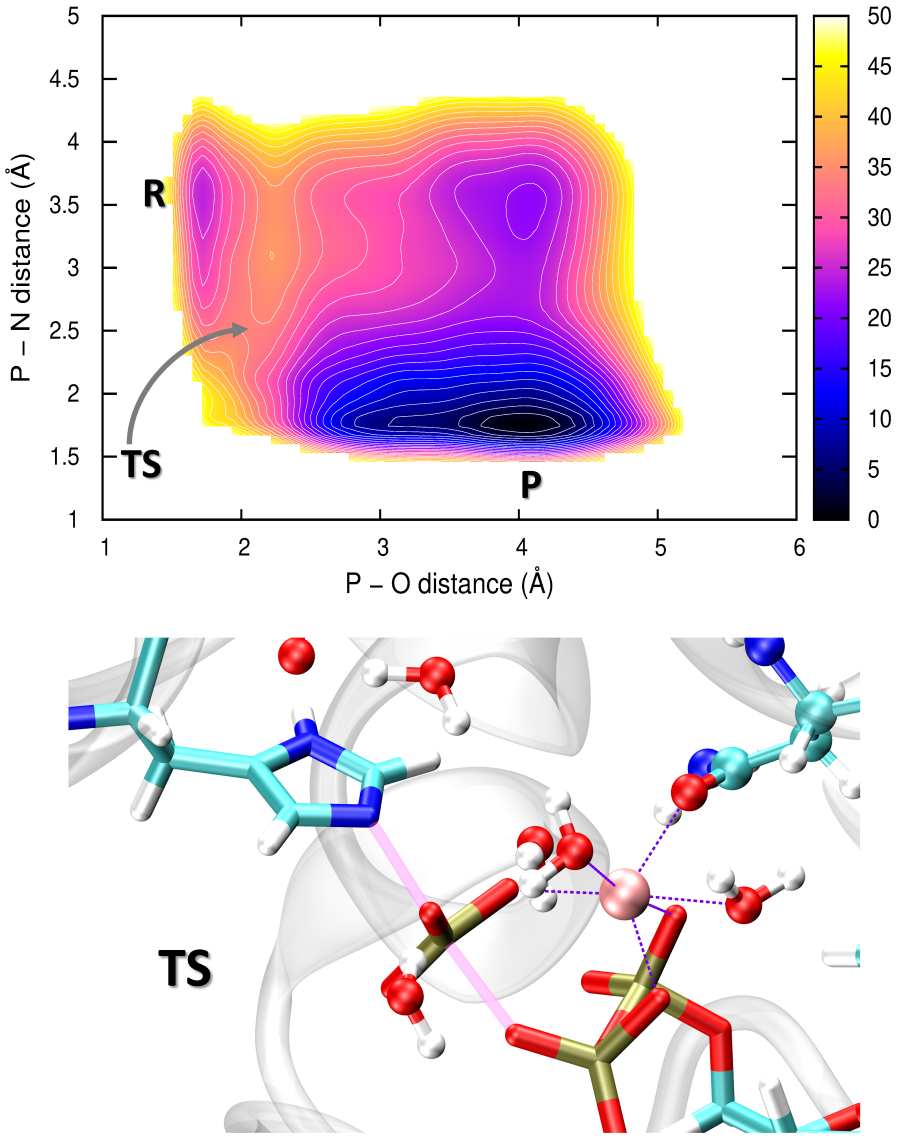

Figure 9.4.: Results from 2D QM/MM metadynamics simulation using the distances P–N and P–O as CV. Contour lines are at 2Kcal/mol, energy units are in Kcal.Top: Resulting potentials of the mean force. Bottom: Representative transition state structure from that simulation; highlighted are P–N and P–O distances (thick solid line) and the six coordination bonds to the Mg2+ ion (thin dashed lines)

.

in Fig. S3 also. Interestingly, the bonding to the phosphorus atom is asymmetric, with the P–O distance of 2.17 Å being markedly shorter than the P–N distance of 2.53 Å.

The Chemical Step Is Exergonic And Base-Dependent

We modelled the reaction considering the conserved Glu392, which is in a close contact with His391, as the proton acceptor. A good convergence of the resulting free energy surface (FES) shown in Fig. 9.5 (left) is indicated by the analysis presented in Fig. S4. Passing over a barrier of 8 kcal/mol, the phosphoryl transfer leads to a protonated phosphohistidine intermediate, lying 20 kcal/mol below the reactant. Then, a nearly barrierless proton transfer from His391 to Glu392 leads to a final product that lies 13 kcal/mol above the protonated intermediate. This indicates that a stronger base is needed as the final proton acceptor to make the reaction sequence exergonic.

On the other side, an OH^- ion was placed near the Hδ atom of His391, which is the proton to be transferred. A new QM/MM metadynamics simulation was performed, see Fig. 9.5 (right) for the resulting free energy surface, and see Fig. S5 for the analysis of convergence. The final product lies less than 1 kcal/mol below the intermediate, and is a global minimum of free energy now. The energy barrier to the reaction sequence of 8 kcal/mol is identical to the previous case where considering a glutamate as the proton acceptor. The reason for this is that the higher barrier applies to the phosphoryl group transfer from ATP to histidine, which is exactly the same process in both cases. The subsequent proton transfer passes over a much lower barrier of 4 kcal/mol.

Stability of the final product

After the reaction took place, we were curious about the stability of the final product, which is protein with phosphohistidine residue. Since it is a non-standard residue we parameterised the forcefield parameters for phosphohistidine, as we discussed above. According to [177] after the autophosphorylation happens, histidine kinase again moves back to kinase inactive state and the ADP leaves from the CA domain. In order investigate the hypothesis we ran two classical simulations, one with ADP and Mg^{2+} ion and other one without ADP and Mg^{2+} ion. RMSD of both simulations after 100ns is shown in 9.7. Protein without ADP and Mg^{2+} ion has got two RMSD states indicating conformational rearrangement might be a reason. On the other hand the structure with ADP and Mg^{2+} ion has single RMSD state, indicating final product with ADP and Mg^{2+} ion is quite stable. The simulation with ADP and Mg^{2+} shows Mg^{2+} ion still coordinating with the phosphate of phosphohistidine and full filled all six coordinations. Overall we can say our parameterised forcefield parameters for phsphohistidie is quite successful and can be used in future simulations.

9.3. Discussion

Regarding the chemical step of histidine phosphorylation, we performed two different QM/MM metadynamics simulations of the process combining the phosphoryl transfer

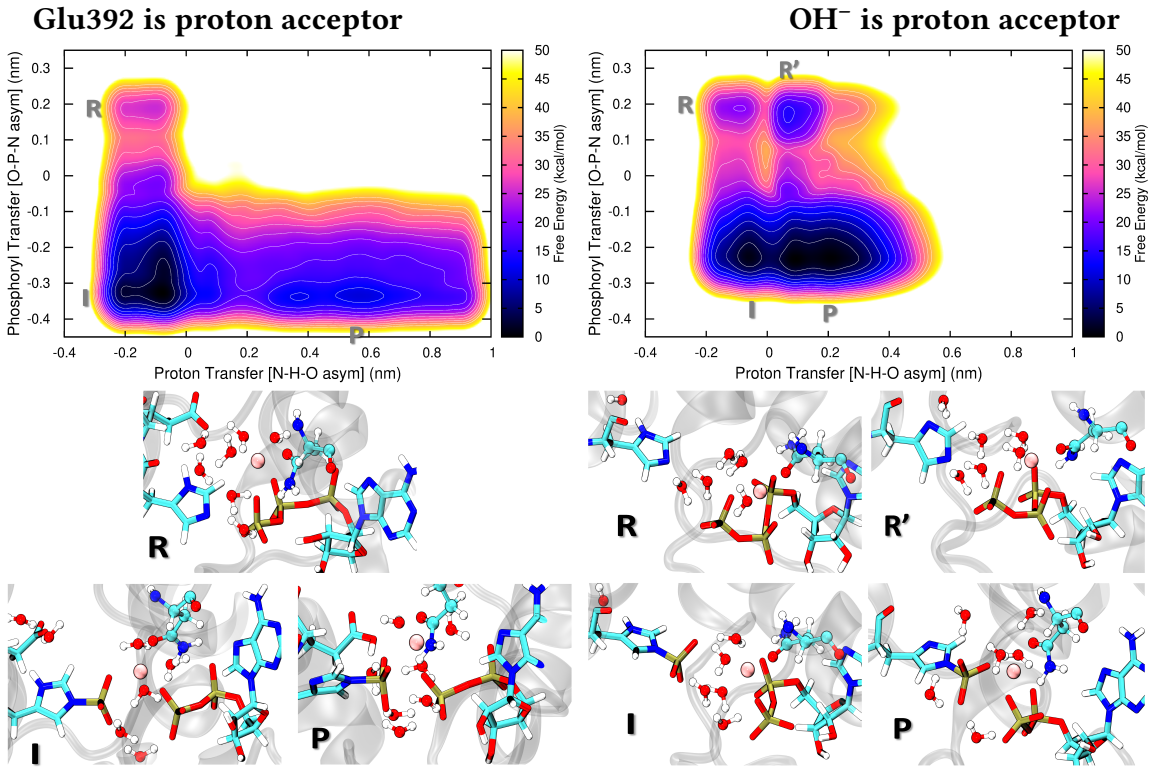

Figure 9.5.: Results from the 2D QM/MM metadynamics simulations of autophosphorylation of WalK, using the antisymmetric stretches N–H–O and O–P–N as CVs. There are two different simulations: one involving the side chain of Glu392 as the proton acceptor, and the other with an OH^- ion playing that role. Top: Potentials of the mean force for the phosphorylation reaction. Bottom: Representative structures from the : R – reactant, R' – intermediate (His391 is deprotonated before its phosphorylation takes place), I – intermediate (protonated phosphorylated His391), P – final product (deprotonated phosphohistidine). The free energy is color-coded, and the spacing of contour lines is 3 kcal/mol

.

from ATP to His391 and the proton transfer from His391 either to the initially deprotonated Glu392 or to an OH^- anion placed near His391. The resulting free energy surfaces converged after simulations were extended to 1 µs. We observe a free energy barrier to the chemical reaction of 8 kcal/mol, independent of the identity of the proton acceptor, leading to a stable intermediate represented by protonated phosphorylated His391.

This value should be added to the free energy change accompanying the conformational transition obtained from classical simulations (work done by Fathia Idris [82]), which is however most likely overestimated. The experimentally reported catalytic rate of 0.027 min^{-1} [34] corresponds, using the transition state theory expression for rate = $kT/h \cdot \exp[-E_A/kT]$, to an activation energy (barrier) of 22 kcal/mol. Apparently, the energy barrier of the chemical step (8 kcal/mol) represents just a small part of the overall barrier. Therefore, the conformational transition appears to be the rate limiting step of

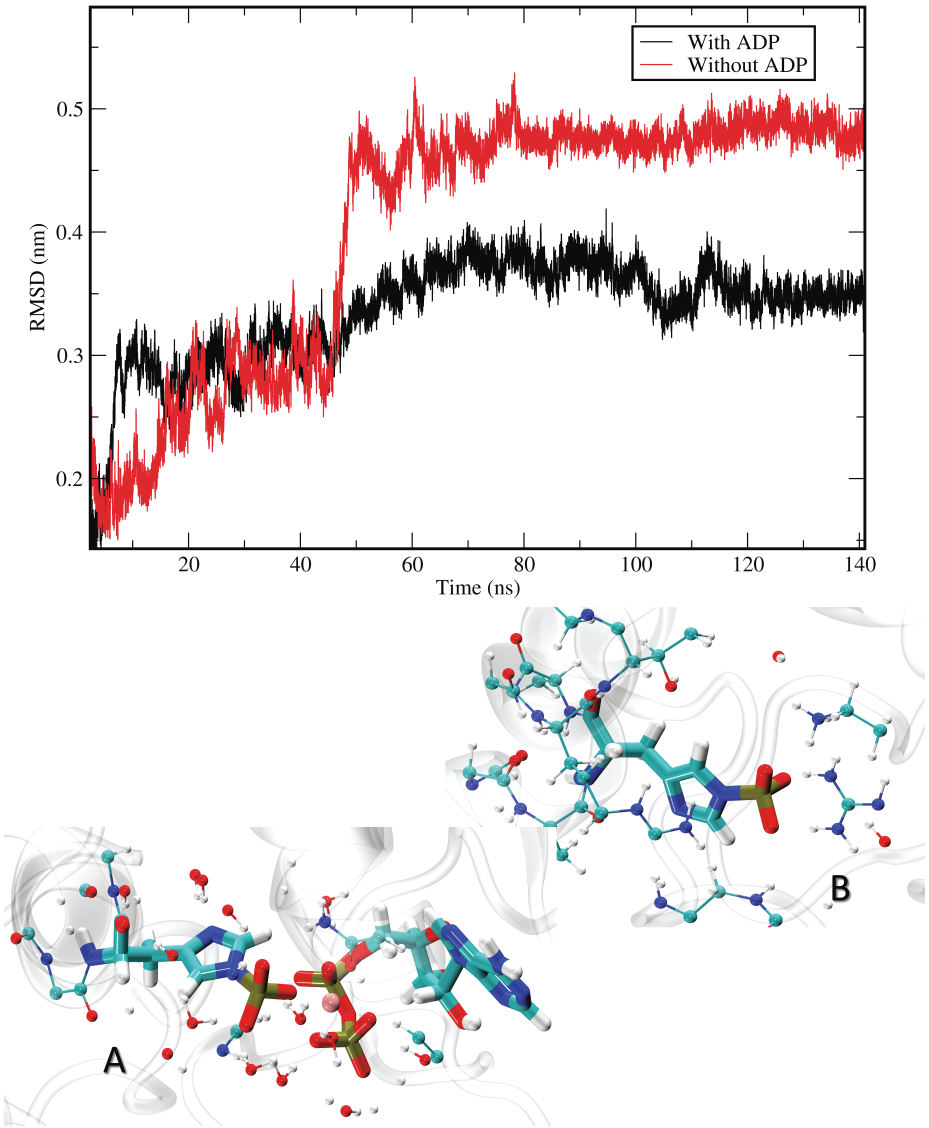

Figure 9.6.: RMSD plot of the protein shown above after the autophosphorylation reaction, obtained from a 100ns free classical MD simulation. One is ADP bound structure and other one is without ADP

.

the whole catalytic activity. Note that this corresponds to the situation in an *in vitro* experiment, whereas the genuine *in vivo* process is being triggered by an energy input from the sensory domain reacting to a stimulus, and this also effectively reduces the barrier.

When terminating the reaction with the protonated Glu392, the product lies above the protonated pHis intermediate in free energy. This suggests that a stronger base has to be present as a real final proton acceptor, and this situation was considered in the second QM/MM metadynamics simulation, which had an OH^- anion located near His391. With this setup, the deprotonated phosphorylated His391 is a stable product, the species lying the lowest in free energy, 20 kcal/mol below the initial reactant. Therefore, for the phosphorylation to occur more readily, there has to be a strong base present as a final proton acceptor somewhere in the system. It is not necessary for it to be directly near the histidine, as long as a proton transfer pathway between the histidine and the proton acceptor is available. No matter what the pathway is like, the reaction energy will likely be similar; note that proton transfer may occur along rather long "water wires" exhibiting low energy barriers [143].

We like to emphasize that the two simulations performed with different proton acceptors serve different purposes: The simulation with Glu392 as the acceptor shows the phosphorylation followed by proton transfer to Glu, which is likely the first step of the potentially complex deprotonation of the histidine. The other simulation with OH^- representing a general strong base aims to estimate the reaction energy of the entire process, involving a real final proton acceptor. We do not claim that an OH^- ion is genuinely present in close proximity to His391; most likely, it is not. The former simulation rules out the deprotonated unphosphorylated His391 (R') as a viable intermediate, as it was suggested earlier to be an "activated" phosphoacceptor [28], because it lies too high in energy in the realistic pathway proceeding via a protonated Glu392. The latter simulation in turn reveals that the chemical step of the autophosphorylation is in fact a down-the-hill process.

Clausen et al. showed that the rate of autophosphorylation in WalK increases with increasing pH, indicating the need for a strong base present as a proton acceptor [34]. In agreement with that, our simulation performed with and without an OH^- present provide a means to quantify the effect of a generic strong base available in the system, on the energetics and kinetics of the reaction. Prior computational studies of WalK and CpxA involved computationally considerably more costly DFT methods, which limited these studies to (sub-nanosecond) time scales insufficient for treating complex proton transfers and making the resulting free energy surfaces undersampled [127, 113]. The main finding on WalK was the "tight coupling" of the chemical step with the preceding conformational transition, whereby the protonation of His391 is prevented, which would otherwise hinder phosphorylation. Our vastly increased sampling in the current study unveils quantitative free energies surfaces covering the reaction mechanism in great detail.

Comparing with previous studies, study of oliveri et. al over-estimated the reaction barrier [127], this could be due to use of poor basis set (speeding up QM/MM simulation) and less sampling. We also argue that short simulations (less than 1 ns) [127, 113] for such complex multistep reaction could introduce large error in computing free energy reaction barrier and reaction energy. In our study, the free energy of the simulation converged in 1 μs in which it is eventually revealed that the intermediate and product minima are deeper

than reactant minimum the fact no other previous studies could simulate because of their expensive computational cost of DFT QM/MM. Here it is worth mentioning that, DFTB parameters used for this reaction were parameterised in chapter 8 specifically for this reaction as a SRP using DFT(B3LYP) as reference therefore we expect the PMF obtained in this work will have same accuracy as B3LYP QM/MM dynamics.

What we have learnt about the autophosphorylation, consisting of the phosphoryl-group transfer from ATP to His391 followed (at some point) by the deprotonation of pHis, leaves us with two possible scenarios of the whole process: One possibility is that pHis deprotonates immediately after the phosphoryl transfer has taken place. The first proton acceptor, Glu392 acts like a proton relay before the proton eventually is transferred to a sufficiently strong base (here represented by an OH^- anion), which need not be located directly next to pHis. *In vivo*, that strong base might be, for instance, a suitable titratable molecule present in the solution, or an area of local basic environment in the cytosol. The other conceivable scenario is that the base does not act in this step, and pHis does not deprotonate. Then, the next phosphorylation reaction, which is the phosphoryl transfer from pHis to the conserved Asp residue on the RR WalR, would have to proceed from the protonated pHis. Future work will need to answer whether this process would run spontaneously and if so, whether the energy barrier would be low enough to allow for reasonably favourable kinetics of the process.

Looking at the big picture of the phosphorylation cascade in TCS, the step subsequent to histidine autophosphorylation will be transfer of the phosphoryl group from His391 to the conserved aspartate residue on the WalR RR. The final product of the reaction sequence in this study, the WalK protein in an active state containing a deprotonated phosphorylated His391, represents a quite stable minimum on the free energy surface, yet still phosphohistidine is a relatively unstable, high-energetic species. This means that the energy of ATP hydrolysis has not been fully released, and is still available to drive consecutive processes, of which the first is a phosphoryl group transfer to the conserved Asp in RR.

It appears likely that the dynamics and energetics of WalK autophosphorylation found here are valid not only for a single HK, rather, broadly representative of the HK autophosphorylation mechanism. We base this on the fact that direct coupling analysis of vast sequence alignments of HK proteins identified highly correlated residue pairings between DHp and CA domains, that later were identified to be in close proximity in individual structural examples of either inactive or active conformation of the HK protein [37]. Thus, the vast majority in sequence alignments included in this study have to have similar active and inactive conformations, in which these contacts can be realized. By extension, we also anticipate similar dynamics in the transition between active and inactive conformations.

9.4. Conclusion

The initial processes in the signalling cascade of the WalK HK have been explored by means of a multi-scale simulation approach. The structure of the activated state of WalK HK served as a basis for an investigation of the autophosphorylation reaction, in which the γ-phosphate group of an ATP molecule bound to the CA domain is transferred to the His391

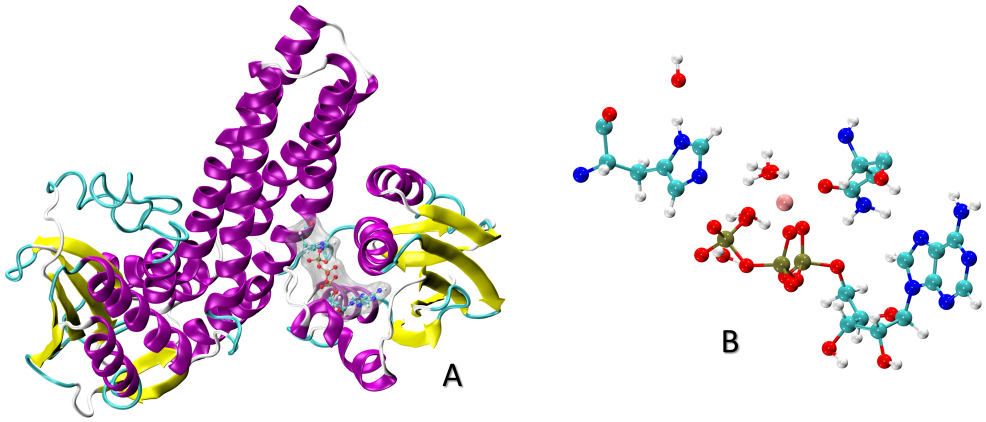

Figure 9.7.: A) Trans HK structure with the region of QM region, based CpxA Histidine Kinase obtained and modelled from pdb id:5lfk. B) Reaction center (QM region) is shown separately

.

residue of the DHp domain. The applied QM/MM MD multiple-walker metadynamics made it possible to achieve microsecond sampling and draw a more reliable picture of the mechanism and energetics of the process. The reaction was shown to proceed via a penta-coordinated transition state to a protonated phosphohistidine intermediate, which is consequently deprotonated in favour of a suitable nearby base. The role of the basicity of the final proton acceptor was also described quantitatively.

Accordingly, the obtained potential of the mean force of the conformational transition indicated an energy barrier of 27 kcal/mol [82]; this estimate however represents an upper bound of the real value due to the properties of the computational method. The phosphorylation step, on the other hand, exhibits down-the-hill energetics, with the exact shape being dependent on the nature of the final proton acceptor. Taken together with the high energy expense of the prior conformational transition, that draws a picture with isoenergetic or slightly exergonic process accompanied by a high energy barrier, being in agreement with and extending the current state of knowledge of the reaction.

9.5. Trans HK and future work

CpxA histidine kinase is an envelope stress sensor protein found in Escherichia coli. Protein misfolding in the periplasm is regulated by this two-component system for the kinase action of CpxA HK and a response regulator protein, CpxR. Here we have modelled a suitable structure based on pdb id 5lfk and identified the reaction center. Future work is to apply the same approach we have developed for WalK Cis histidine kinase.

10. Improving P-N pair Potential in DFTB3 using Neural Network

10.1. Inroduction

In this chapter we present another method to improve DFTB3 energies. In this approach we use Neural Network to improve DFTB energies (DFTB-NN), like the previous work [66], done on di-sulfide exchange reaction. Goal is simple, we calculate the DFTB energy and DFT(B3LYP) energy of the same system/molecule and then compute the difference between them and feed the difference in a neural network. Therefore the output of the neural network will give difference corrected energies which will be the same as DFT(B3LYP) level. Comparing the computational cost of DFT-hybrid functional (B3LYP) level QM/MM simulations, which is only accessible up to 100 picoseconds at max, DFTB-NN level QM/MM Simulations are faster, which can be accessible up to few hundreds of nanoseconds and gives the same results as DFT QM/MM. Thus this approach is very reliable for simulations of large biochemical system. The first such ML potential was introduced by Doren et al. in 1995, who fitted a DFT PES with the help of an Artificial Neural Network (ANN). However, the model was limited to a few atoms, and it would take 12 more years before larger atomic systems could be described. Here we use Behler–Parrinello method for constructing neural network.

Semi empirical(SE) methods can also be combined with ML algorithms, which are designed correct the energy difference between the SE method and a high level QM method one of such algorithm is called Δ-ML approach [141]. First ever implementation Δ-ML in a QM/MM framework was done by Shen and Yang in 2018 with an ANN to correct the PES of peptide building blocks [150]. Since then, similar approaches have been further developed. Some notable examples are by Böselt et al. [22],Gastegger et al.[56],etc

Being based on the PBE functional, 3OB parameters inherits DFT-PBE differences and sometimes complex reactions, which often needs to be described by better exchange correlation interactions lacks behind to give accurate estimation of reaction energy and reaction barrier. The reason why there are several SRP (special reaction parameters) were developed for some non-trivial reactions. Two of those SRP has been discussed in detail in chapter 8 and chapter 11. Uses of these SRPs also make the usage of DFTB3 more complex. This is where Δ-ML could help DFTB3 to give a practical solution. In this work, we aim to develop a Δ-ML based approach for the description of TEPA hydrolysis reaction within a QM/MM framework, where unlike in chapter 8 (using SRP) DFTB3 energies can be trained only using 3OB parameters without using any SRP and can be corrected to B3LYP level of energy. Later we also aim to extend this framework to implement a general phosphate reaction DFTB-NN model where no longer SRPs are required.

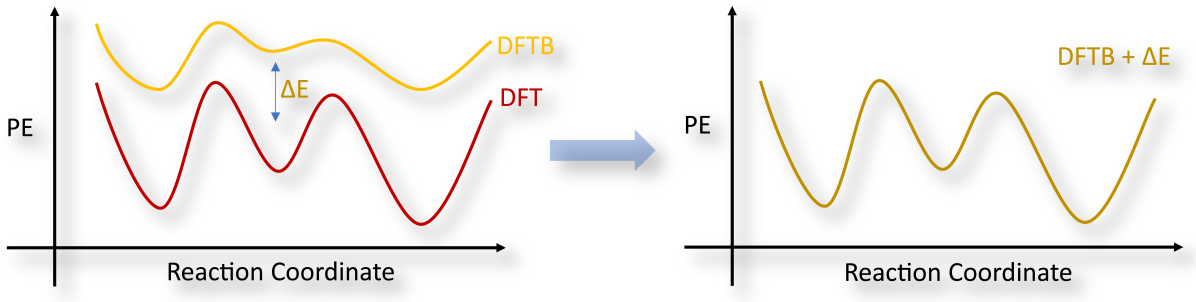

Figure 10.1.: Schematic diagram of Δ Machine Learning

10.2. Methodology

Δ-Machine

Concept of Δ-Machine is trivial, It is a ANN model that learns only the difference of a high level quantum energy (usually B3LYP or CC) and a low level quantum energy (in this case DFTB3/3OB) of a the same system (set of the same structures) and after training it takes low level energy as a input and predicts the higher level energy corresponding to the same structure, see figure 10.1. Therefore using this DFTB+Δ we can precisely reconstruct higher QM level potential energy surface.

Data generation

For the training of the ANN only gas-phase energies are required. In order to get the structures and the corresponding energies we took gas-phase metadynamics of the TEPA hydrolysis reaction we already computed for chapter 8 (see fig 8.10) and let the metadynamics run till it fill all the local minima and reach at the diffusive region, over-converged. Now structures are extracted from the trajectory after every 10ps. By this process we ended up making a total numbers of 16000 structures.

After obtaining structures, single point calculations are performed on every structure both in DFTB/3OB parameter using DFTB+ and in B3LYP/aug-cc-pVTZ. On the other hand we take another molecule of TEPA and scanned through the potential energy surface along P-N distance and P-O distance from 1.9 Å to 4 Å and used every structure to generate a 2ps MD trajectory which again generated 10000 structures. We then performed single point calculations on these structures both in DFTB/3OB parameter using DFTB+ and in B3LYP/aug-cc-pVTZ. Now in total we have 26000 structures for the training process. For comparison how good P-N SRP is for this particular reaction we also make a separate list of DFTB single point energies using both P–N and P–O SRPs.

Neural Network

In this work we have used a Behler-Parrinello ANN [15] for Δ-Machine. In Behler-Parrinello formalism, the quantity of a molecular system is expressed as the sum of

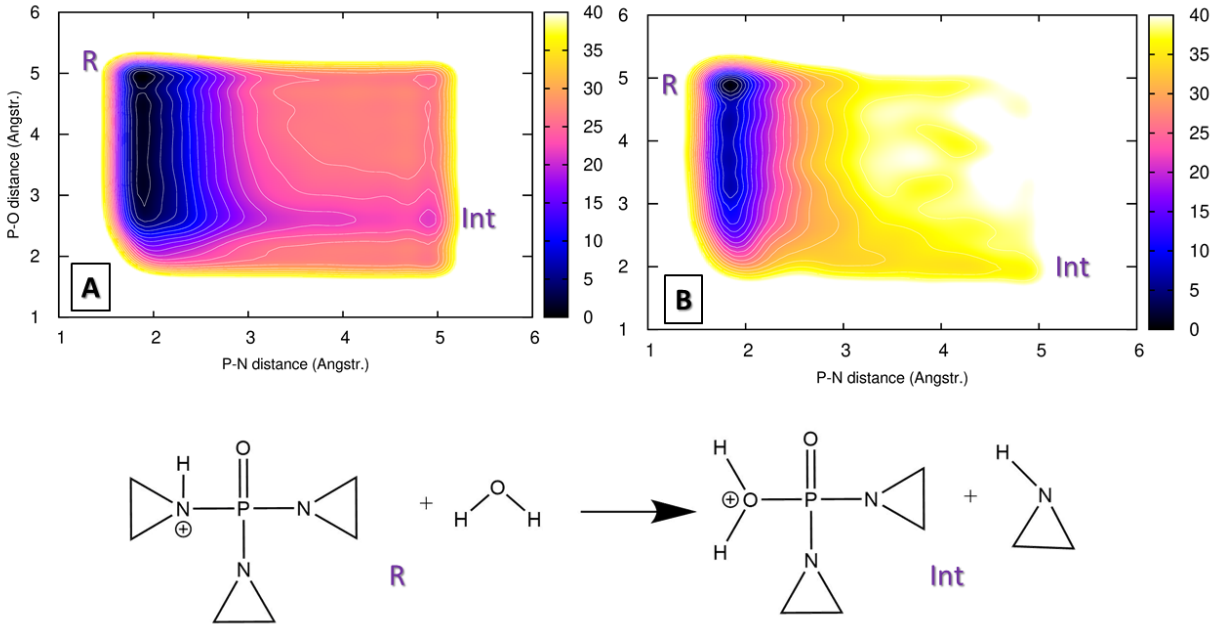

Figure 10.2.: PMF of gas-phase TEPA- hydrolysis reaction both in A)DFTB/3OB+PN+PO B) B3LYP/aug-cc-pVTZ

atomic contributions, ΔE_i. Each of the quantities ΔE_i has to be obtained from atomic neural network (also called subnet). In order to build the subnet molecular structure has to be converted from Cartesian coordinate to atomic symmetry functions (ACSF) (already discussed in detail in chapter 7). The implemented feed-forward neural network consists of a three-layer subnets for each atom with the tanh activation function. The descriptors are defined in terms of radial and angular symmetry functions as the input parameter. Each hidden layer consists of 34 neurons whose weights have been initialized by the NguyenWidrow initialization procedure.

10.3. Results

In figure 8.5 PMF shown in both DFTB and B3LYP, though B3LYP–PMF is not converged it can give a clear idea about the positions of the minima. In this plot DFTB is equipped with both P-N repulsive SRP and P-O repulsive SRP along with other 3OB parameters. In spite of that it can be observed from P–O bond, which is the "Int" minimum is quite displaced (2.5 Å) from the reference PMF (B3LYP) (2.0 Å),(P–O bonds are longer than it should be) indicates P-O repulsive underestimated P-O^+ bond. This makes room for the role of the Δ–machine to improve the P–O^+ interactions.

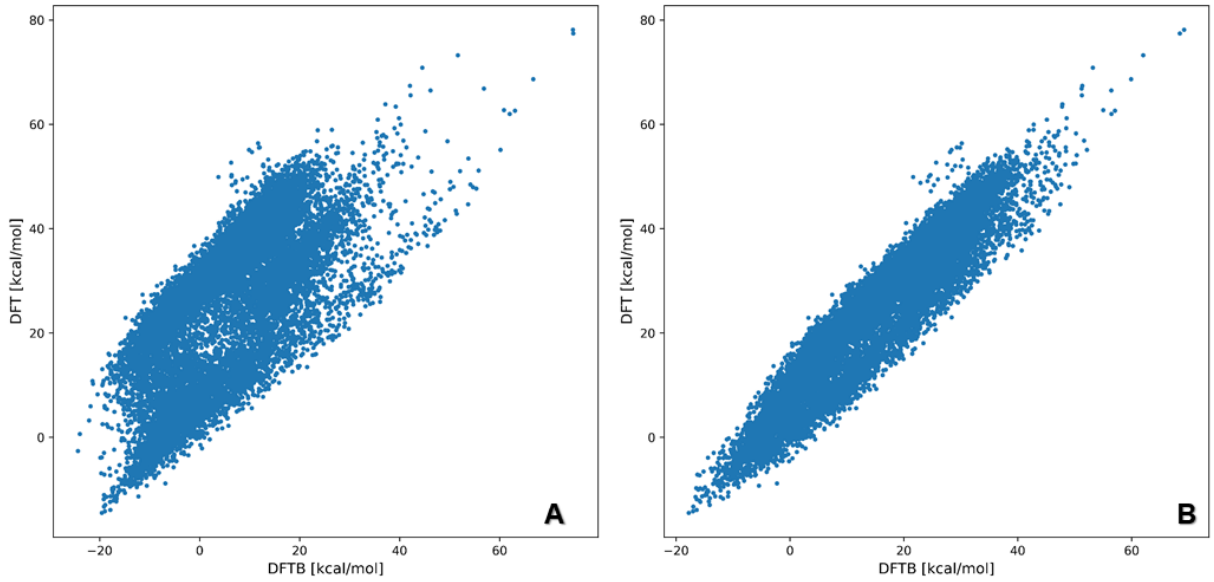

Figure 10.3.: Distribution of data points in data-set 1(A) and 2(B)

Training

In total we used four data sets 1)structures from metadynamics only (16000 data points) using DFTB/3OB energies as lower level, 2)structures from metadynamics only (16000 data points) using DFTB/3OB and SRP energies as lower level 3) structures from scanned structures(10000 data points) using DFTB/3OB energies as lower level 3) structures from both metadynamics and scanned structures (26000 points) using DFTB/3OB energies as lower level. In each data-sets 10% of the total data points were taken for test-set for test prediction. In all data sets B3LYP energies were kept as higher level. Symmetry functions (ACSF) are chosen with radial cut-off 5 Å and angular cut-off 4 Å for all data sets. Learning rate was set at 5×10^3.

At first we look at data-sets 1 (3OB) and 2 (3OB + SRP) 10.3, data-set 1 looks more scattered, which means energy range of the difference between DFTB/3ob–B3LYP energies are bigger, MAE (Mean Absolute Error) for data-set 1 is 20.72 kcal/mol on the other hand data-set 2 is less scattered indicating that re-parametrisation helped to reduce the difference between DFTB and B3LYP, where MAE is 9.43 kcal/mol.

After training (looking at 10.4) this difference reduced down drastically, for data-set 1 MAE appears to be 0.53 kcal/mol and for data-set 2 0.54 kcal/mol corresponding RMSEs are 0.69 kcal/mol and 0.70 kcal/mol. We can say the training worked really well for data-set 1 and 2.

Now in order to get more possibly diverse structure close to potential energy surface we include the scanned structures as well, which is data-set 3 and data-set 4. looking at 10.6 A and B it is evident that there are some potentially high energy structures which were

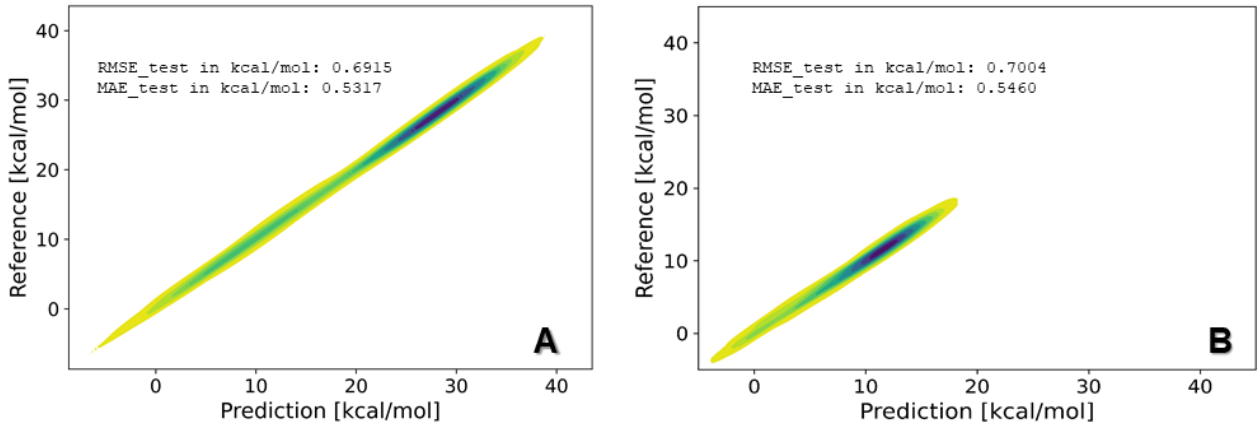

Figure 10.4.: test-set prediction for data-set 1 (A) and 2 (B)

left out in the previous data-sets, which could be essential for training the whole chemical space. For data-set 3 MAE turns out to be 34.10 kcal/mol and for data-set 4 32.06 kcal/mol. In both of these data-sets only 3OB parameters were used with DFTB3 to generate low level energy.

After Training with the same hyper parameters and 800 epochs, MAE comes down to 0.87 kcal/mol and 0.8879 kcal/mol respectively for data-set 3 and 4. In other word trained model now can take DFTB energy and predict the corresponding B3LYP energy of the same structure with a negligible difference of less than 1kcal/mol. This tiny difference will be irrelevant for QM/MM simulation of big biochemical systems.

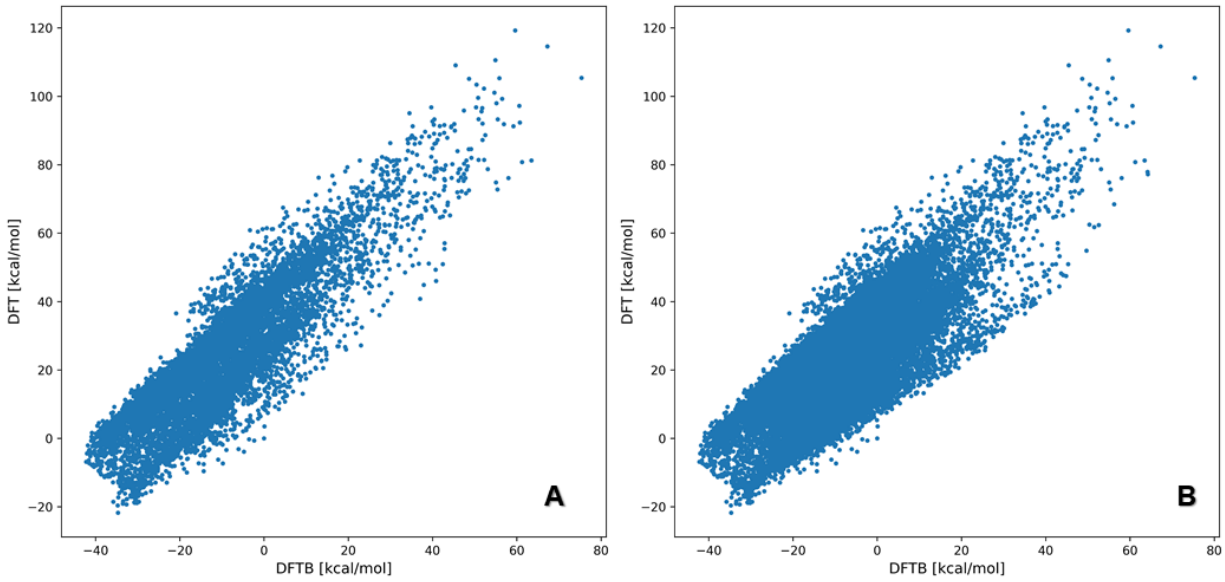

Figure 10.5.: Distribution of data points in data-set 3(A) and 4(B)

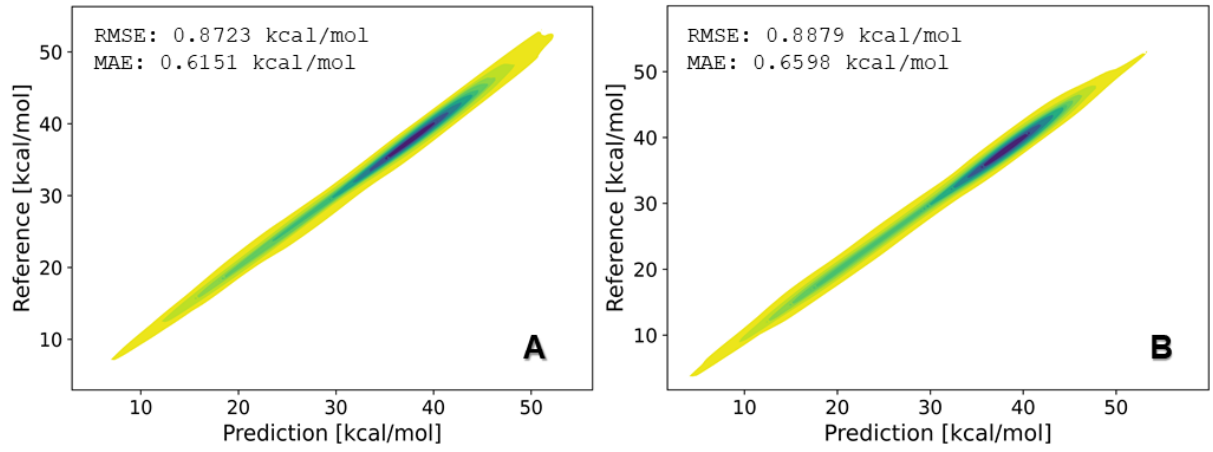

Figure 10.6.: test-set predictions for data-set 3 (A) and 4 (B)

10.4. Conclusion

In this chapter a neural network model has been demonstrated following the steps of our previous work [66]. We have shown that the method is very functional and efficient for gas-phase reactions. Now the goal is to study the hydrolysis reaction using QM/MM simulation equipped with the trained model. We have drawn a contrast between old reparameterisation method (SRP) and NN prediction. Although the use SRPs qualitatively improve DFTB energies compared to 3OB, there are still some differences exist w.r.t the higher level method (B3LYP) but with NN models this differences are drastically reduced down to less than 1 kcal/mol, which makes it more reliable for accuracy in the QM/MM simulations. Thus we can say, for future DFTB problems NN models will be an acceptable solution.

11. Re-parameterisation of Sulfur-Sulfur repulsive Potential for disulfide -thiol exchange reaction in DFTB3

Chapter 11 is reproduced in parts from Ref [66]:

Author Contributions:

This work was done in cooperation with Claudia Leticia Goméz-Flores and Denis Maag. Claudia Leticia Goméz-Flores optimized and trained the neural network. Denis Maag performed QM/MM simulations. Mayukh Kansari reparametrised the sulfur-sulfur parameters. Tomáš Kubař implemented the neural network in DFTB+.

11.1. Introduction

With DFTB3, the 3OB set of parameters are most commonly used for organic and biological systems. However, a few transferability issues were found for some complex chemical reactions. This led to incorrect reaction energetics, we already discussed such a case in Phosphorus-Nitrogen in chapter 8. Here we discuss another such problem for thiol–disulfide exchange reaction. In this chapter we propose another SRP to improve the performance of DFTB3/3OB.

The thiol–disulfide exchange reaction is one of the few special cases where the general 3OB parameter set exhibits considerable errors. Disulfide bonds are essential for the structure and functionality of many proteins, they are formed between two intra- or intermolecular cysteines and thus act as cross-links connecting secondary/tertiary protein structures. Moreover, they direct protein folding, stabilize proteins, catalyse and regulate enzymatic reactions, protect against oxidative damage and participate in electron transfer processes across membranes and in the secretory pathway of proteins. In the recent years, it has become more and more evident that disulfide bonds in proteins are not only static and stable but can also be dynamic and labile, able to rearrange by intra- or intermolecular thiol–disulfide exchange reactions.[115, 33] A thiol–disulfide exchange is an S_N2 reaction between a thiolate R1–S^- and a disulfide bond R2–S–S–R3 which results in the formation of a new disulfide bond, either R1–S–S–R2 or R1–S–S–R3.[69] The attacking sulfur will be referred to as S_{nuc}, the attacked sulfur as S_{ctr} and the leaving sulfur as S_{lg} and the transition state appears to be tri-sulfide like complex. This type of reaction is of great importance for many chemical and biological applications, motivating a variety of experimental and theoretical studies aiming to uncover the mechanistic details.[63, 184, 48] To improve the performance, we reparamterised the sulfur–sulfur pair repulsive parameter.

As for other S_N2 reactions, a hydrophobic environment is catalytic because the charge of the sulfurs is more delocalised.[152] In the gas phase, the charge is completely delocalised along the three sulfurs when the molecules are symmetric (R1=R2=R3) and form a nearly linear trisulfide complex.[9, 48] In a polar environment, e.g. in water and/or a protein, the charge is more localized. Consequently, the thiolate and the disulfide states are stabilized whereas the trisulfide state is the transition state.

11.2. Methodology

Here we again follow the same procedure, explained in chapter 8.

11.2.1. Re-parameterisation scheme

We begin with the total energy of DFTB3:

$$E = E^{(1)} + E^{(2)} + E^{(3)} + E^{\text{rep}} \tag{11.1}$$

$$= \sum_{i}^{\text{MO}} n_i \sum_{a,b}^{\text{atoms}} \sum_{\mu \in a}^{\text{AO}} \sum_{\nu \in b}^{\text{AO}} c_{\mu i} c_{\nu i} H^0_{\mu\nu} + \frac{1}{2} \sum_{a,b}^{\text{atoms}} \Delta q_a \Delta q_b \gamma_{ab} + \frac{1}{3} \sum_{a,b}^{\text{atoms}} \Delta q_a^2 \Delta q_b \Gamma_{ab} + \frac{1}{2} \sum_{A,B}^{\text{atoms}} V_{AB}^{rep} \tag{11.2}$$

The repulsive part representing the repulsive potential expressed in terms of pair potentials V_{ab}^{rep} which are specific to respective pairs of chemical elements and depend on interatomic distance but not on atomic charges. The V_{ab}^{rep} are determined by fitting the repulsive potentials as spline functions to a selected set of reference atomization energies, molecular geometries. The parameterisation procedure is carried out according to a standard, partially automatized protocol [61].

New repulsive potential for S–S interaction is created in this work, by means of a fit for Sulfur-Sulfur bond containing molecules (including all electronic parameters). In this case disulfide and trisulfide, since it is very specific for di-sulfide exchange reaction. Molecules considered were, a dimethyl-disulfide molecule (10 atoms) and a trimethyl-trisulfide anion (15 atoms). For both molecules geometries were optimized at B3LYP/aug-cc-pVTZ level of theory, molecules shown in 11.1.The atomization energies for both of them were obtained from G3B3 [8] single point calculations.

Now following the standard protocol described in Ref. [60] (also discussed in chapter 8) linear equation set were created and solved for determining the repulsive pair potential spline using suitable division points (grid points). The exponential form of the spline is maintained by so-called additional equation V', introduced at the beginning. An overview of all reference systems and values that lead to the repulsive potential related to S–S is provided in Tab. 11.1. However since all 3OB parameters has some overbinding, new s-s parameter could not reproduce the topography of the PES 11.3(A), so it was evident that S–S also needs some overbinding. Thus, the atomisation energy of the obtained potential was shifted by -28 kcal/mol, i.e., we overbinded the S–S bond. The final S–S repulsive pair potential is able to reproduce the B3LYP/def2-TZVPP PES with an error of ca. 1 kcal/mol,

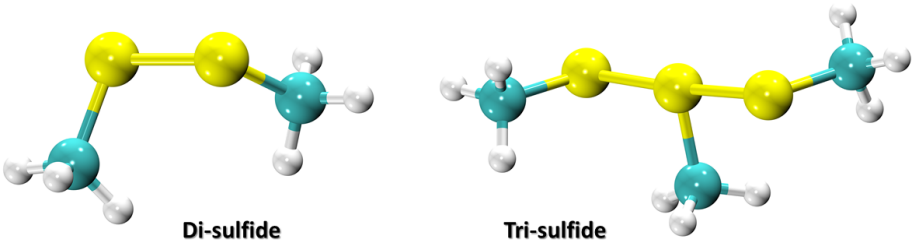

Figure 11.1.: Molecules taken for parameterisation

11.3(D). A comparison of the old 3OB S–S repulsive potential and the new S–S repulsive potential can be found in 11.2

Table 11.1.: Parameters defining the repulsive potential. Atomization energies of a dimethyl disulfide and a trimethyl trisulfide anion, used to reparametrise the S–S repulsive potential using 4 spline division points and an additional equation

Molecule	Charge (e)	E^{at} (kcal/mol)
$H_3CS–SCH_3$	0	856.4
$H_3CS–S(CH_3)–SCH_3$	−1	1292.6

Potential	Division points (au)	Additional equations (au)
S–S	3.7, 4.1, 5.5, 6.5	$V''(1.958\,\text{Å}) = 0.21$ a.u.

11.3. Results

For compare the performance of the SRP in this work, we chose the B3LYP/def2-TZVPP potential energy surface (PES) scan calculations preformed along two S-S distances in trisulfide anion molecule 11.3B from Putzu et al. [135] as reference. Scanned PES is shown in fig. 11.3A. The energy of the two molecules at infinite distances, i.e., the sum of energies of isolated methylthiolate and isolated dimethyl disulfide, was set to zero. We took these structures and repeat single point calculations in DFTB+ using old 3OB parameter 11.3 C and with new S–S repulsive potential 11.3 D.

It is important to notice here that in gasphase, the potential energy profile is inverted because the charge is delocalised between the sulfur atoms. Thus, the linear “trisulfide” complex is no longer a transition state but a minimum.

Comparing Three surfaces it is noticeable that new SRP almost reproduce the reference B3LYP scan. With 3OB parameters T1 minimum used to be more deeper and appeared in a wrong position, indicates longer S–S bonds. This phenomenon correlates with the behavior of the DFT-LDA and DFT-GGA approaches [123]. Being based on the PBE functional, DFTB thus seems to reproduce the DFT-PBE errors. DFT-PBE not only fails to give an accurate description of structures, but also exhibits an error in energies of ca. 7 kcal/mol.

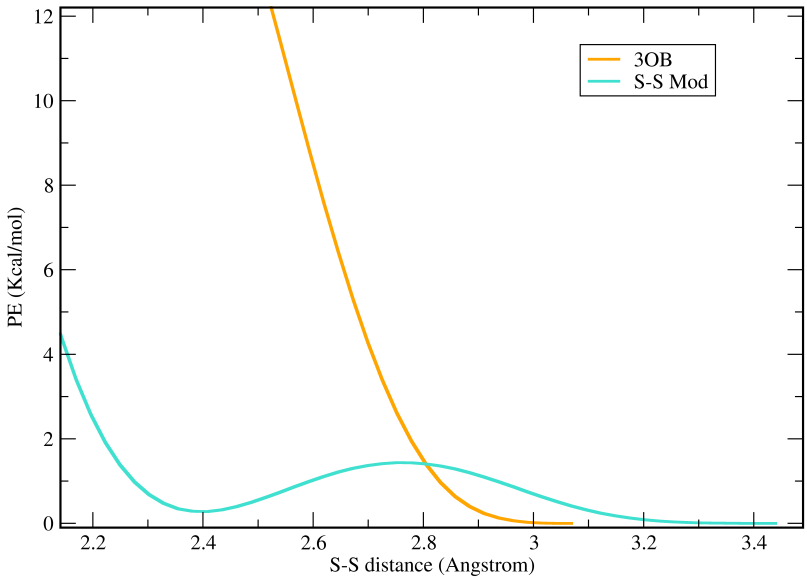

Figure 11.2.: Comparison of old 3OB and New S-S repulsive potential

Therefore, it seems that DFTB/3OB very much inherits the DFT-PBE problems. B3LYP performs much better but still slightly overestimates the bond lengths in the minimum, with an error of 3 kcal/mol. The largest qualitative difference are apparent for high-energy structures, which are hardly relevant in typical applications.

As discussed above, the PES of thiol–disulfide exchange for a solvated system differs significantly from a gas-phase system, and a transition state appears where there is a minimum in the gas phase. To investigate the performance of the the new repulsive spline for solvated systems, we performed QM/MM metadynamics simulations of a dimethyl disulfide–methylthiolate system immersed in water that was described by an MM force field. The metadynamics setup was designed to sample all three disulfide bond patterns, i.e. S1–S2, S1–S3, and S2–S3 with the respective third sulfur in a deprotonated anionic state. The free energy profile of the exchange reactions is completely symmetric and therefore ideally suited for comparing the different levels of theory. The 2D representations of the three-dimensional free energy landscape, expressed as a function of the S1–S2 and S1–S3 distances with the S2–S3 distance integrated out, are shown in Fig. 11.4 together with exemplary molecular structures and pathways. All PMFs are symmetrical and show the three expected minima of equal depth. Moreover, the transition states within the respective PMFs have the same energy, which illustrates the good convergence of the simulations.

The PMF obtained with uncorrected DFTB/3OB (Fig. 11.4A) shows two significant problems: (i) the bonds S1–S2 and S1–S3 in the transition state geometries are too long with ca. 2.8 Å, and (ii) the transition state geometries exhibit shallow minima on the free energy landscape, rather than saddle points[135].

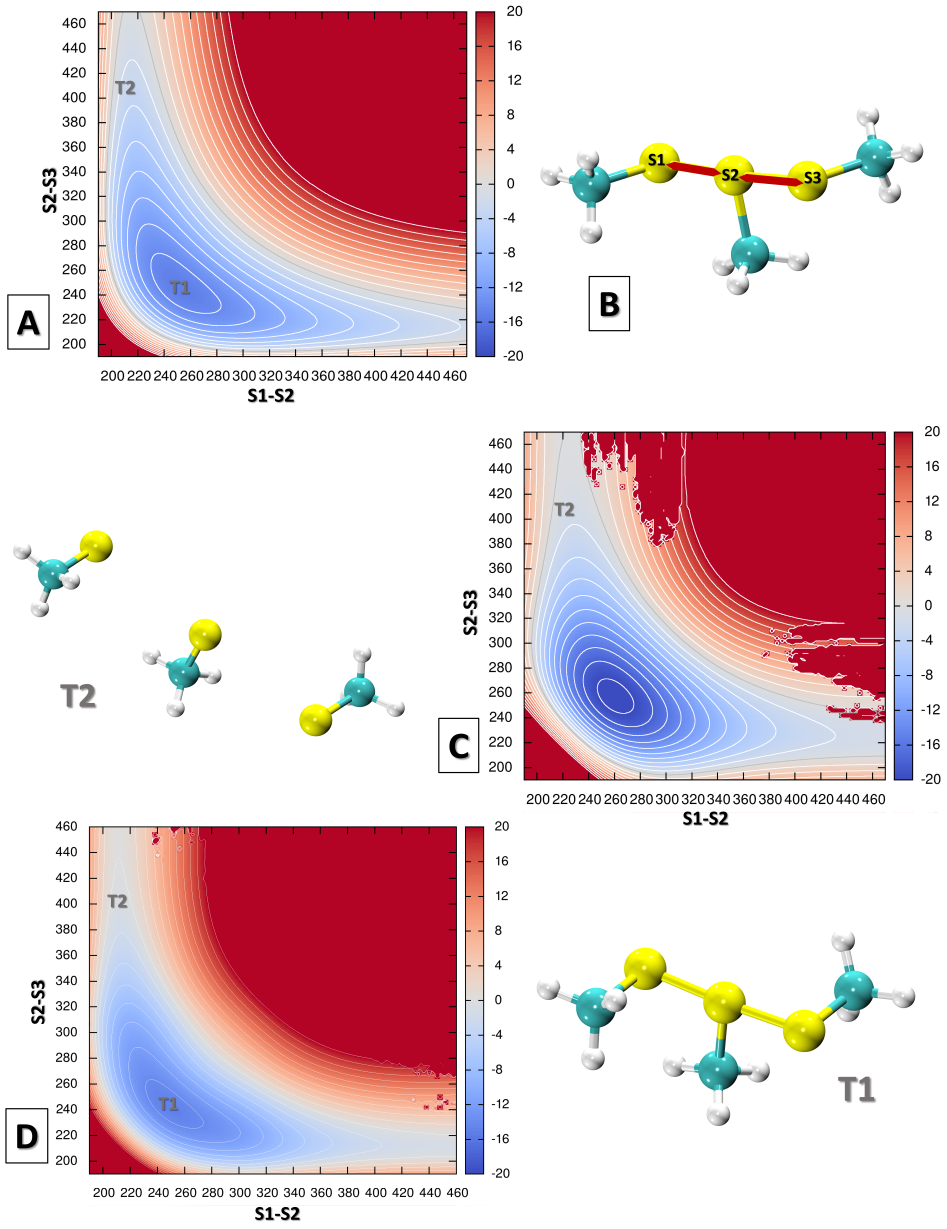

Figure 11.3.: Gas-phase potential energy surfaces, representing the total energy as a function of B) S1–S2 and S1–S3 bond length in a linear configuration exhibited using different level of theories A) BLYP/aug-cc-pVTZ, C) DFTB/3OB, D) DFTB/3OB+ new S–S SRP

Both problems are resolved by reparametrising the S–S repulsive potentials 11.4B The transition states now appear as saddle points at shorter bond lengths This correlates with the B3LYP potential energy scan in vacuum.

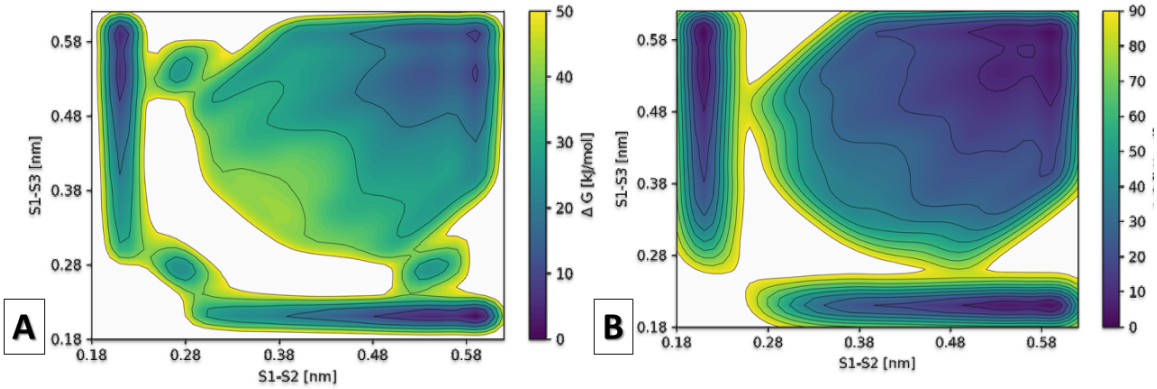

Figure 11.4.: PMF obtained from QM/MM simulation in MM water A) Using 3OB sets of parameters B) Using newly created S-S modified parameter

11.4. Conclusion

Disulfide bonds have an important role for the function of many proteins, therefore, being able to address these reactions using accurate computational approaches is of great importance. These reactions were shown to be quite challenging for DFT methods requiring costly computational approaches to be applied. Sampling, however, is then out of reach, which poses a further restriction on the accuracy of the results. DFTB is 3–4 orders of faster than DFT-GGA using moderately sized basis sets, however, they may run into even greater difficulties for challenging reactions. Reparameterisation of sulfur-sulfur repulsive potential has solved the problem and reduce the error qualitatively.

12. Summary

We have developed new SRP for better describing Phosphorus-Nitrogen interaction in Histidine phosphorylation in chapter 8. We performed QM/MM simulation on model imidazole-phosphate reaction using new SRP and showed that our SRP can reproduce B3LYP level reaction energies and barrier only with a tiny difference (ca 2kcal/mol). Which is sufficient for reactions for large biological systems like histidine kinase. Though this SRP is made specifically for the purpose of Histidine phosphorylation, we further tested it on other similar gas-phase reactions to evaluate accuracy and performance. After that we extended our study further to investigate cancer drug (TEPA) hydrolysis using the new parameter and demonstrated the full mechanism of the hydrolysis for the first time using QM/MM multiple-walker well-tempered metadynamics.

We studied the whole mechanism of Histidine autophosphorylation by using new parameter developed in chapter 8. We applied QM/MM MD multiple-walker well-tempered metadynamics made it possible to achieve microsecond scale sampling and to draw a reliable picture of the mechanism and energetics of the process. The reaction was found to proceed via a penta-coordinated transition state to a protonated phosphohistidine intermediate, which then gets deprotonated in favour of a suitable nearby base. The role of the basicity of the final proton acceptor was also described quantitatively. Further we investigated the role of Magnesium ion in the simulation. We compared our study with available experiment and we found substantial agreement of our observation with the outcomes of the experimental observation. Though it is difficult to relate experimental (in vitro) biochemical reactions with the same reaction in biological (in vivo) condition, this makes computer simulation, only way to investigate such mechanism in great detail. We also developed forcefield parameters for phosphorylated histidine residue (non-standard residue) in this work. These parameters will help studying next step of the phospho-relay cascade of WalK/R Two component system.

In chapter 11 we developed SRP for disulfide-thiol exchange reaction by reparameterising sulfur-sulfur repulsive potential. Earlier with 3OB sulfur-sulfur bonds were unusually long in trisulfide ion which led wrong minima instead of transition state in free energy simulation. This made difficult to assess reaction barriers. With new SRP the problem is resolved, transition state appeared in its position. The new SRP can reproduce the same energy profile as B3LYP (with ca 1kcal/mol error).

In the chapter 10 we tried to use Artificial Neural Network to improve DFTB energies without using any SRP. We used gas-phase structures of TEPA hydrolysis reaction for training and applied Δ−machine algorithm equipped with Behler-Parrinello Neural Network to construct the model. After training the error (MAE) drastically reduce under 1 kcal/mol. To give a comparison, this error (MAE) was 9 kcal/mol using new SRP. Thus, we can conclude that, though SRP could reduce the qualitative error of DFTB3 Neural Network can improve this error much more efficiently.

Bibliography

[1] Mark James Abraham et al. “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 1–2 (2015), pp. 19–25. DOI: 10.1016/j.softx.2015.06.001.

[2] Daniela Albanesi et al. “Structural plasticity and catalysis regulation of a thermosensor histidine kinase”. In: *Proceedings of the National Academy of Sciences* 106.38 (2009), pp. 16185–16190.

[3] Michael P Allen et al. “Introduction to molecular dynamics simulation”. In: *Computational soft matter: from synthetic polymers to proteins* 23.1 (2004), pp. 1–28.

[4] Johan Åqvist. “Ion-water interaction potentials derived from free energy perturbation simulations”. In: *Journal of Physical Chemistry* 94.21 (1990), pp. 8021–8024.

[5] L Aravind and Chris P Ponting. “The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins”. In: *FEMS microbiology letters* 176.1 (1999), pp. 111–116.

[6] Fatima Ardito et al. “The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy”. In: *International journal of molecular medicine* 40.2 (2017), pp. 271–280.

[7] Orr Ashenberg, Amy E Keating, and Michael T Laub. “Helix bundle loops determine whether histidine kinases autophosphorylate in cis or in trans”. In: *J. Mol. Biol.* 425.7 (2013), pp. 1198–1209.

[8] Anwar G Baboul et al. “Gaussian-3 theory using density functional geometries and zero-point energies”. In: *The Journal of chemical physics* 110.16 (1999), pp. 7650–7657.

[9] Robert D. Bach, Olga Dmitrenko, and Colin Thorpe. “Mechanism of Thiolate-Disulfide Interchange Reactions in Biochem.” In: *J. Org. Chem,* 73.1 (Jan. 2008), pp. 12–21. ISSN: 0022-3263, 1520-6904. DOI: 10.1021/jo702051f. (Visited on 03/29/2021).

[10] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. “Well-tempered metadynamics: a smoothly converging and tunable free-energy method”. In: *Phys. Rev. Lett.* 100.2 (2008), p. 020603.

[11] Eric Batchelor and Mark Goulian. “Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system”. In: *Proceedings of the National Academy of Sciences* 100.2 (2003), pp. 691–696.

[12] Christopher I Bayly et al. “A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model”. In: *The Journal of Physical Chemistry* 97.40 (1993), pp. 10269–10280.

[13] Jörg Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials". In: *The Journal of chemical physics* 134.7 (2011), p. 074106.

[14] Jörg Behler. "Four generations of high-dimensional neural network potentials". In: *Chemical Reviews* 121.16 (2021), pp. 10037–10072.

[15] Jörg Behler and Michele Parrinello. "Generalized neural-network representation of high-dimensional potential-energy surfaces". In: *Physical review letters* 98.14 (2007), p. 146401.

[16] Agnieszka E Bem et al. "Bacterial histidine kinases as novel antibacterial drug targets". In: *ACS chemical biology* 10.1 (2015), pp. 213–224.

[17] Romualdo Benigni and Cecilia Bossa. "Mechanisms of chemical carcinogenicity and mutagenicity: a review with implications for predictive toxicology". In: *Chemical reviews* 111.4 (2011), pp. 2507–2536.

[18] Melissa T Berhow, Noboru Hiroi, and Eric J Nestler. "Regulation of ERK (extracellular signal regulated kinase), part of the neurotrophin signal transduction cascade, in the rat mesolimbic dopamine system by chronic exposure to morphine or cocaine". In: *Journal of Neuroscience* 16.15 (1996), pp. 4707–4715.

[19] Herbert Bestian. "Über einige Reaktionen des Äthylen-imins". In: *Justus Liebigs Annalen der Chemie* 566.2 (1950), pp. 210–244.

[20] Lisa Bleul, Patrice Francois, and Christiane Wolz. "Two-Component Systems of *S. aureus*: Signaling and Sensing Mechanisms". In: *Genes* 13.1 (2022), p. 34.

[21] Nils Blüthgen et al. "Effects of sequestration on signal transduction cascades". In: *The FEBS journal* 273.5 (2006), pp. 895–906.

[22] Lennard Böselt, Moritz Thürlemann, and Sereina Riniker. "Machine learning in QM/MM molecular dynamics simulations of condensed-phase systems". In: *Journal of Chemical Theory and Computation* 17.5 (2021), pp. 2641–2658.

[23] Alejandro Buschiazzo and Felipe Trajtenberg. "Two-component sensing and regulation: how do histidine kinases talk with response regulators at the molecular level?" In: *Annu. Rev. Microbiol.* 73 (2019), pp. 507–528.

[24] Giovanni Bussi, Davide Donadio, and Michele Parrinello. "Canonical sampling through velocity rescaling". In: *J. Chem. Phys.* 126.1 (2007), p. 014101.

[25] Giovanni Bussi and Alessandro Laio. "Using metadynamics to explore complex free-energy landscapes". In: *Nature Reviews Physics 2020 2:4* 2 (4 Mar. 2020), pp. 200–212. ISSN: 2522-5820. DOI: 10.1038/s42254-020-0153-0. URL: https://www.nature.com/articles/s42254-020-0153-0.

[26] Yongfei Cai et al. "Conformational dynamics of the essential sensor histidine kinase WalK". In: *Acta Crystallogr. D: Struct. Biol.* 73.10 (2017), pp. 793–803.

[27] Emily J Capra and Michael T Laub. "Evolution of two-component signal transduction systems". In: *Annual review of microbiology* 66 (2012), pp. 325–347.

[28] Patricia Casino, Laura Miguel-Romero, and Alberto Marina. "Visualizing autophosphorylation in histidine kinases". In: *Nat. Commun.* 5 (2014), p. 3258.

[29] Reha Celikel et al. “ATP forms a stable complex with the essential histidine kinase WalK (YycG) domain”. In: *Acta Crystallographica Section D: Biological Crystallography* 68.7 (2012), pp. 839–845.

[30] Susana K Checa, Matias D Zurbriggen, and Fernando C Soncini. “Bacterial signaling systems as platforms for rational design of new generations of biosensors”. In: *Current opinion in biotechnology* 23.5 (2012), pp. 766–772.

[31] Ryan R Cheng et al. “Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information”. In: *Proceedings of the National Academy of Sciences USA* 111.5 (2014), E563–E571.

[32] Jonah Cheung and Wayne A Hendrickson. “Sensor domains of two-component regulatory systems”. In: *Current opinion in microbiology* 13.2 (2010), pp. 116–123.

[33] Joyce Chiu and Philip J. Hogg. “Allosteric Disulfides: Sophisticated Molecular Structures Enabling Flexible Protein Regulation”. In: *J. Biol. Chem.* 294.8 (Feb. 2019), pp. 2949–5908. ISSN: 0021-9258. DOI: 10.1074/jbc.REV118.005604.

[34] Valerie A Clausen et al. “Biochemical characterization of the first essential two-component signal transduction system from *Staphylococcus aureus* and *Streptococcus pneumoniae*”. In: *J. Mol. Microbiol. Biotechnol.* 5.4 (2003), pp. 252–260.

[35] Michel A. Cuendet and Wilfred F. van Gunsteren. “On the calculation of velocity-dependent properties in molecular dynamics simulations using the leapfrog integration algorithm”. In: *Journal of Chemical Physics* 127 (18 Nov. 2007), p. 184102. ISSN: 0021-9606. DOI: 10.1063/1.2779878. URL: https://aip.scitation.org/doi/abs/10.1063/1.2779878.

[36] Qiang Cui. “Perspective: Quantum mechanical methods in biochemistry and biophysics”. In: *J. Chem. Phys.* 145.14 (2016), p. 140901.

[37] Angel E Dago et al. “Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis”. In: *Proc. Natl. Acad. Sci. USA* 109.26 (2012), E1733–E1742.

[38] Tom Darden, Darrin York, and Lee Pedersen. “Particle–mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems”. In: *J. Chem. Phys.* 98.12 (1993), pp. 10089–10092.

[39] Ernest R Davidson and David Feller. “Basis set selection for molecular calculations”. In: *Chemical Reviews* 86.4 (1986), pp. 681–696.

[40] Terri Davis-Smyth et al. “The second immunoglobulin-like domain of the VEGF tyrosine kinase receptor Flt-1 determines ligand binding and may initiate a signal transduction cascade.” In: *The EMBO journal* 15.18 (1996), pp. 4919–4927.

[41] Igor Dikiy et al. “Insights into histidine kinase activation mechanisms from the monomeric blue light sensor EL346”. In: *Proc. Natl. Acad. Sci. USA* 116.11 (2019), pp. 4963–4972.

[42] Thom Dunning, P Jeffrey Hay, et al. “Gaussian basis sets for molecular calculations”. In: *Methods of electronic structure theory*. Springer, 1977, pp. 1–27.

[43] Rinku Dutta, Ling Qin, and Masayori Inouye. "Histidine kinases: diversity of domain organization". In: *Molecular microbiology* 34.4 (1999), pp. 633–640.

[44] Yoko Eguchi et al. "Angucycline antibiotic waldiomycin recognizes common structural motif conserved in bacterial histidine kinases". In: *The Journal of antibiotics* 70.3 (2017), pp. 251–258.

[45] Marcus Elstner et al. "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties". In: *Physical Review B* 58.11 (1998), p. 7260.

[46] Ulrich Essmann et al. "A smooth particle–mesh Ewald method". In: *Journal of Chemical Physics* 103 (19 Aug. 1998), p. 8577. ISSN: 0021-9606. DOI: 10.1063/1.470117. URL: https://aip.scitation.org/doi/abs/10.1063/1.470117.

[47] Md Fakhruzzaman et al. "Study on in vivo effects of bacterial histidine kinase inhibitor, Waldiomycin, in Bacillus subtilis and Staphylococcus aureus". In: *The Journal of General and Applied Microbiology* 61.5 (2015), pp. 177–184.

[48] Pedro Alexandrino Fernandes and Maria João Ramos. "Theoretical Insights into the Mechanism for Thiol/Disulfide Exchange". In: *Chem. Eur. J* 10.1 (Jan. 2004), pp. 257–266. ISSN: 0947-6539, 1521-3765. DOI: 10.1002/chem.200305343. URL: http://doi.wiley.com/10.1002/chem.200305343 (visited on 03/29/2021).

[49] Hedda U Ferris et al. "The mechanisms of HAMP-mediated signaling in transmembrane receptors". In: *Structure* 19.3 (2011), pp. 378–385.

[50] Martin J Field. *A practical introduction to the simulation of molecular systems*. Cambridge University Press, 1999.

[51] Charlotte Froese Fischer. "Hartree–Fock method for atoms. A numerical approach". In: (1977).

[52] Stewart L Fisher et al. "Cross-talk between the Histidine Protein Kinase VanS and the Response Regulator PhoB: CHARACTERIZATION AND IDENTIFICATION OF A VanS DOMAIN THAT INHIBITS ACTIVATION OF PhoB ()". In: *Journal of Biological Chemistry* 270.39 (1995), pp. 23143–23149.

[53] Stewart L Fisher et al. "Kinetic comparison of the specificity of the vancomycin resistance kinase VanS for two response regulators, VanR and PhoB". In: *Biochemistry* 35.15 (1996), pp. 4732–4740.

[54] Marie E Fraser et al. "A detailed structural description of Escherichia coli succinyl-CoA synthetase". In: *Journal of molecular biology* 285.4 (1999), pp. 1633–1653.

[55] Daan Frenkel and Berend Smit. *Understanding molecular simulation: from algorithms to applications*. Vol. 1. Elsevier, 2001.

[56] Michael Gastegger, Kristof T Schütt, and Klaus-Robert Müller. "Machine learning of solvent effects on molecular spectra and reactions". In: *Chemical science* 12.34 (2021), pp. 11473–11483.

[57] Michael Gaus. "Extension and Parametrization of an Approximate Density Functional Method for Organic and Biomolecules". Karlsruhe, Karlsruher Institut für Technologie (KIT), Diss., 2011. PhD thesis. Karlsruhe, 2011. URL: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000024141%20;%20http://d-nb.info/1014817803/34%20;%20http://nbn-resolving.de/urn:nbn:de:swb:90-241415.

[58] Michael Gaus, Qiang Cui, and Marcus Elstner. "Density functional tight binding: application to organic and biological molecules". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.1 (2014), pp. 49–61.

[59] Michael Gaus, Qiang Cui, and Marcus Elstner. "DFTB3: extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB)". In: *J. Chem. Theory Comput.* 7.4 (2011), pp. 931–948.

[60] Michael Gaus, Albrecht Goez, and Marcus Elstner. "Parametrization and benchmark of DFTB3 for organic molecules". In: *J. Chem. Theory Comput.* 9.1 (2013), pp. 338–354.

[61] Michael Gaus et al. "Automatized Parametrization of SCC-DFTB Repulsive Potentials: Application to Hydrocarbons". In: *J. Phys. Chem. A* 113.43 (2009), pp. 11866–11881. DOI: 10.1021/jp902973m.

[62] Michael Gaus et al. "Parameterization of DFTB3/3OB for sulfur and phosphorus for chemical and biological applications". In: *J. Chem. Theory Comput.* 10.4 (2014), pp. 1518–1537.

[63] Hiram F. Gilbert. "Molecular and Cellular Aspects of Thiol-Disulfide Exchange". In: *Advances in Enzymology - and Related Areas of Molecular Biology*. Ed. by Alton Meister. Hoboken, NJ, USA: John Wiley & Sons, Inc., Nov. 2006, pp. 69–172. DOI: 10.1002/9780470123096.ch2. URL: http://doi.wiley.com/10.1002/9780470123096.ch2 (visited on 04/06/2021).

[64] Thomas L Gilbert. "Hohenberg-Kohn theorem for nonlocal external potentials". In: *Physical Review B* 12.6 (1975), p. 2111.

[65] Peter MW Gill et al. "The performance of the Becke—Lee—Yang—Parr (B—LYP) density functional theory with various basis sets". In: *Chemical Physics Letters* 197.4-5 (1992), pp. 499–505.

[66] Claudia L Gómez-Flores et al. "Accurate Free Energies for Complex Condensed-Phase Reactions Using an Artificial Neural Network Corrected DFTB/MM Methodology". In: *Journal of Chemical Theory and Computation* (2022).

[67] Anna R Greenswag et al. "Conformational transitions that enable histidine kinase autophosphorylation and receptor array integration". In: *Journal of molecular biology* 427.24 (2015), pp. 3890–3907.

[68] Eli S Groban et al. "Kinetic buffering of cross talk between bacterial two-component sensors". In: *Journal of molecular biology* 390.3 (2009), pp. 380–393.

[69] Trevor A. Hamlin, Marcel Swart, and F. Matthias Bickelhaupt. "Nucleophilic Substitution (SN2): Dependence on Nucleophile, Leaving Group, Central Atom, Substituents, and Solvent". In: *ChemPhysChem.* 19.11 (2018), pp. 1315–1330. ISSN: 1439-7641. DOI: https://doi.org/10.1002/cphc.201701363. (Visited on 02/11/2021).

[70] John A Hartley et al. "DNA sequence selectivity of guanine-N7 alkylation by three antitumor chloroethylating agents". In: *Cancer research* 46.4 Part 2 (1986), pp. 1943–1947.

[71] Berk Hess et al. "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation". In: *J. Chem. Theory Comput.* 4.3 (2008), pp. 435–447. DOI: 10.1021/ct700301q.

[72] Berk Hess et al. "LINCS: a linear constraint solver for molecular simulations". In: *J. Comput. Chem.* 18.12 (1997), pp. 1463–1472.

[73] Steven M Hill. "Receptor crosstalk: communication through cell signaling pathways". In: *The Anatomical Record: An Official Publication of the American Association of Anatomists* 253.2 (1998), pp. 42–48.

[74] P Hohenberg and WJPR Kohn. "Density functional theory (DFT)". In: *Phys. Rev* 136 (1964), B864.

[75] Thomas D Horn et al. "Observations and Proposed Mechanism of N, N', N?-Triethylenethiophosphoramide (Thiotepa)-Induced Hyperpigmentation". In: *Archives of dermatology* 125.4 (1989), pp. 524–527.

[76] Guanhua Hou and Qiang Cui. "QM/MM analysis suggests that alkaline phosphatase (AP) and nucleotide pyrophosphatase/phosphodiesterase slightly tighten the transition state for phosphate diester hydrolysis relative to solution: implication for catalytic promiscuity in the AP superfamily". In: *Journal of the American Chemical Society* 134.1 (2012), pp. 229–246.

[77] Ben Hourahine et al. "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations". In: *J. Chem. Phys.* 152.12 (2020), p. 124101.

[78] DE Hultquist, R Wo Moyer, and PD Boyer. "The preparation and characterization of 1-phosphohistidine and 3-phosphohistidine". In: *Biochemistry* 5.1 (1966), pp. 322–331.

[79] TuAnh Ngoc Huynh and Valley Stewart. "Negative control in two-component signal transduction by transmitter phosphatase activity". In: *Molecular microbiology* 82.2 (2011), pp. 275–286.

[80] S Huzinaga. "Analytical methods in Hartree-Fock self-consistent field theory". In: *Physical Review* 122.1 (1961), p. 131.

[81] S Huzinaga. "GTO basis sets for heavier elements". In: *The Journal of Chemical Physics* 66.9 (1977), pp. 4245–4245.

[82] Fathia Idiris. "Multiscale Molecular Dynamics Simulations of Histidine Kinase Activity". PhD thesis. 2022.

[83] Masayuki Igarashi et al. “Waldiomycin, a novel WalK-histidine kinase inhibitor from *Streptomyces sp.* MK844-mF10”. In: *J. Antibiot.* 66.8 (2013), pp. 459–464.

[84] Michele M Igo et al. “Phosphorylation and dephosphorylation of a bacterial transcriptional activator by a transmembrane receptor.” In: *Genes Dev.* 3.11 (1989), pp. 1725–1734.

[85] Françoise Jacob-Dubuisson et al. “Structural insights into the signalling mechanisms of two-component systems”. In: *Nat. Rev. Microbiol.* 16.10 (2018), pp. 585–593.

[86] Milly E de Jonge et al. “Simultaneous quantification of cyclophosphamide, 4-hydroxycyclophosphamide, N, N, N -triethylenethiophosphoramide (thiotepa) and N, N, N -triethylenephosphoramide (tepa) in human plasma by high-performance liquid chromatography coupled with electrospray ionization tandem mass spectrometry”. In: *Journal of mass spectrometry* 39.3 (2004), pp. 262–271.

[87] William L Jorgensen et al. “Comparison of simple potential functions for simulating liquid water”. In: *J. Chem. Phys.* 79.2 (1983), pp. 926–935.

[88] Akinori Kato et al. “Characterization of H-box region mutants of WalK inert to the action of waldiomycin in Bacillus subtilis”. In: *The Journal of General and Applied Microbiology* 63.4 (2017), pp. 212–221.

[89] John Keener and Sydney Kustu. “Protein kinase and phosphoprotein phosphatase activities of nitrogen regulatory proteins NTRB and NTRC of enteric bacteria: roles of the conserved amino-terminal domain of NTRC”. In: *Proceedings of the National Academy of Sciences* 85.14 (1988), pp. 4976–4980.

[90] Djaffar Kheffache and Ourida Ouamerali. “Some physicochemical properties of the antitumor drug thiotepa and its metabolite tepa as obtained by density functional theory (DFT) calculations”. In: *Journal of molecular modeling* 16.8 (2010), pp. 1383–1390.

[91] Dong-jin Kim and Steven Forst. “Genomic analysis of the histidine kinase family in bacteria and archaea”. In: *Microbiology* 147.5 (2001), pp. 1197–1212.

[92] Walter Kohn and Lu Jeu Sham. “Self-consistent equations including exchange and correlation effects”. In: *Physical review* 140.4A (1965), A1133.

[93] Pekka Koskinen and Ville Mäkinen. “Density-functional tight-binding for beginners”. In: *Computational Materials Science* 47.1 (2009), pp. 237–253.

[94] Tino Krell et al. “Bacterial sensor kinases: diversity in the recognition of environmental signals”. In: *Annu. Rev. Microbiol.* 64 (2010), pp. 539–559.

[95] T. Kubař. *DFTB+ – modified QM/MM interface*. `https://github.com/tomaskubar/dftbplus`. last accessed 18 March 2022. 2022.

[96] T. Kubař. *Gromacs – QM/MM interface for DFTB+*. `https://github.com/tomaskubar/gromacs-dftbplus`. last accessed 18 March 2022. 2022.

[97] Tomáš Kubař, Kai Welke, and Gerrit Groenhof. “New QM/MM implementation of the DFTB3 method in the Gromacs package”. In: *J. Comput. Chem.* 36.26 (2015), pp. 1978–1989. DOI: `10.1002/jcc.24029`.

[98] Rui Lai and Qiang Cui. "What Does the Brønsted Slope Measure in the Phosphoryl Transfer Transition State?" In: *ACS catalysis* 10.23 (2020), pp. 13932–13945.

[99] Alessandro Laio and Michele Parrinello. "Escaping free-energy minima". In: *Proceedings of the National Academy of Sciences* 99.20 (2002), pp. 12562–12566.

[100] DC Langreth and JP Perdew. "The gradient approximation to the exchange-correlation energy functional: A generalization that works". In: *Solid State Communications* 31.8 (1979), pp. 567–571.

[101] Michael T Laub and Mark Goulian. "Specificity in two-component signal transduction pathways". In: *Annu. Rev. Genet.* 41 (2007), pp. 121–145.

[102] Anne Lecroisey et al. "Phosphorylation mechanism of nucleoside diphosphate kinase: 31P-nuclear magnetic resonance studies". In: *Biochemistry* 34.38 (1995), pp. 12445–12450.

[103] Timothy S Lewis, Paul S Shapiro, and Natalie G Ahn. "Signal transduction through MAP kinase cascades". In: *Advances in cancer research* 74 (1998), pp. 49–139.

[104] Haichen Li et al. "A density functional tight binding layer for deep learning of chemical Hamiltonians". In: *Journal of chemical theory and computation* 14.11 (2018), pp. 5764–5776.

[105] Kresten Lindorff-Larsen et al. "Improved side-chain torsion potentials for the Amber ff99SB protein force field". In: *Proteins: Struct. Funct. Bioinf.* 78.8 (2010), pp. 1950–1958.

[106] Alessio Lodola and Marco De Vivo. "The increasing role of QM/MM in drug discovery". In: *Advances in protein chemistry and structural biology* 87 (2012), pp. 337–362.

[107] Augusto F Lois et al. "Autophosphorylation and phosphatase activities of the oxygen-sensing protein FixL of Rhizobium meliloti are coordinately regulated by oxygen." In: *Journal of Biological Chemistry* 268.6 (1993), pp. 4370–4375.

[108] Xiya Lu et al. "QM/MM free energy simulations: Recent progress and challenges". In: *Molecular simulation* 42.13 (2016), pp. 1056–1078.

[109] P Lykos and GW Pratt. "Discussion on the Hartree-Fock approximation". In: *Reviews of Modern Physics* 35.3 (1963), p. 496.

[110] Maria J van Maanen et al. "A search for new metabolites of N, N, N -triethylenethiophosphoramide". In: *Cancer research* 59.18 (1999), pp. 4720–4724.

[111] Maria J van Maanen et al. "Stability of thioTEPA and its metabolites, TEPA, monochloroTEPA and thioTEPA-mercapturate, in plasma and urine". In: *International Journal of Pharmaceutics* 200.2 (2000), pp. 187–194.

[112] Alberto Marina, Carey D Waldburger, and Wayne A Hendrickson. "Structure of the entire cytoplasmic portion of a sensor histidine-kinase protein". In: *The EMBO journal* 24.24 (2005), pp. 4247–4259.

[113] Franco Marsico et al. “Multiscale approach to the activation and phosphotransfer mechanism of CpxA histidine kinase reveals a tight coupling between conformational and chemical steps”. In: *Biochem. Biophys. Res. Commun.* 498.2 (2018), pp. 305–312.

[114] Kristin L Meagher, Luke T Redman, and Heather A Carlson. “Development of polyphosphate parameters for use with the AMBER force field”. In: *J. Comput. Chem.* 24.9 (2003), pp. 1016–1025.

[115] Clive Metcalfe et al. “Labile Disulfide Bonds Are Common at the Leucocyte Cell Surface”. In: *Open Biol.* 1.3 (2011), p. 110010. DOI: 10.1098/rsob.110010.

[116] Andreas Möglich, Rebecca A Ayers, and Keith Moffat. “Structure and signaling mechanism of Per-ARNT-Sim domains”. In: *Structure* 17.10 (2009), pp. 1282–1294.

[117] Manuel Montenegro et al. “A QM/MM study of the phosphoryl transfer to the Kemptide substrate catalyzed by protein kinase A. The effect of the phosphorylation state of the protein on the mechanism”. In: *Physical Chemistry Chemical Physics* 13.2 (2011), pp. 530–539.

[118] S Morera et al. “Mechanism of phosphate transfer by nucleoside diphosphate kinase: X-ray structures of the phosphohistidine intermediate of the enzymes from Drosophila and Dictyostelium”. In: *Biochemistry* 34.35 (1995), pp. 11062–11070.

[119] Elina Multamäki et al. “Comparative analysis of two paradigm bacteriophytochromes reveals opposite functionalities in two-component signaling”. In: *Nat. Commun.* 12 (2021), p. 4394.

[120] J Munoz-Dorado et al. “Autophosphorylation of nucleoside diphosphate kinase from Myxococcus xanthus”. In: *Journal of bacteriology* 175.4 (1993), pp. 1176–1181.

[121] Kim M Murray et al. “Stability of thiotepa (lyophilized) in 0.9% sodium chloride injection”. In: *American journal of health-system pharmacy* 54.22 (1997), pp. 2588–2591.

[122] Kwangho Nam et al. “Specific reaction parametrization of the AM1/d Hamiltonian for phosphoryl transfer reactions: H, O, and P atoms”. In: *Journal of Chemical Theory and Computation* 3.2 (2007), pp. 486–504.

[123] Rui P. P. Neves et al. “Benchmarking of Density Functionals for the Accurate Description of Thiol–Disulfide Exchange”. In: *J. Chem. Theory Comput.* 10.11 (Nov. 2014), pp. 4842–4856. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/ct500840f. (Visited on 03/29/2021).

[124] Alexander J Ninfa et al. “Crosstalk between bacterial chemotaxis signal transduction proteins and regulators of transcription of the Ntr regulon: evidence that nitrogen assimilation and chemotaxis are controlled by a common phosphotransfer mechanism”. In: *Proceedings of the National Academy of Sciences* 85.15 (1988), pp. 5492–5496.

[125] David M Noll, Tracey McGregor Mason, and Paul S Miller. “Formation and repair of interstrand cross-links in DNA”. In: *Chemical reviews* 106.2 (2006), pp. 277–301.

[126] Ario Okada et al. "Walkmycin B targets WalK (YycG), a histidine kinase essential for bacterial cell growth". In: *J. Antibiot.* 63.2 (2010), pp. 89–94.

[127] Federico A Olivieri et al. "Conformational and Reaction Dynamic Coupling in Histidine Kinases: Insights from Hybrid QM/MM Simulations". In: *J. Chem. Inf. Model.* 60.2 (2020), pp. 833–842.

[128] John S Parkinson and Eric C Kofoid. "Communication modules in bacterial signaling proteins". In: *Annual review of genetics* 26.1 (1992), pp. 71–112.

[129] Robert G Parr. "Density functional theory". In: *Annual Review of Physical Chemistry* 34.1 (1983), pp. 631–656.

[130] JP Perdew. "Orbital functional for exchange and correlation: self-interaction correction to the local density approximation". In: *chemical physics letters* 64.1 (1979), pp. 127–130.

[131] JP Perdew, K Burke, and M Ernzerhof. "Perdew, burke, and ernzerhof reply". In: *Physical Review Letters* 80.4 (1998), p. 891.

[132] Eric F Pettersen et al. "UCSF Chimera -— a visualization system for exploratory research and analysis". In: *J. Comput. Chem.* 25.13 (2004), pp. 1605–1612.

[133] Michael C Pirrung. "Histidine kinases and two-component signal transduction systems". In: *Chem. Biol.* 6.6 (1999), R167–R175.

[134] Dirk Porezag et al. "Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon". In: *Physical Review B* 51.19 (1995), p. 12947.

[135] Marina Putzu et al. "On the Mechanism of Spontaneous Thiol–Disulfide Exchange in Proteins". In: *Phys. Chem. Chem. Phys.* 20.23 (2018), pp. 16222–16230. ISSN: 1463-9076, 1463-9084. DOI: 10.1039/C8CP01325J. URL: http://xlink.rsc.org/?DOI=C8CP01325J (visited on 08/06/2019).

[136] Ling Qin, Takeshi Yoshida, and Masayori Inouye. "The critical role of DNA in the equilibrium between OmpR and phosphorylated OmpR mediated by EnvZ in Escherichia coli". In: *Proceedings of the National Academy of Sciences* 98.3 (2001), pp. 908–913.

[137] Awwad Radwan and Gamal M Mahrous. "Docking studies and molecular dynamics simulations of the binding characteristics of waldiomycin and its methyl ester analog to Staphylococcus aureus histidine kinase". In: *PloS one* 15.6 (2020), e0234215.

[138] Krishnan Raghavachari. "Perspective on "Density functional thermochemistry. III. The role of exact exchange"". In: *Theoretical Chemistry Accounts* 103.3 (2000), pp. 361–363.

[139] Elisa J M Raineri, Dania Altulea, and Jan Maarten van Dijl. "Staphylococcal trafficking and infection—from 'nose to gut' and back". In: *FEMS Microbiol. Rev.* 46.1 (2021), fuab041. DOI: 10.1093/femsre/fuab041.

[140] Paolo Raiteri et al. "Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics". In: *J. Phys. Chem. B* 110.8 (2006), pp. 3533–3539.

[141] Raghunathan Ramakrishnan et al. “Big data meets quantum chemistry approximations: the Δ-machine learning approach”. In: *Journal of chemical theory and computation* 11.5 (2015), pp. 2087–2096.

[142] Rodrigo Recabarren et al. “Mechanistic insights into the phosphoryl transfer reaction in cyclin-dependent kinase 2: A QM/MM study”. In: *PloS one* 14.9 (2019), e0215793.

[143] Demian Riccardi et al. ““Proton Holes” in Long-Range Proton Transfer Reactions in Solution and Enzymes: A Theoretical Analysis”. In: *J. Am. Chem. Soc.* 128.50 (2006), pp. 16302–16311. DOI: 10.1021/ja065451j.

[144] Daniel Roston et al. “Analysis of phosphoryl-transfer enzymes with QM/MM free energy simulations”. In: *Methods in enzymology*. Vol. 607. Elsevier, 2018, pp. 53–90.

[145] Victor Rühle. “Berendsen and nose-hoover thermostats”. In: *Am. J. Phys* (2007), pp. 1–4.

[146] Alexander Schug et al. “High-resolution protein complexes from integrating genomic information with molecular simulation”. In: *Proc. Natl. Acad. Sci. USA* 106.52 (2009), pp. 22124–22129.

[147] Joel C Selcher et al. “Protein kinase signal transduction cascades in mammalian associative conditioning”. In: *The Neuroscientist* 8.2 (2002), pp. 122–131.

[148] Hans Martin Senn and Walter Thiel. “QM/MM methods for biological systems”. In: *Atomistic approaches in modern biology* (2006), pp. 173–290.

[149] S SEVERIN. “E., GEORGIEVSKAYA, EF, AND IVANOV, VI”. In: *Biokhimiya* 12 (1947), p. 35.

[150] Lin Shen and Weitao Yang. “Molecular dynamics simulations with quantum mechanics/molecular mechanics and adaptive neural networks”. In: *Journal of chemical theory and computation* 14.3 (2018), pp. 1442–1455.

[151] Ting Shi et al. “Mechanism for the Autophosphorylation of CheA Histidine Kinase: QM/MM Calculations”. In: *J. Phys. Chem. B* 115.41 (2011), pp. 11895–11901. DOI: 10.1021/jp203968d.

[152] Rajeeva Singh and George M. Whitesides. “Comparisons of Rate Constants for Thiolate-Disulfide Interchange in Water and in Polar Aprotic Solvents Using Dynamic Proton NMR Line Shape Analysis”. In: *J. Am. Chem. Soc.* 112.3 (Jan. 1990), pp. 1190–1197. ISSN: 0002-7863. DOI: 10.1021/ja00159a046. (Visited on 05/05/2021).

[153] Albert Siryaporn and Mark Goulian. “Cross-talk suppression between the CpxA–CpxR and EnvZ–OmpR two-component systems in E. coli”. In: *Molecular microbiology* 70.2 (2008), pp. 494–506.

[154] Albert Siryaporn et al. “Evolving a robust signal transduction pathway from weak cross-talk”. In: *Molecular systems biology* 6.1 (2010), p. 452.

[155] Jeffrey M Skerker et al. “Rewiring the specificity of two-component signal transduction systems”. In: *Cell* 133.6 (2008), pp. 1043–1054.

[156] Jeffrey M Skerker et al. "Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis". In: *PLoS biology* 3.10 (2005), e334.

[157] John C Slater. "A simplification of the Hartree-Fock method". In: *Physical review* 81.3 (1951), p. 385.

[158] Gregory K Smith et al. "Insights into the phosphoryl transfer mechanism of cyclin-dependent protein kinases from ab initio QM/MM free-energy studies". In: *The journal of physical chemistry B* 115.46 (2011), pp. 13713–13722.

[159] Thyagarajan Srikantha et al. "The two-component hybrid kinase regulator CaNIKl of Candida albicans". In: *Microbiology* 144.10 (1998), pp. 2715–2729.

[160] Thomas Steinbrecher and Marcus Elstner. "QM and QM/MM simulations of proteins". In: *Biomolecular Simulations* (2013), pp. 91–124.

[161] David R Stevens and Sharon Hammes-Schiffer. "Exploring the role of the third active site metal ion in DNA polymerase η with QM/MM free energy simulations". In: *Journal of the American Chemical Society* 140.28 (2018), pp. 8965–8969.

[162] Ann M Stock, Victoria L Robinson, and Paul N Goudreau. "Two-component signal transduction". In: *Annual review of biochemistry* 69.1 (2000), pp. 183–215.

[163] JB Stock, AJ Ninfa, and AM372749 Stock. "Protein phosphorylation and regulation of adaptive responses in bacteria". In: *Microbiological reviews* 53.4 (1989), pp. 450–490.

[164] Jeffry B Stock, Ann M Stock, and James M Mottonen. "Signal transduction in bacteria". In: *Nature* 344.6265 (1990), pp. 395–400.

[165] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[166] Hiraku Takada and Hirofumi Yoshikawa. "Essentiality and function of WalK/WalR two-component system: the past, present, and future of research". In: *Biosci. Biotechnol. Biochem.* 82.5 (2018), pp. 741–751.

[167] Edward Teller. "On the stability of molecules in the Thomas-Fermi theory". In: *Reviews of Modern Physics* 34.4 (1962), p. 627.

[168] Hedieh Torabifard and Alireza Fattahi. "DFT study on Thiotepa and Tepa interactions with their DNA receptor". In: *Structural Chemistry* 24.1 (2013), pp. 1–11.

[169] Hedieh Torabifard and Alireza Fattahi. "Mechanisms and kinetics of thiotepa and tepa hydrolysis: DFT study". In: *Journal of molecular modeling* 18.8 (2012), pp. 3563–3576.

[170] Gareth A Tribello et al. "PLUMED 2: New feathers for an old bird". In: *Computer Physics Communications* 185.2 (2014), pp. 604–613.

[171] Nicholas A Turner et al. "Methicillin-resistant *Staphylococcus aureus*: an overview of basic and clinical research". In: *Nat. Rev. Microbiol.* 17.4 (2019), pp. 203–218.

[172] Jeffrey A Ubersax and James E Ferrell Jr. "Mechanisms of specificity in protein phosphorylation". In: *Nature reviews Molecular cell biology* 8.7 (2007), pp. 530–541.

[173] Jeffrey A Ubersax et al. “Targets of the cyclin-dependent kinase Cdk1”. In: *Nature* 425.6960 (2003), pp. 859–864.

[174] Luke E Ulrich and Igor B Zhulin. “The MiST2 database: a comprehensive genomics resource on microbial signal transduction”. In: *Nucleic Acids Res.* 38.suppl_1 (2010), pp. D401–D407.

[175] MJ Van Maanen, CJM Smeets, and JH Beijnen. “Chemistry, pharmacology and pharmacokinetics of N, N, N-triethylenethiophosphoramide (ThioTEPA)”. In: *Cancer treatment reviews* 26.4 (2000), pp. 257–268.

[176] Karl Volz. “Structural conservation in the CheY superfamily”. In: *Biochemistry* 32.44 (1993), pp. 11741–11753.

[177] Chen Wang et al. “Mechanistic insights revealed by the crystal structure of a histidine kinase with signal transducer and sensor domains”. In: *PLoS Biol.* 11.2 (2013), e1001493.

[178] Junmei Wang et al. “Antechamber: an accessory software package for molecular mechanical calculations”. In: *J. Am. Chem. Soc* 222 (2001), U403.

[179] Loo Chien Wang et al. “The inner membrane histidine kinase EnvZ senses osmolality via helix-coil transitions in the cytoplasm”. In: *The EMBO journal* 31.11 (2012), pp. 2648–2659.

[180] Benjamin Webb and Andrej Sali. “Comparative protein structure modeling using MODELLER”. In: *Current Protocols in Bioinformatics* 54.1 (2016), pp. 5–6.

[181] Philip A Webb et al. “The crystal structure of human nucleoside diphosphate kinase, NM23-H2”. In: *Journal of molecular biology* 251.4 (1995), pp. 574–587.

[182] Christopher Weidenmaier, Christiane Goerke, and Christiane Wolz. “*Staphylococcus aureus* determinants for nasal colonization”. In: *Trends Microbiol.* 20.5 (2012), pp. 243–250.

[183] Ann H West and Ann M Stock. “Histidine kinases and response regulator proteins in two-component signaling systems”. In: *Trends in biochemical sciences* 26.6 (2001), pp. 369–376.

[184] Janet M. Wilson, Robert J. Bayer, and D. J. Hupe. “Structure-Reactivity Correlations for the Thiol-Disulfide Interchange Reaction”. In: *J. Am. Chem. Soc.* 99.24 (Nov. 1977), pp. 7922–7926. ISSN: 0002-7863. DOI: 10.1021/ja00466a027. URL: https://doi.org/10.1021/ja00466a027 (visited on 02/23/2021).

[185] Peter M Wolanin, Peter A Thomason, and Jeffry B Stock. “Histidine protein kinases: key signal transducers outside the animal kingdom”. In: *Genome Biol.* 3.10 (2002), p. 3013.1.

[186] Steven N Wolff et al. “High-dose N, N’, N"-triethylenethiophosphoramide (thiotepa) with autologous bone marrow transplantation: phase I studies.” In: *Seminars in Oncology*. Vol. 17. 1 Suppl 3. 1990, pp. 2–6.

[187] Lu Yu et al. “Role of Mg^{2+} ions in protein kinase phosphorylation: insights from molecular dynamics simulations of ATP-kinase complexes”. In: *Mol. Simul.* 37.14 (2011), pp. 1143–1150. DOI: 10.1080/08927022.2011.561430.

[188] Jingjing Zheng, Xuefei Xu, and Donald G Truhlar. "Minimally augmented Karlsruhe basis sets". In: *Theoretical Chemistry Accounts* 128.3 (2011), pp. 295–305.

[189] Christopher P Zschiedrich, Victoria Keidel, and Hendrik Szurmant. "Molecular mechanisms of two-component signal transduction". In: *J. Mol. Biol.* 428.19 (2016), pp. 3752–3775.

A. Appendix

A.1. Chapter 8

Metadynamics parameters for model reactions

- CV used: P–N and P–O Distances
- Gaussian height 0.2 kj/mol, Gaussian width 0.02
- 16 Walkers used, walker stride: 500
- bias frequency: 100

Restraints

- P–N Distance was restrained to values lower than 3.5 Å
- P–O Distance was restrained to values lower than 3.5 Å
- N–P–O Angle P–N Distance was restrained to values at 178°

Metadynamics parameters for TEPA reaction

- CV used: 1) phos-nitro_hyd[P–N distance – P–O distance], 2) N–H distance
- initial Gaussian height 2.8 kj/mol, Gaussian width 0.02, bias factor 65
- 24 Walkers used, using walker stride: 800
- bias frequency: 500

Restraints

- P–N Distance was restrained to values lower than 5.0 Å
- P–O Distance was restrained to values lower than 5.0 Å
- N–H Distance was restrained to values lower than 3.0 Å
- All other Nitrogens in the TEPA ring was restricted with protonations from water

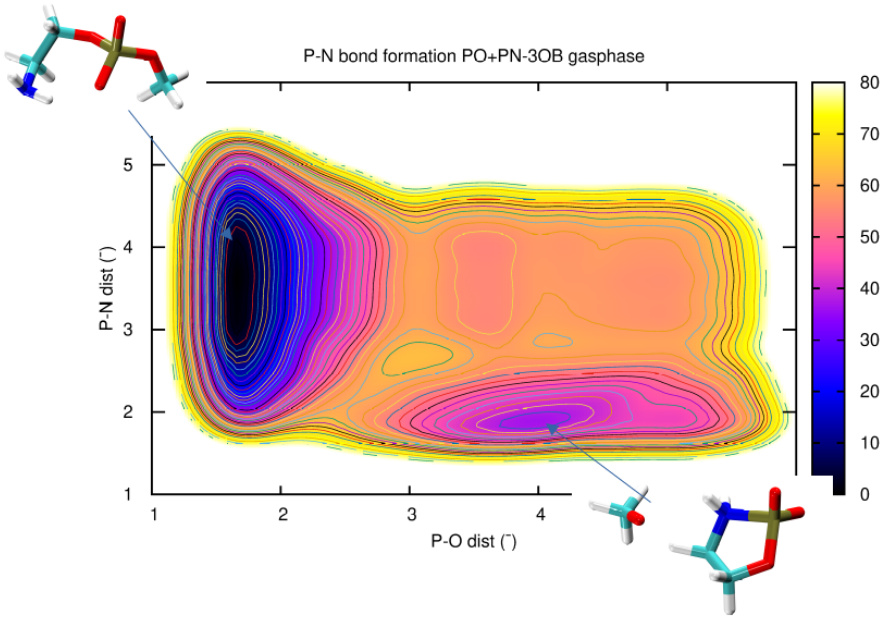

Figure A.1.: P-N bond formation in gasphase, shown for one of the benchmark reaction, transition state is visible

.

A.2. Chapter 9

Restraints

The following additional harmonic restrains were applied in the QM/MM simulation:

- All O–H bonds of QM water molecules were restrained to 0.1 nm length with a force constant of 15,000 kJ mol^{-1} nm^{-2}.
- The angle Nδ(His391)–P(γ-phosphate of ATP)–O(β-phosphate of ATP) was restrained to values higher than 172° with a force constant of 1500 kJ mol^{-1} rad^{-2} ('lower wall' of PLUMED).
- The distance P(ɣ-phosphate of ATP)–Nε(His391) was restrained to values lower than 0.35 nm ('upper wall' of PLUMED) with a force constant of 15,000 kJ mol^{-1} nm^{-2}.
- The distance P(ɣ-phosphate of ATP)–O(β-phosphate of ATP) was restrained to values lower than 0.40 nm ('upper wall') with a force constant of 15,000 kJ mol^{-1} nm^{-2}.
- In the simulation of System 1: the proton transfer CV (N–H–O antisymmetric stretch) was restrained to the interval between −0.2 and 0.7 nm ('lower' and 'upper walls') with a force constant of 1500 kJ mol^{-1} nm^{-2}.
- In the simulation of System 2: the proton transfer CV (N–H–O antisymmetric stretch) was restrained to the interval between −0.2 and 0.2 nm ('lower' and 'upper walls') with a force constant of 1500 kJ mol^{-1} nm^{-2}.

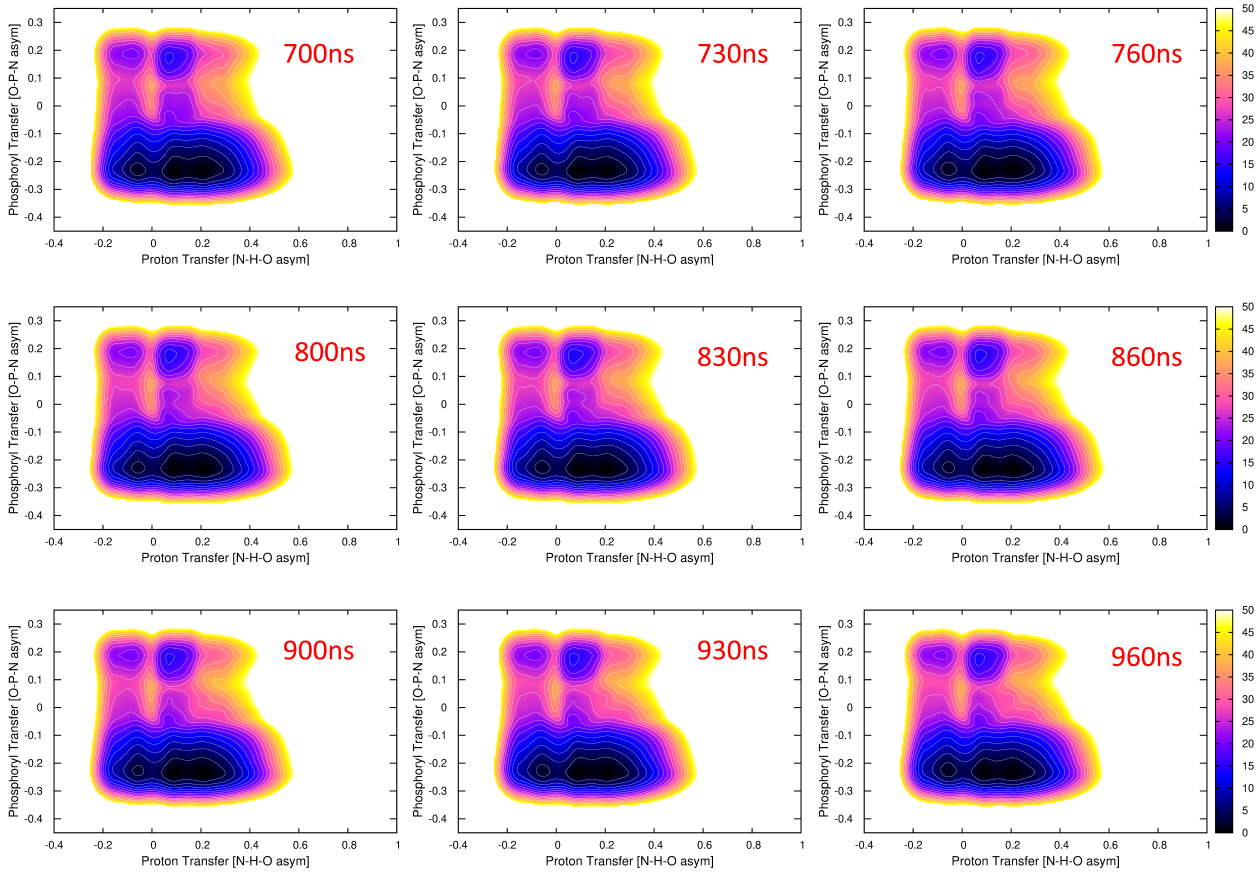

Figure A.2.: Convergence of the potentials of the mean force in the QM/MM metadynamics simulation of the chemical step of the autophosphorylation, considering a hydroxyl ion as the proton acceptor. Distances in nm, free energies color-coded in kcal/mol.H

.

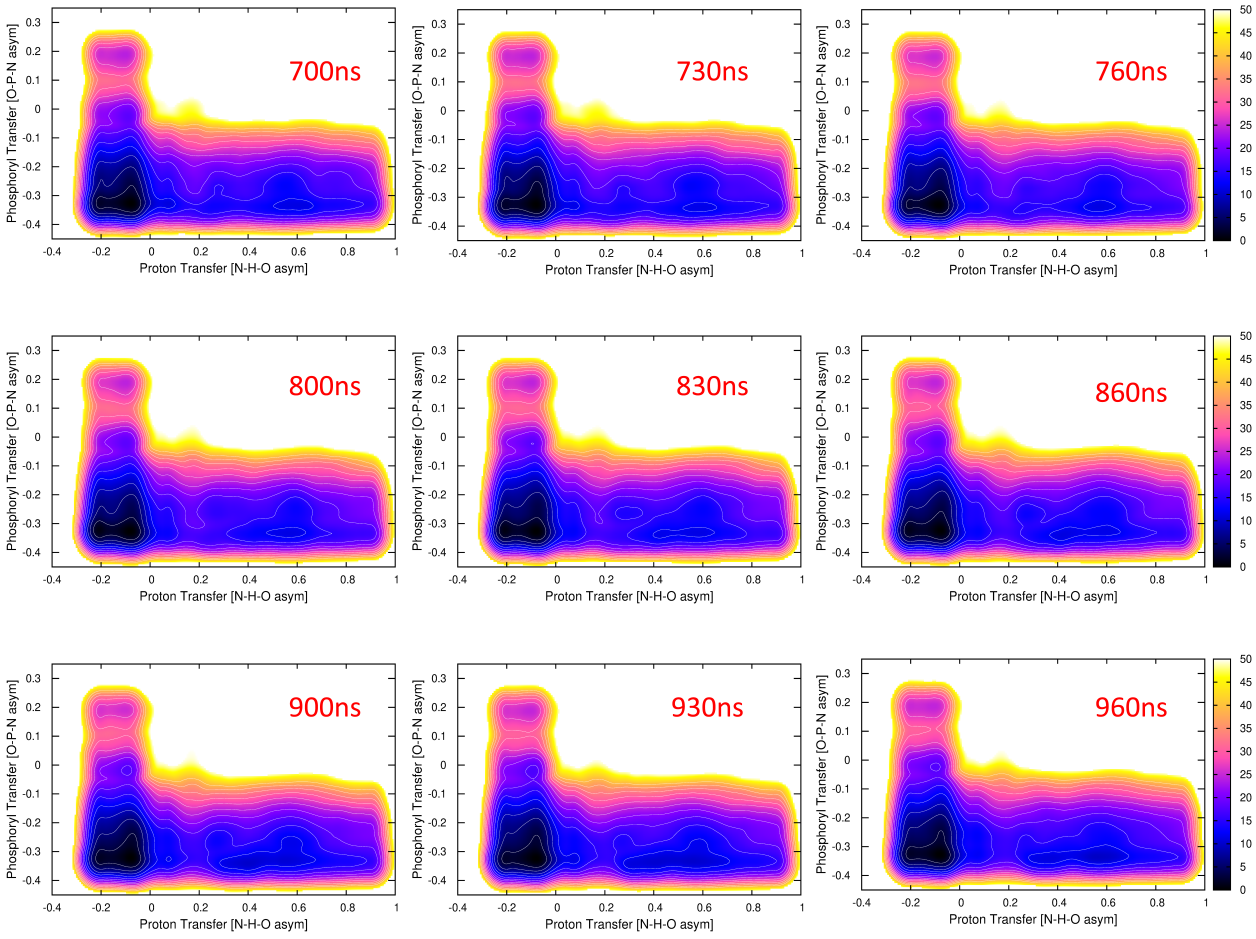

Figure A.3.: Convergence of the potentials of the mean force in the QM/MM metadynamics simulation of the chemical step of the autophosphorylation, considering the side chain of Glu392 as the proton acceptor. Distances in nm, free energies color-coded in kcal/mol

.

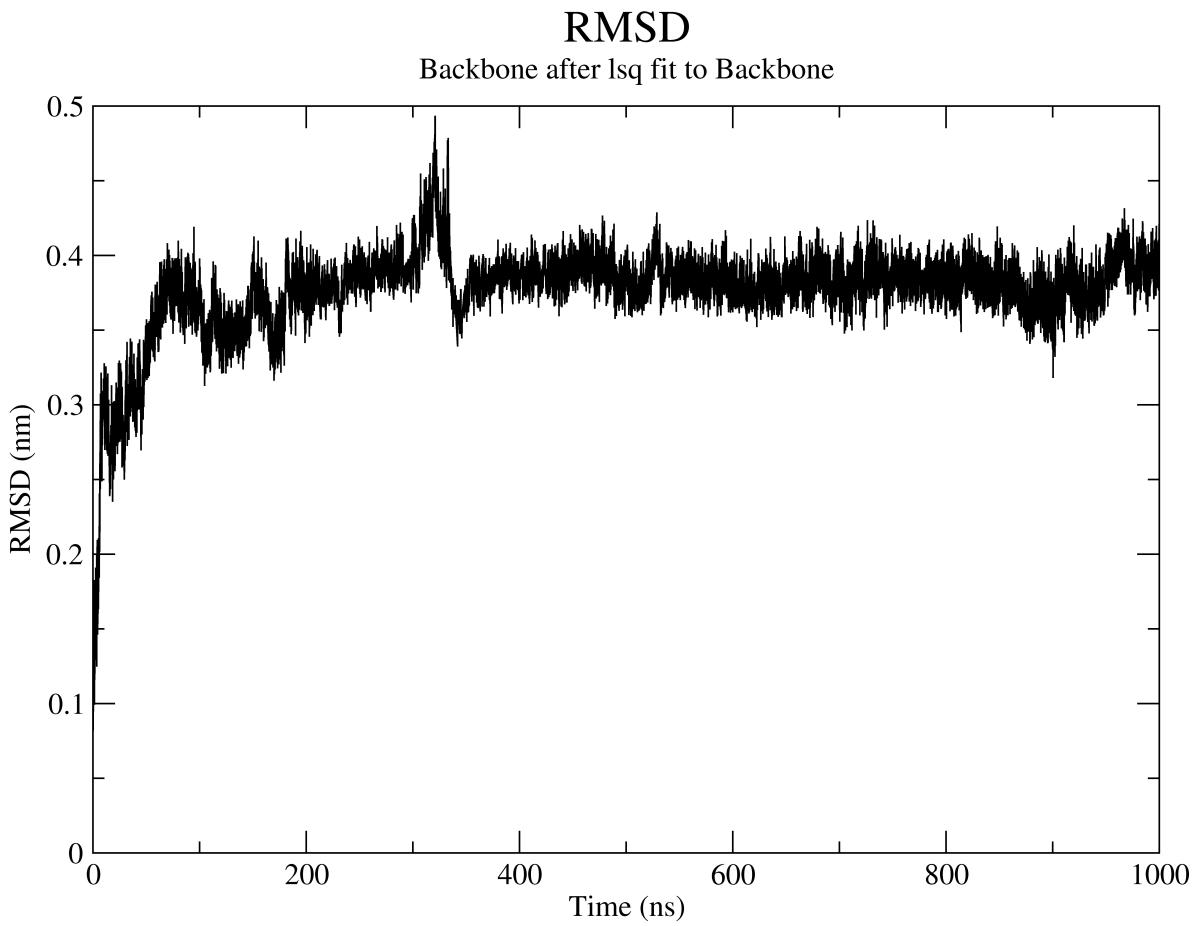

Figure A.4.: 1 Microsecond RMSD of autophosphorylated cis-kinase bound with ADP

.

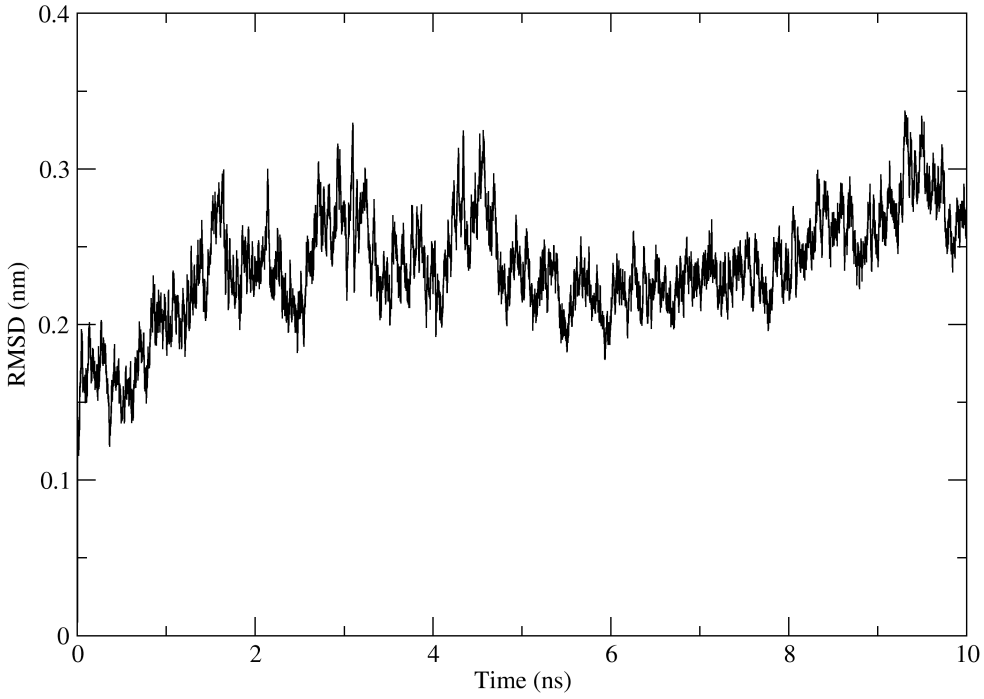

Figure A.5.: RMSD of trans histidine kinase

.

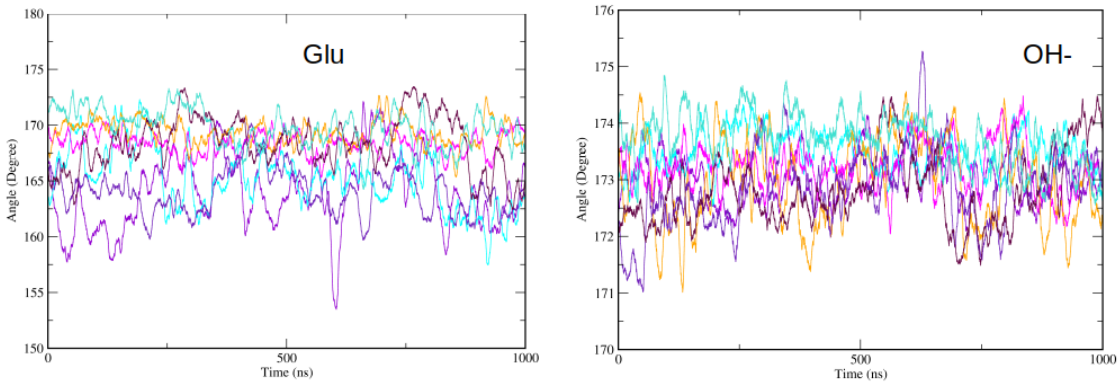

Figure A.6.: P-N-O angle(reaction angle), shown for both OH- and Glu assisted proton transfer simulation in first few walkers

.

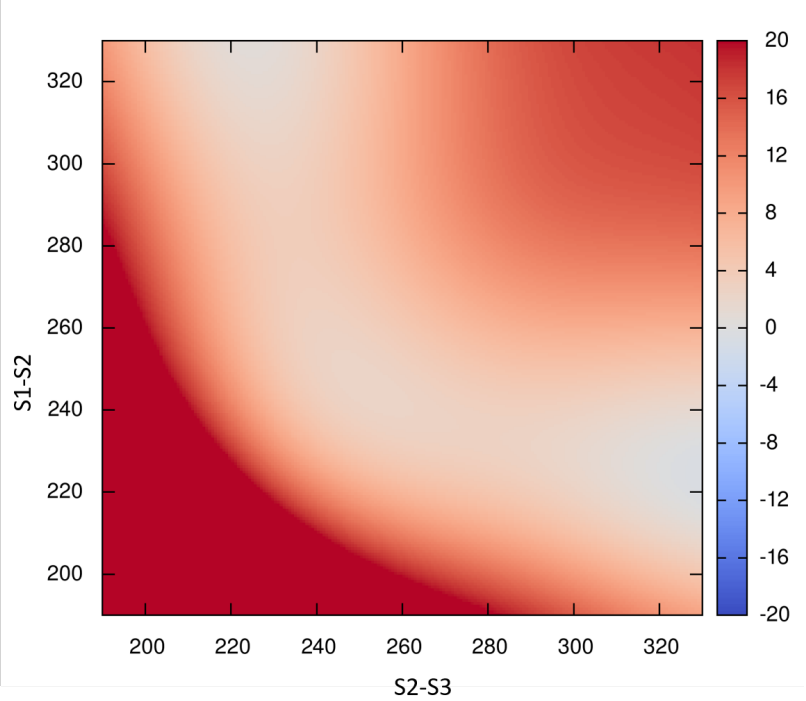

Figure A.7.: Potential energy plot, obtained from the repulsive spline where no overbinding enrgy used

.

Erklärung zur Dissertation

Ich erkläre hiermit, dass ich die vorliegende Dissertation selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Weiterhin versichere ich, dass ich die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis beachtet habe.

Mayukh Kansari

01.06.2022