Master's thesis

Master's Programme in Theoretical and Computational Methods

Investigation of Proton Transfer Reactions with Hybrid QM/MM Free Energy Calculations

Mahdi Torabi

12th March 2026

Supervisor(s): Prof. Vivek Sharma
Dr. Oleksii Zdorevskyi

Examiner(s): Prof. Vivek Sharma
Dr. Waldemar Kulig

University of Helsinki
Faculty of Science

P. O. Box 64 (Gustaf Hällströmin katu 2)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Koulutusohjelma — Utbildningsprogram — Degree programme
Faculty of Science	Master's Programme in Theoretical and Computational Methods

Tekijä — Författare — Author

Mahdi Torabi

Työn nimi — Arbetets titel — Title

Investigation of Proton Transfer Reactions with Hybrid QM/MM Free Energy Calculations

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidantal — Number of pages
Master's thesis	12th March 2026	85

Tiivistelmä — Referat — Abstract

Proton transfer (PT) reactions play an important role in biology. Understanding PT pathways and energetics can help uncover the molecular function of energy-converting enzymes. To study PT reactions in a complex protein environment, a powerful technique, hybrid quantum mechanics/molecular mechanics (QM/MM), can be employed as it models both chemical reactions and the protein environment, offering a reasonable compromise between accuracy and computational cost. However, conventional QM/MM does not capture slow PT reactions. Therefore, it is often combined with various enhanced sampling techniques. Nonetheless, the choice of enhanced sampling method is often ambiguous and can produce unexpectedly diverse results depending on the calculation type and the reaction coordinate. Therefore, the first goal of this thesis is to optimize available enhanced sampling approaches for studying PT reactions within the QM/MM framework. The second goal is to investigate how the choice of reaction coordinate affects the resulting free energy surface.

One of the commonly used enhanced sampling methods is metadynamics. Due to low computational cost, metadynamics is an attractive choice for studying PT reactions. By choosing a particular PT pathway, the optimal parameters for metadynamics within the QM/MM framework are determined. A Gaussian height of 0.3 kcal/mol, a Gaussian width of 0.3 Å and a deposition rate of 10 fs are optimal parameters for studying the energetics of PT reactions. Using these parameters, plausible free energy profiles that agree with umbrella sampling simulations are obtained. Furthermore, parameters are validated across different PT pathways with varying QM-region compositions.

In addition, the modified center of excess charge (mCEC) collective variable (CV) is implemented and verified on various PT pathways. The resulting free energy surface is compared with those obtained using a linear combination of hydrogen-bonding distances (LinComb) CV. The results show that mCEC can exhibit more favorable energetics for PT reactions, allowing greater flexibility in possible proton pathways than the LinComb CV.

The results of this thesis help researchers computationally investigate PT reactions at lower computational cost and avoid problems that might arise from using the LinComb CV by employing the mCEC CV as the reaction coordinate.

Avainsanat — Nyckelord — Keywords

Respiratory Complex I, Metadynamics, Umbrella Sampling, Center of Excess Charge

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Rarely do we arrive at the summit of truth without running into extremes; we have frequently to exhaust the part of error, and even of folly, before we work our way up to the noble goal of tranquil wisdom.

— Friedrich Schiller

Preface

I thank the Center for Scientific Computing (CSC), Finland, for giving access to large-scale computational resources. I appreciate the University of Helsinki for all the facilities that made this work possible. I thank Dr. Erik Endres and Luka Simsive for providing the data needed to perform simulations from their respective projects. I thank Dr. Oleksii Zdorevskyi for his help, support, and excellent discussions during the entire thesis work. I'm grateful to Prof. Vivek Sharma for his supervision, guidance, and precious advice. I thank my family for their continued support and all my friends for their joyful presence and rewarding coffee-break discussions.

Helsinki, 12th March 2026

Mahdi Torabi

Contents

1. Introduction

Proton solvation and transfer play an important role in many areas, such as material science, where the motion of protons is central in hydrogen fuel cells, and in atmospheric science, where the acidified aerosols impose health risks on people [1]. In biology, pH gradients across lipid bilayers, formed by virtue of excess protons on one side, provide an effective means for biochemical energy storage [2]. In many enzymes facilitating chemical reactions, proton transfer (PT) can occur between distant active sites. This long-range PT occurs along a series of hydrogen bonds created by intervening water molecules or titratable protein residues and plays a crucial role in biology, for example, in ATP synthesis [3].

Although the mechanism of PT reactions localized to the enzyme active site is well studied, long-range PT in biological contexts remains difficult to quantify [3]. Experimental methods based on mutations and kinetic measurements can be used to study PT pathways. However, the results can be ambiguous and difficult to interpret due to perturbation of the protein and water molecules introduced by mutations [3]. PT reactions are described by the Grotthuss mechanism [4]. The excess charge is transferred not as a single proton diffusion, but through the set of proton hopping events from the donor to the acceptor along the water chain [1].

To acquire atomistic-level information and gain insights into the mechanism of PT across various biomolecular systems, theoretical and computational methods are employed. The phenomenon of PT involves bond formation/breakage, which can be accurately described by quantum mechanics (QM). As large biological systems are beyond the reach of a full QM treatment, the hybrid method of quantum mechanics/molecular mechanics (QM/MM) must be employed to decrease the computational load [5, 6]. In this framework, only a small number of atoms are treated quantum-mechanically, while the highly scalable molecular mechanics (MM) method accounts for the remainder of the system. To track the dynamical evolution of the system, molecular dynamics (MD) is used to obtain atomic trajectories from Newton's equations of motion. A QM/MM molecular dynamics simulation alone is insufficient, as sampling the PT is often infeasible on the simulation timescale. Therefore, enhanced sampling methods have to be used. There are various enhanced sampling methods, such as replica-exchange [7],

adaptive biasing force [8], umbrella sampling (US) [9, 10], metadynamics (MetaD) [11], and more. MetaD has the advantage of exploring low-energy regions first and does not require the knowledge of the "exact" reaction path [12].

The input parameters in a MetaD simulation are not optimized for the QM/MM simulations involving PT. Therefore, the first goal of this research is to optimize the input parameters to construct the underlying free energy surface (FES) of the PT in a specific scenario with an optimal balance between accuracy and computational cost. Furthermore, in the study of PT, the choice of the reaction coordinate (RC) is not uniquely determined, and there is flexibility in selecting a particular collective variable (CV) as a RC. Choosing an appropriate RC is challenging because long-range PT reaction pathways are often non-unique or reform during biased MD simulations with different waters. Therefore, in the second part of the research, a collective variable, the modified center of excess charge (mCEC), is introduced, which serves as a RC for long-range PT, since the exact pathway is not specified, and the proton is transferred along the optimal path. The implementation is verified via both the US and the MetaD enhanced sampling methods.

This thesis has the following structure. In Chapter 2, the methods used to treat the system of interest are described in detail. This includes QM, MM, QM/MM, and enhanced sampling methods such as US and MetaD. In Chapter 3, a brief overview of the proton transfer in solution and in biomolecules is presented, followed by a brief description of the biological system of interest, respiratory complex I. In Chapter 4, the results of investigations are presented and discussed in detail with concluding remarks in Chapter 5.

2. Computational Methods

2.1 Quantum Mechanics

Quantum mechanics (QM) is a fundamental theory that accurately describes various types of matter and interactions at the atomic scale [13]. To describe the atomic structure and consequently investigate its properties the solution of the Schrödinger equation of the system is required. In QM, the state of the system is described by a wave function of all the nuclei and electrons in the system $\Psi(\mathbf{R}_1, \cdots, \mathbf{R}_N, \mathbf{r}_1, \cdots, \mathbf{r}_n, t)$, where N is the number of nuclei and n the number of electrons, and t is time. This wave function is the solution of a wave equation, called the time-dependent Schrödinger equation [13]:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{R}_1, \ldots, \mathbf{R}_N, \mathbf{r}_1, \ldots, \mathbf{r}_n, t) = \hat{\mathcal{H}}\Psi(\mathbf{R}_1, \ldots, \mathbf{R}_N, \mathbf{r}_1, \ldots, \mathbf{r}_n, t), \tag{2.1}$$

where $i = \sqrt{-1}$, $\hbar$ is the reduced Planck's constant equal to $\frac{h}{2\pi}$, and the Hamiltonian operator is [13]:

$$\hat{\mathcal{H}} = \sum_{A=1}^{N}\frac{\hat{P}_A^2}{2M_A} + \sum_{i=1}^{n}\frac{\hat{p}_i^2}{2m_i} - \sum_{i=1}^{n}\sum_{A=1}^{N}\frac{Z_A e^2}{\|\mathbf{R}_A - \mathbf{r}_i\|} + \sum_{A<B}^{N}\frac{Z_A Z_B e^2}{\|\mathbf{R}_A - \mathbf{R}_B\|} + \sum_{i<j}^{n}\frac{e^2}{\|\mathbf{r}_i - \mathbf{r}_j\|}. \tag{2.2}$$

The first term is the sum over the kinetic energy of all the nuclei where $\hat{P}_A$ and M_A are the momentum operator and the mass of the A^{th} nucleus, respectively. The second term is the sum over the kinetic energy of all the electrons, where $\hat{p}_i$ and m_i are the momentum operator and the mass of the i^{th} electron, respectively. The third term is the nucleus-electron Coulomb potential energy, where $\mathbf{R}_A$, $\mathbf{r}_i$ are the positions of the A^{th} nucleus and the i^{th} electron, respectively. Z_A is the atomic number of the A^{th} nucleus and e is the charge of an electron. $\|\cdot\|$ is the norm of the vector. The remaining two terms are the nucleus-nucleus and electron-electron Coulomb potential energy, respectively [13].

Solving the Schrödinger equation to get the wave function is a task of immense difficulty for complex molecules and large systems. Therefore, several approximations

are suggested. The first well-known approximation is the Born–Oppenheimer (BO) approximation [14], based on the fact that the proton-to-electron mass ratio is approximately 1836, meaning that electrons move much faster than the nuclei. This allows the separation of the Hamiltonian into a nuclear and an electronic part, since the nuclear-electronic interaction potential now has only parametric dependence on $\mathbf{R}$. The expression for the electronic Hamiltonian is as follows [13]:

$$\begin{aligned}\hat{\mathcal{H}}_e &= \sum_{i=1}^{n} \frac{\hat{p}_i^2}{2m_i} - \sum_{i=1}^{n}\sum_{A=1}^{N} \frac{Z_A e^2}{\|\mathbf{R}_A - \mathbf{r}_i\|} + \sum_{A<B}^{N} \frac{Z_A Z_B e^2}{\|\mathbf{R}_A - \mathbf{R}_B\|} + \sum_{i<j}^{n} \frac{e^2}{\|\mathbf{r}_i - \mathbf{r}_j\|} \\ &= \hat{T}_e(\mathbf{R}) - \hat{V}_{eN}(\mathbf{r};\mathbf{R}) + \hat{V}_{NN}(\mathbf{R}) + \hat{V}_{ee}(\mathbf{R}),\end{aligned} \tag{2.3}$$

where $\mathbf{R}$ is now a fixed parameter. $\hat{T}_e(\mathbf{R})$ is the total electron kinetic energy operator, $\hat{V}_{eN}(\mathbf{r};\mathbf{R})$ is the nuclear-electron Coulomb potential operator, $\hat{V}_{NN}(\mathbf{R})$ is the nuclear-nuclear interaction potential operator, and $\hat{V}_{ee}(\mathbf{R})$ is the electron-electron Coulomb interaction operator. The importance of BO approximation comes from the fact that now the total wave function itself can be separated into two wave functions, one electronic Φ_e and the other nuclear ϕ_n [13]:

$$\Psi(\mathbf{R}_1,\ldots,\mathbf{R}_N,\mathbf{r}_1,\ldots,\mathbf{r}_n,t) = \Phi_e(\mathbf{r}_1,\ldots,\mathbf{r}_n;\mathbf{R}_1,\ldots,\mathbf{R}_N)\phi_n(\mathbf{R}_1,\ldots,\mathbf{R}_N,t). \tag{2.4}$$

Using the electronic Hamiltonian and the electronic wave function the time-independent electronic Schrödinger equation is defined as follows [13]:

$$\hat{\mathcal{H}}_e\Phi_e(\mathbf{r}_1,\ldots,\mathbf{r}_n;\mathbf{R}_1,\ldots,\mathbf{R}_N) = \mathcal{E}_{\mathrm{BO}}(\mathbf{R}_1,\ldots,\mathbf{R}_N)\Phi_e(\mathbf{r}_1,\ldots,\mathbf{r}_n;\mathbf{R}_1,\ldots,\mathbf{R}_N). \tag{2.5}$$

The solution to Eq. (2.5) is the ground state electronic wave function that minimizes the BO energy $\mathcal{E}_{\mathrm{BO}}$ [13]. The nuclear Schrödinger equation, on the other hand, is [13]:

$$i\hbar\frac{\partial}{\partial t}\phi_n(\mathbf{R}_1,\ldots,\mathbf{R}_N,t) = \hat{\mathcal{H}}_n\phi_n(\mathbf{R}_1,\ldots,\mathbf{R}_N,t), \tag{2.6}$$

where the nuclear Hamiltonian is defined as follows:

$$\hat{\mathcal{H}}_n = \sum_{A=1}^{N} \frac{\hat{P}_A^2}{2M_A} + \mathcal{E}_{\mathrm{BO}}(\mathbf{R}_1,\ldots,\mathbf{R}_N). \tag{2.7}$$

From Eqs. (2.6) and (2.7), it is evident that $\mathcal{E}_{\mathrm{BO}}$ acts as a potential energy surface that determines the nuclear motion [13].

2.1.1 Hartree-Fock Method

The electronic Schrödinger equation is only solvable analytically for one-electron systems such as the hydrogen atom and He^+. Solving Eq. (2.5) for real systems that are made of a considerable number of atoms can only be carried out after further approximations.

Within the Hartree-Fock (HF) method, the Hartree product approximation [15] assumes that electrons move independently from each other and experience the other electrons in an average manner (mean field), neglecting correlation effects. This allows writing the electronic wave function as the product of one-electron functions $\psi(\mathbf{r})$ [16]:

$$\Phi_{HP}(\mathbf{r}_1, \ldots, \mathbf{r}_n) = \psi_1(\mathbf{r}_1)\psi_2(\mathbf{r}_2)\ldots\psi_n(\mathbf{r}_n), \tag{2.8}$$

where n is the number of electrons, and the parametric dependence on $\mathbf{R}$ is dropped from the notation for convenience. Introducing spin states to Eq. (2.8), the wavefunction is rewritten in terms of space-spin coordinates. Space-spin coordinates denote the three spatial coordinates and the intrinsic spin denoted by $\mathbf{x} = (\mathbf{r}, \sigma)$ where σ can be α or β (i.e., up or down). So the Hartree product for the wave function defined on the space-spin coordinates is:

$$\tilde{\Phi}_{HP}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \varphi_1(\mathbf{x}_1)\varphi_2(\mathbf{x}_2)\ldots\varphi_n(\mathbf{x}_n). \tag{2.9}$$

The caveat here is that this wave function is not antisymmetric. Antisymmetry means that the wave function of fermions (e.g., electrons) must be equal to its negative if the space-spin coordinates of two electrons are interchanged. To make Eq. (2.9) antisymmetric, the Slater determinant [17] must be employed. In this case, the wave function is written as follows:

$$\tilde{\Phi} = \frac{1}{\sqrt{n!}} \begin{vmatrix} \varphi_1(\mathbf{x}_1) & \varphi_2(\mathbf{x}_1) & \cdots & \varphi_n(\mathbf{x}_1) \\ \varphi_1(\mathbf{x}_2) & \varphi_2(\mathbf{x}_2) & \cdots & \varphi_n(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1(\mathbf{x}_n) & \varphi_2(\mathbf{x}_n) & \cdots & \varphi_n(\mathbf{x}_n) \end{vmatrix}, \tag{2.10}$$

where the fraction before the determinant is for normalization. The electronic Hamiltonian can be rewritten as follows [16]:

$$\hat{\mathcal{H}}_e = \sum_i \hat{h}(i) + \sum_{i<j} \hat{v}(i,j) + \hat{V}_{NN}\,, \tag{2.11}$$

where $\hat{h}(i) = -\frac{1}{2}\nabla_i^2 - \sum_A \frac{Z_A}{r_{iA}}$ is the one-electron operator and $\hat{v}(i,j) = r_{ij}^{-1}$ is the two electron operator in atomic units. $\hat{V}_{NN}$ is a function of $\mathbf{R}$, and here $\mathbf{R}$ is just a parameter and therefore this term only shifts the energy by a constant [16]. Now the energy can be calculated as follows:

$$\mathcal{E}_{\mathrm{HF}} = \langle\tilde{\Phi}|\hat{\mathcal{H}}_e|\tilde{\Phi}\rangle\,. \tag{2.12}$$

To obtain the energy $\mathcal{E}_{\mathrm{HF}}$, the variational method is employed to minimize the energy of the system to get the wave function that best describes the state of the system.

Employing this method yields the following equation determining the one-electron orbitals that minimize the energy:

$$\hat{h}(\mathbf{x}_1)\varphi_i(\mathbf{x}_1) + \sum_{j\neq i}\left[\int d\mathbf{x}_2\,|\varphi_j(\mathbf{x}_2)|^2 r_{12}^{-1}\right]\varphi_i(\mathbf{x}_1) - \sum_{j\neq i}\left[\int d\mathbf{x}_2\,\varphi_j^*(\mathbf{x}_2)\varphi_i(\mathbf{x}_2) r_{12}^{-1}\right]\varphi_j(\mathbf{x}_1) = \epsilon_i\,\varphi_i(\mathbf{x}_1). \tag{2.13}$$

Eq. (2.13) can be written in terms of operators acting on $\varphi_i(\mathbf{x}_1)$. The first operator is the Coulomb operator:

$$\hat{\mathcal{J}}_j(\mathbf{x}_1) = \int d\mathbf{x}_2\,|\varphi_j(\mathbf{x}_2)|^2 r_{12}^{-1}\,, \tag{2.14}$$

and the second operator is the exchange operator:

$$\hat{\mathcal{K}}_j(\mathbf{x}_1)\varphi_i(\mathbf{x}_1) = \left[\int d\mathbf{x}_2\,\varphi_j^*(\mathbf{x}_2)\varphi_i(\mathbf{x}_2) r_{12}^{-1}\right]\varphi_j(\mathbf{x}_1). \tag{2.15}$$

Furthermore, $\hat{\mathcal{J}}$ and $\hat{\mathcal{K}}$ operators are combined into a new operator called the Fock operator:

$$\hat{f}(\mathbf{x}_1) = \hat{h}(\mathbf{x}_1) + \sum_j \hat{\mathcal{J}}_j(\mathbf{x}_1) - \hat{\mathcal{K}}_j(\mathbf{x}_1), \tag{2.16}$$

transforming Eq. (2.13) into the following compact form:

$$\hat{f}(\mathbf{x}_1)\varphi_i(\mathbf{x}_1) = \epsilon_i\varphi_i(\mathbf{x}_1). \tag{2.17}$$

To solve Eq. (2.17), the orbitals can be expanded in terms of a basis set, where $\{\chi_\mu\}$ is a set of orbitals centered on individual atoms (atomic orbitals):

$$\varphi_i = \sum_{\mu=1}^{K} C_{\mu i}\chi_\mu\,, \tag{2.18}$$

where K determines the number of basis functions used and $C_{\mu i}$ are the coeeficients determined by energy minimization. Substituting Eq. (2.18) in Eq. (2.17) gives:

$$\hat{f}(\mathbf{x}_1)\sum_\nu C_{\nu i}\chi_\nu(\mathbf{x}_1) = \epsilon_i\sum_\nu C_{\nu i}\chi_\nu(\mathbf{x}_1). \tag{2.19}$$

This is an operator equation which can be transformed into a matrix equation by multiplying the equation from the left by χ_μ^* and integrating:

$$\sum_\nu C_{\nu i}\int dx_1\chi_\mu^*(\mathbf{x}_1)\hat{f}(\mathbf{x}_1)\chi_\nu(\mathbf{x}_1) = \epsilon_i\sum_\nu C_{\nu i}\int dx_1\chi_\mu^*(\mathbf{x}_1)\chi_\nu(\mathbf{x}_1), \tag{2.20}$$

leading to $\sum_\nu F_{\mu\nu}C_{\nu i} = \epsilon_i\sum_\nu S_{\mu\nu}C_{\nu i}$, where it can be written as:

$$\mathbf{FC} = \mathbf{SC}\epsilon. \tag{2.21}$$

This is a generalized eigenvalue equation that can be solved using linear algebra routines. The operator matrix $\mathbf{F}$ depends on its solution, and that is why it has to be solved iteratively, leading to what is called the self-consistent-field method [16].

Although the HF method is considered a breakthrough in quantum chemistry, the neglect of the correlation effects (i.e., an electron only experiences the average field generated by the other electrons) made the HF method unsuitable for even simple chemical reactions [18]. To take into account correlation energy, post-HF methods [19] are developed, with the disadvantage of being computationally expensive (scaling larger than $O(N^4)$) [20]. To describe PT in biomolecules, however, a more accurate approach than HF is needed that includes electron correlation while being less demanding on computational resources. Density functional theory (DFT) is an alternative to HF and post-HF methods, offering a reasonable balance between accuracy and speed.

2.1.2 Density Functional Theory

DFT attempts to calculate molecular properties in terms of the density, thereby providing a different perspective from wavefunction methods, such as Hartree-Fock (HF). The electron density can be calculated by integrating the absolute-squared electronic wavefunction $\Phi_e(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_n)$ [21]:

$$\rho(\mathbf{r}) = n \int \mathrm{d}\mathbf{r}_2 \mathrm{d}\mathbf{r}_3 \ldots \mathrm{d}\mathbf{r}_n \, |\Phi_e(\mathbf{r}, \mathbf{r}_2, \mathbf{r}_3, \ldots, \mathbf{r}_n)|^2. \tag{2.22}$$

The very first attempt to model kinetic and exchange energy was made by Thomas, Fermi, and Dirac, where a uniform electron gas energy density was employed, setting the foundation for the first exchange functionals [22, 23, 24]. The electronic Hamiltonian of n electrons is defined in Eq. (2.3). The Hamiltonian in this case is fully determined by the number and positions of electrons, and the positions and charge of the nuclei. Consequently, the ground-state wave function, energy, and density are obtainable [25].

In 1964, Hohenberg and Kohn [26] proved two theorems that are essential for putting DFT on a firm basis. The first theorem states that the electron density uniquely determines the energy functional. The second is that the energy from an approximate density cannot be lower than the exact ground-state energy. This means that the search for the electron density that minimizes the energy must be carried out, since the optimal electron density does not change the energy when varied [27]. The next breakthrough came with the introduction of the local density approximation by Kohn and Sham [28]. They first wrote the ground state energy of an inhomogeneous gas in a stationary potential $v(\mathbf{r})$ as follows:

$$E = \int v(\mathbf{r})\rho(\mathbf{r})\mathrm{d}\mathbf{r} + \frac{1}{2} \iint \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}\mathbf{r}\mathrm{d}\mathbf{r}' + T_\mathrm{s}[\rho] + E_\mathrm{xc}[\rho], \tag{2.23}$$

where the first term is the interaction energy between the electrons and the external potential (for example, nuclei). The second term is the classical Coulomb interaction potential, where the factor $\frac{1}{2}$ prevents double counting of interactions. The third term $T_{\mathrm{s}}[\rho]$ is the kinetic energy of a system made of non-interacting electrons, and $E_{\mathrm{xc}}[\rho]$ is the exchange and correlation (XC) energy of an interacting system that also includes the difference between the kinetic energy of interacting electrons and a non-interacting one [29]. The last term accounts for important physical effects: the correction for self-interaction energy in the second term, where an electron interacts with itself, the Pauli exclusion principle, and the Coulomb correlation between individual electrons [30]. The form of the XC energy functional is unknown. However, by assuming that the density is slowly varying [28], the last term can be written as the following:

$$E_{\mathrm{xc}}[\rho] = \int \rho(\mathbf{r})\epsilon_{\mathrm{xc}}(\rho(\mathbf{r}))\mathrm{d}\mathbf{r}, \tag{2.24}$$

where $\epsilon_{\mathrm{xc}}(\rho(\mathbf{r}))$ is the exchange and correlation energy per electron in a homogeneous gas. Finally, they derived what are called the Kohn-Sham (KS) equations [28, 31]:

$$\Big[-\frac{1}{2}\nabla^2 + \Big[v(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\mathrm{d}\mathbf{r}' + \mu_{xc}(\rho(\mathbf{r}))\Big]\Big]\psi_i(\mathbf{r}) = \epsilon_i\psi(\mathbf{r}), \tag{2.25}$$

where the number of particles is conserved. The first term is the kinetic energy operator of the electrons, where ∇^2 is the Laplacian operator. The second is the external potential, the third is the repulsive Coulomb potential, and the last is the exchange-correlation contribution to the chemical potential of a uniform gas having density $\rho(\mathbf{r})$ [28].

Despite its success, DFT has some shortcomings. This ranges from incorrect estimates of spin-state energetics to significant errors in energy-barrier estimates for torsional rotation in certain cases [32]. However, a DFT drawback of concern here is the self-interaction energy, where the electron interacts with its own density [33]. Since 1981, many attempts to solve this problem have appeared in the literature [34]. Another important shortcoming is the failure to account for weak long-range interactions, known as dispersion effects [32]. Nonetheless, DFT is the workhorse nowadays in chemistry, biophysics, and materials science due to its sufficiently accurate results, despite the approximations in the exchange-correlation energy functional, and its speed, thanks to enhanced numerical algorithms on modern computers [21]. DFT has been used successfully in many biomolecular studies involving proton transfer as well [35, 36, 37].

2.1.3 Basis Sets

In Eq. (2.18), the orbital is expanded in terms of a basis set where the coefficients are determined by either the HF or DFT method to minimize the energy. The basis set is

composed of a set of atom-centered Gaussian functions, having the form [38]:

$$\chi_i = N(x-X)^k(y-Y)^l(z-Z)^m e^{-\zeta_i(r-R)^2}, \tag{2.26}$$

where N is the normalization constant. The (X, Y, Z) are the coordinates of the function center, which is usually the nucleus. The sum of $k+l+m$ is the angular momentum, and ζ_i is the exponent that determines the radial extension of this function. The accuracy of this expansion depends on the number of considered basis functions in the sum and the exponent value ζ_i [38]. χ_i is called a primitive function that is usually contracted into m fixed linear combinations:

$$\mathcal{K}_j = \sum_{i=1}^{m} d_{ij}\chi_i, \tag{2.27}$$

$$\psi = \sum_{j=1}^{k} c_j\mathcal{K}_j. \tag{2.28}$$

where d_{ij} is a fixed coefficient and c_j are determined by minimizing the energy. The reason for contraction is simply computational efficiency, as the number of coefficients included in the calculations is reduced to k. The basis functions are usually classified by the number of contracted functions, as this reflects the basis set's flexibility in capturing different molecular environments [38].

2.1.4 Exchange-Correlation Functionals

As mentioned in Section 2.1.2, the form of the XC functional is not known. However, approximations are suggested and categorized in the order of complexity as follows: (1) local density approximation (LDA) where XC functional depends only on the local density and not its spatial variation, (2) generalized-gradient approximation (GGA) where it depends on the derivative of the density, (3) meta-GGA, where there is dependence on the Laplacian of the density, (4) hybrid functional, that in addition incorporates a HF exchange functional [18].

One of the most commonly used XC functionals for modeling biological systems is B3LYP [39, 40, 41, 42], which has the following form:

$$E_{\mathrm{xc}} = a_0E_{\mathrm{x}}^{\mathrm{HF}} + (1-a_0)E_{\mathrm{x}}^{\mathrm{LDA}} + a_{\mathrm{x}}\Delta E_{\mathrm{x}}^{\mathrm{B88}} + a_{\mathrm{c}}E_{\mathrm{c}}^{\mathrm{LYP}} + (1-a_{\mathrm{c}})E_{\mathrm{c}}^{\mathrm{VWN}}, \tag{2.29}$$

where the constants are $a_0 = 0.2$, $a_{\mathrm{x}} = 0.72$, $a_{\mathrm{c}} = 0.81$. The first term is the exchange from HF, the second is the exchange correction from LDA, and the third term is the exchange correction from the gradient. The fourth is the Lee-Yang-Parr correlation functional, and the last term is the Vosko-Wilk-Nusair local correlation functional [42].

2.1.5 Dispersion Effects

As mentioned in Section 2.1.2, one important disadvantage of DFT is the absence of dispersion interactions. Dispersion forces are attractive interactions that arise from the response of electrons in a specific region to the charge density fluctuations in another region [43]. The standard XC functionals (LDA, GGA, and hybrid) do not consider dispersion forces, as there is no account for the instantaneous changes of density, and only consider local properties in the calculation of XC energy. A simple example of the failure of DFT to report the correct energy is the energy of the stacked configuration of DNA base pairs, as the major part of this interaction is dispersion forces [43].

To overcome this issue, one simple approach is to add an energy-correction term that reproduces the correct behavior of dispersion interaction in the gas phase. As the dispersion interactions fall with the distance as a $\frac{1}{r^6}$ function, the correction term takes the following form [43]:

$$E_{\text{dispersion}} = -\sum_{\text{A,B}} \frac{C_6^{\text{AB}}}{r_{\text{AB}}^6}. \tag{2.30}$$

C_6^{AB} is the dispersion coefficient of a specific A, B atom pair. In this thesis, the D3 method developed by Grimme [44] with Becke-Johnson correction [45] is used, where the dispersion coefficients are obtained from element-specific atomic reference data and continuously adjusted according to local coordination numbers [44].

2.2 Molecular Mechanics

The success of QM as a fundamental theory for treating atomic and molecular systems is hindered by the complexity of calculations and the high computational cost. Consequently, QM cannot be used to study large systems such as biomolecules. Therefore, an alternative framework characterized by high scalability and low computational cost must be introduced to attain the simulation of large systems. This approach is called molecular mechanics (MM), where the treatment is purely classical (Newtonian). The theoretical justification of approximating the dynamics using classical (non-QM) considerations is furnished by the Ehrenfest theorem, which is stated using the following equation [13]:

$$M_i \frac{d^2\langle \mathbf{R}_i \rangle}{dt^2} = -\langle \nabla_i \mathcal{E}_{\text{BO}}(\mathbf{R}_1, \ldots, \mathbf{R}_N) \rangle, \tag{2.31}$$

where M_i and $\mathbf{R}_i$ are the mass and the position of the i^{th} nucleus, respectively. The bracket is the average using the nuclear wavefunction ϕ_n:

$$\langle \cdots \rangle = \int d\mathbf{R}_1 \ldots \int d\mathbf{R}_N \, \phi_n^*(\mathbf{R}_1, \ldots, \mathbf{R}_N, t) \cdots \phi_n(\mathbf{R}_1, \ldots, \mathbf{R}_N, t). \tag{2.32}$$

Considering that nuclei are heavy and the temperatures are relatively high (∼310 K), the average in Eq. (2.31) can be rewritten as follows:

$$\langle \nabla_i \mathcal{E}_{\mathrm{BO}}(\mathbf{R}_1, \ldots, \mathbf{R}_N) \rangle \approx \nabla_i \mathcal{E}_{\mathrm{BO}}(\langle \mathbf{R}_1 \rangle, \ldots, \langle \mathbf{R}_N \rangle), \tag{2.33}$$

which allows Eq. (2.31) to be written as:

$$M_i \frac{d^2 \langle \mathbf{R}_i \rangle}{dt^2} = -\nabla_i \mathcal{E}_{\mathrm{BO}}(\langle \mathbf{R}_1 \rangle, \ldots, \langle \mathbf{R}_N \rangle). \tag{2.34}$$

The form of Eq. (2.34) resembles the form of Newton's second law, where the right-hand side is the force acting on the i^{th} nucleus. MM makes use of Eq. (2.34) to approximate $\mathcal{E}_{\mathrm{BO}}$ with a potential function made of simple and differentiable terms called the force field and denoted as $U(\mathbf{R}_1, \ldots, \mathbf{R}_N)$, where N is the number of nuclei [13].

The potential function $U(\mathbf{R}_1, \ldots, \mathbf{R}_N)$ in the MM scheme is called the force field, and it is classified into 3 classes. Class I force fields are made of two main terms, named bonded and non-bonded interaction energies. The functional form of the bonded energy term is as follows:

$$\begin{aligned} E_{\mathrm{bonded}} = \sum_{\text{bonds}} K_b (b - b_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\substack{\text{improper} \\ \text{dihedrals}}} K_\varphi (\varphi - \varphi_0)^2 \\ + \sum_{\text{dihedrals}} \sum_{n=1}^{6} K_{\phi,n} (1 + cos(n\phi - \delta_n)). \end{aligned} \tag{2.35}$$

The first two terms are bond stretching/compression and angle bending, which are modeled by harmonic potentials. For each one of them, there is an equilibrium value b_0 and θ_0 and a force constant K_b and K_θ, respectively. The third term is the improper dihedral, used to account for the energetics of out-of-plane motion and to maintain planarity of certain atom groups. It is also a harmonic function with equilibrium out-of-plane angle φ_0 and force constant K_φ [46]. The last term is the torsional energy, which accounts for the rotation of dihedral angles, and due to the periodicity of this rotation, it is modeled by a sum of cosine functions with different multiplicities n. The choice of n depends on the functional group. For example, ethane dihedral requires $n = 3$, and ethene requires $n = 2$. δ_n is the phase which is usually either 0° or 180° so that different enantiomers have the same energy [46].

The non-bonded energy term is defined as follows:

$$E_{\text{non-bonded}} = \sum_{\substack{\text{non-bonded} \\ \text{pairs } ij}} \frac{q_i q_j}{4\pi D r_{ij}} + \sum_{\substack{\text{non-bonded} \\ \text{pairs } ij}} \varepsilon_{ij} \left[\left(\frac{R_{\min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{\min,ij}}{r_{ij}} \right)^{6} \right]. \tag{2.36}$$

The first term is the Coulomb potential, which describes the electrostatic interaction between fixed point charges (partial charges) q_i and q_j, where r_{ij} is the separation

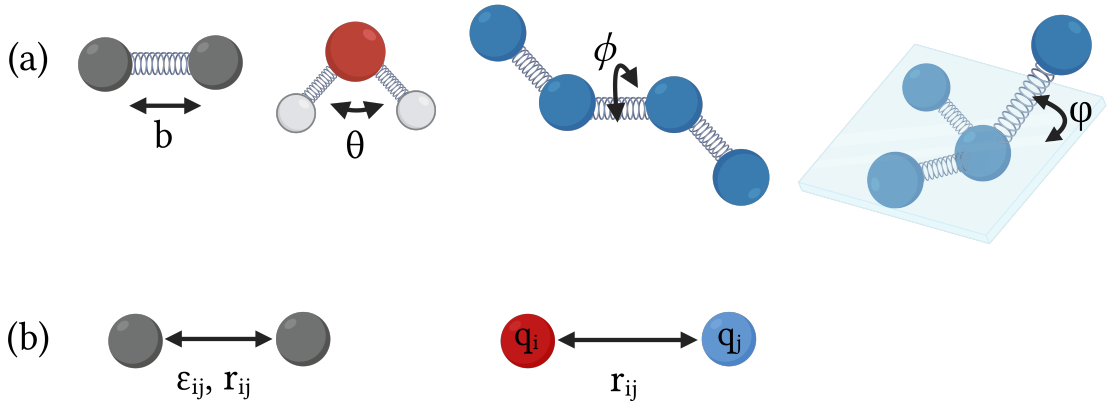


Figure 2.1: (a) Illustration of bonded interactions in the molecular mechanics scheme. From left: the covalent bond, the angle, the dihedral angle, and finally the improper dihedral angle. (b) Illustration of non-bonded interactions. From left: Lennard-Jones and electrostatic interactions.

distance. The total electrostatic energy of the system is simply the sum of individual interactions. D is the electric constant. The last term is called the Lennard-Jones (LJ) potential, which accounts for the van der Waals interaction with a cut off $R_{\text{min},ij}$, and a well-depth ε_{ij} [46]. The system's potential energy is the sum of two terms: $U = E_{\text{bonded}} + E_{\text{non-bonded}}$. An illustration of both bonded and non-bonded interactions between different atoms is shown in Fig. 2.1.

The long-range electrostatics pose a problem. The electrostatic potential energy scales as $O(N^2)$, thereby increasing the computational burden of the simulation beyond the capabilities of most modern computers. A simple cutoff, although computationally efficient, proved to cause problems [47]. This issue can be overcome with the Ewald summation method [48], which scales as $O(N^{3/2})$. Further improvement came with the particle mesh Ewald (PME) method [49], which uses fast Fourier transform (FFT) techniques to calculate the reciprocal space part of the Ewald summation method, and it achieves a scaling of $O(N \log N)$ [49, 50].

Notably, the class I force field considers only harmonic terms, neglecting higher-order terms. In class II and III, higher order (anharmonic) and cross terms are added to the potential energy [46]. Furthermore, the partial charges are fixed and cannot account for the polarization effect. There are different methods to overcome this problem. The first being simply the use of off-center charge placement, where partial charges are put around the nucleus as well. Other attempts, such as the fluctuating charge model, the Drude oscillator, and the induced dipole model, aim to incorporate polarization effects [51]. In this thesis, the fixed-charge CHARMM36 force field is used [52, 53, 54].

From the MM scheme, the form of the potential function $U(\mathbf{R}_1, \ldots, \mathbf{R}_N)$ is obtained, which can be used to study biomolecular systems of interest. However, this

requires constructing the model system, comprising the biomolecule and its environment, which includes water, ions, and possibly lipid membranes or carbohydrates. Furthermore, all biological systems are dynamic. To investigate these systems, the static model must be supplemented with dynamics that trace the time evolution of the biomolecule and its environment.

2.3 Molecular Dynamics

Using the potential function from the molecular mechanics scheme and computing the corresponding forces, the motion of a system of particles can be simulated. The forces are substituted into Newton's equations of motion, yielding the new atomic coordinates. This is referred to as an MD simulation [55]. The output is called a trajectory, which is essentially a set of atom positions as a function of time. This trajectory is used to answer questions about atomic-level details that experiments cannot address [56].

Newton's equation is a second-order differential equation:

$$-\nabla_{\mathbf{R}_i} U(\mathbf{R}) = M_i \ddot{\mathbf{R}}_i(t). \tag{2.37}$$

where $\nabla_{\mathbf{R}_i}$ is the gradient with respect to the position of the i^{th} atom, M_i is the mass, and $\ddot{\mathbf{R}}_i(t)$ is the acceleration of the i^{th} atom [55].

Several important aspects have to be taken into account before performing MD simulations. The first is periodic boundary conditions (PBC). PBC is a computationally efficient technique for accounting for bulk effects in reality and avoiding surface artifacts arising from the finite size of the simulated system. This means that molecules exiting one side appear on the other. In practice, this is realized via the minimum image convention, where any component of the distance between two atoms is upper-bounded by half of the cell length along that specific component [57].

The second is energy minimization. This step moves the system along the potential energy surface to locate the global energy minimum with respect to the atomic positions, thereby providing a reliable starting point for subsequent MD simulations. The potential energy surface could be determined using a force field (MM) or a BO energy surface (QM). There are various methods for this task, but the most commonly used are steepest descent and conjugate gradient [58]. Steepest descent is a first-order derivative method that uses the forces on the atoms to locate the nearest energy minimum [59].

2.3.1 Integrators

In MD simulations, the differential equations (see Section 2.3, Eq. (2.37)) are solved numerically. This is achieved by transforming the differential equations into finite-difference equations, which are then solved iteratively using an integrator algorithm [55].

The main requirements for an integrator are: (1) reproduce the original differential equation when the step size approaches zero, (2) be stable and accurate for longer time intervals, (3) be time-reversible, (4) be computationally efficient, (5) be symplectic, meaning that it must preserve the phase space volume and conserve energy [55]. There are many integrators, each with its own advantages and disadvantages [55]. Examples of commonly used integrators include the leapfrog and Velocity Verlet algorithms [60].

2.3.2 Temperature & Pressure Control

Physiological processes in biological systems take place at constant temperature and pressure. Furthermore, biochemical experiments are usually performed at constant pressure [61]. Therefore, in MD simulations of biological systems, it is desirable to maintain constant temperature and pressure.

For temperature control, atomic velocities are altered; for pressure control, atomic coordinates are rescaled. To attain a constant temperature, the system is coupled to an external heat bath. The details of the heat bath and its thermal interaction with the system should not affect the system's equilibrium properties. There are different ways to achieve coupling, namely, constraints, extended systems, and stochastic methods. The latter achieves coupling by modeling the thermal fluctuations as stochastic dynamics [62]. The Langevin thermostat [63] is a popular stochastic thermostat that produces the correct NVT ensemble (for more details on ensembles see Section 2.5.1), i.e., the atomic trajectories sample the canonical distribution [64].

To maintain constant pressure during an MD simulation, methods must be employed to adjust particle motion. These methods are called barostats. There are various barostats, such as Berendsen [65], Parrinello-Rahman [66], and Langevin piston [67]. When both temperature and pressure are held constant, the simulation samples the isothermal-isobaric (NPT) ensemble. Barostats must minimally disturb Newtonian dynamics and sample the NPT ensemble [68].

MD simulations are much more efficient and scalable relative to QM calculations, but still suffer from the shortcomings of the MM force field. Namely, charge polarization and bond formation/breakage are missing. Therefore, to study PT in enzymes, both the accuracy of QM and the scalability of MD are required.

2.4 Hybrid QM/MM Method

If studying chemical reactions or any change in electronic structure, such as enzyme reaction mechanisms, electron transfer, electronic transitions, or the spin states of a metalloprotein's active site, is of interest, the system should be studied quantum-

mechanically. However, DFT-based QM approaches are expensive and limited to a few hundred atoms. This limitation is particularly problematic for the computational study of biomolecules with thousands of atoms, making a full QM treatment impractical. To address this problem while maintaining an accurate description of the biological environment, the hybrid quantum mechanics/molecular mechanics (QM/MM) method combines classical and quantum approaches within a single computational framework. The region where chemistry occurs is treated with QM, and the rest of the system is treated with MM [69]. The QM region must be large enough to avoid QM/MM boundary artifacts [70].

In QM/MM simulations, there are three types of interactions: interactions within the MM region, interactions within the QM region, and interactions between atoms of the two regions. The latter ones are the most challenging to describe. There are two main approaches to modeling the interaction between the two regions: the subtractive and additive schemes. The subtractive scheme is described as follows [71]:

$$V_{\mathrm{QM/MM}} = V_{\mathrm{MM}}(\mathrm{MM{+}QM}) + V_{\mathrm{QM}}(\mathrm{QM}) - V_{\mathrm{MM}}(\mathrm{QM}), \tag{2.38}$$

where the first term represents the entire system (MM+QM) modeled with molecular mechanics. The second term corresponds to the QM region energy, and the last term is the QM region modeled solely by molecular mechanics. This scheme is relatively easy to implement as there is no information exchange between the QM and MM routines, but it comes with some caveats: (1) the QM region is treated in an isolated way, which means that the charges in the MM region does not affect the charge distribution in the QM region, (2) the force field of the MM region must be capable of handling the chemical changes in the QM region during the reaction process [71].

The additive scheme, on the other hand, is as follows:

$$V_{\mathrm{QM/MM}} = V_{\mathrm{MM}}(\mathrm{MM}) + V_{\mathrm{QM}}(\mathrm{QM}) + V_{\mathrm{QM\text{-}MM}}(\mathrm{QM{+}MM}), \tag{2.39}$$

where the first term is the MM region modeled by molecular mechanics, and the second term is the QM region modeled by quantum mechanics. The last term explicitly represents the interaction between the two regions.

In both approaches, bonded and van der Waals (Lennard-Jones) interactions are treated similarly to those in the MM scheme, and the same force field parameters are used in QM/MM simulations. For electrostatic interactions, more sophisticated considerations are needed. Mechanical embedding is where the electronic wavefunction is calculated for an isolated QM region, meaning the charges in the QM region are not polarized due to the partial charges in the MM region. The next level of sophistication is electrostatic embedding, in which the MM point charges are incorporated into the QM Hamiltonian, and the electronic wavefunction calculation accounts for polarization

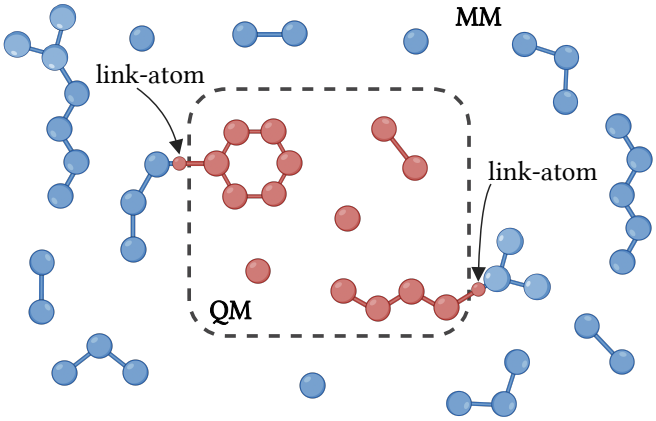


Figure 2.2: A schematic representation of QM and MM regions in QM/MM simulations. A bond crossing the QM region has to be capped by a link atom.

effects. In polarization embedding, the MM atoms are polarized, which necessitates the use of a polarizable force field [71]. In this thesis, additive electrostatic embedding is used as implemented in NAMD/ORCA [72, 73].

Furthermore, in QM/MM simulations, special attention is given to covalent bonds that cross the boundary between the two regions, as cutting a covalent bond will create an unpaired electron [71]. However, the choice of covalent bonds is usually restricted, and one must avoid cutting through polar bonds and instead separate the MM and QM regions through nonpolar C−C bonds [74]. To overcome this problem, an atom (usually hydrogen) is put at a suitable position along the bond, called the "link atom". Furthermore, the charge of an atom in the MM region, located close to a link atom, must be rescaled (or completely ignored) to prevent strong artificial repulsion [72]. Fig. 2.2 illustrates region partitioning and link atoms in QM/MM simulations.

The hybrid QM/MM method is a powerful tool for investigating PT. However, it often fails to capture slow PT reactions within the simulation timescales. Hence, to study the kinetics and energetics of a specific PT pathway, enhanced sampling techniques and free energy calculation methods are employed.

2.5 Free Energy Calculations

One objective of MD simulations is to calculate expectation values of different quantities. This can be a feasible task for some mechanical quantities, such as pressure or internal energy, but a difficult task for some statistical quantities, such as entropy and free energy, as these quantities cannot be expressed as ensemble averages [10]. In MD simulations, both the potential energy of the particles and thermal fluctuations (entropy) are present. Free energy accounts for both, making it a useful quantity to calculate. Free energy (F) is an interplay between the potential energy (internal energy, U) and the entropy

(S), and is given by the following formula:

$$F = U - TS, \tag{2.40}$$

where T is the absolute temperature.

The quantity inquired in practice is the difference between the free energies of two states at specific conditions. The two states can be bound or unbound ligand to a receptor, mixed or unmixed fluids, or structured or unstructured biomolecules [75]. If two states, A and B, can be distinguished via an experimental method, the ratio of probabilities of finding the system in each state (P_A and P_B) is calculated from the ratio of times spent in each state. The free energy difference between the two states, A and B is [76]:

$$\Delta F_{A,B} = -\beta^{-1} \ln \frac{P_A}{P_B}, \tag{2.41}$$

where $\beta = \frac{1}{k_B T}$ with k_B being the Boltzmann constant.

On the other hand, the main quantity sought from simulations is the configurational integral or the configurational partition function:

$$Z = \int \exp\left(-\beta U(\mathbf{x})\right) \mathrm{d}\mathbf{x}, \tag{2.42}$$

where $U(\mathbf{x})$ is the potential energy of the system. Many quantities can be calculated from Z, such as heat capacity and free energy. Once free energy is known, all thermodynamic quantities can be retrieved from the knowledge of the free energy and its derivative [77]. Nonetheless, calculating Z from MD simulations is extremely difficult.

Free energy calculations are challenging as it is strenuous to sample the less stable state as much as the more stable one [75]. Therefore, enhanced sampling methods have been developed to address this problem. To describe these methods, certain fundamental concepts of statistical physics must be reviewed.

2.5.1 Phase Space, Ensemble, & Ergodicity

Phase space is the set of all $(\mathbf{q}, \mathbf{p})$ points where $\mathbf{q}$ are the generalized coordinates and $\mathbf{p}$ are the generalized momenta. Configuration space is simply the set of all the generalized coordinates $\mathbf{q}$. Throughout this text, $(\mathbf{x}, \mathbf{p})$ is the phase space if the coordinates are Cartesian, and consequently, the corresponding configuration space is $\mathbf{x}$. Furthermore, a microstate is defined as a point in configuration space. A macrostate is a collection of microstates (a region in configuration space) where a probability can be assigned to each microstate contained within [76].

The concept of an ensemble stems from the idea that many different microstates can relate to the same macrostate. Rigorously speaking, an ensemble is "a collection of

systems described by the same set of microscopic interactions and sharing a common set of macroscopic properties" [78]. Examples of ensembles used in MD simulations are: (1) NVE, where the number of particles, volume, and energy are constant, (2) NVT, where the number of particles, volume, and temperature are constant, (3) NPT, where the number of particles, pressure, and temperature are constant.

An important assumption in MD simulations is ergodicity. An ergodic dynamical system is one in which ensemble averages can be obtained from long-time averages, i.e., the samples are taken from an infinitely long trajectory. Mathematically, ergodicity means that the ensemble average of quantity O can be calculated in the following way:

$$\langle O \rangle = \lim_{M \to \infty} \frac{1}{M} \sum_{t=1}^{M} O(\mathbf{x}_t, \mathbf{p}_t), \tag{2.43}$$

where $(\mathbf{x}_t, \mathbf{p}_t)$ is a discrete dynamics that is ergodic [76]. Of course, M never approaches infinity as the MD trajectories are finite, and therefore, ergodicity is an assumption.

2.5.2 Free Energy

The free energy in the canonical ensemble (NVT) is called Helmholtz free energy, and it is defined as follows:

$$F = -\beta^{-1} \ln Z_\Sigma = -\beta^{-1} \ln \int_\Sigma e^{-\beta U(\mathbf{x})} \, \mathrm{d}\mathbf{x}, \tag{2.44}$$

where Σ is a subset of the configuration space corresponding to a macrostate [76].

If the ensemble of choice is isothermal-isobaric (NPT), the Gibbs free energy is the corresponding quantity defined by the following formula [78]:

$$G = -\beta^{-1} \ln \Delta = -\beta^{-1} \ln \frac{1}{V_0} \int_0^\infty e^{-\beta P V} Z_\Sigma \, \mathrm{d}V, \tag{2.45}$$

where Δ is the isothermal-isobaric partition function and V_0 is the reference volume.

In simulations, both F and G are often functions of what are called collective variables (CV). The reasons for using a CV are as follows: (1) it defines a unique pathway along the free energy landscape, which is straightforward to sample and easy to visualize, (2) easier comparison with the experimental results, as experiments provide a macroscopic view, and the models are designed at the molecular level [79].

2.5.3 Collective Variables

Collective variables (CV) are a lower-dimensional representation defined by differentiable functions of 3N coordinates [76, 79]:

$$\mathbf{s} = \boldsymbol{\eta}(\mathbf{x}) = \boldsymbol{\eta}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N), \tag{2.46}$$

where **s** denotes a CV vector. In practice, a single CV, s, is considered as a function of many fewer arguments, $\eta(z^{(1)}(\mathbf{x}), z^{(2)}(\mathbf{x}), \ldots, z^{(n)}(\mathbf{x}))$. z is called a basis function and can be a simple function, such as distance or angle, or a more complicated one like the principal component of the covariance matrix [79].

To utilize CVs in simulations, an existing library is coupled to the simulation package. However, if the CV intended for use is not already present in the library, it must be introduced as a script that the simulation package reads. This comes with the caveat that not only the functional form of the CV needs to be implemented, but also the Jacobian matrix must be evaluated. The reason is that gradients are required for bias-force calculations [79]. Limited timescales in MD simulations hinder sampling of less favorable states; therefore, simulations are biased to ensure proper sampling. The simulation is said to be biased if an extra energy term is added to the potential energy function. The resulting potential function can be written as follows:

$$\tilde{U}(\mathbf{x}) = U(\mathbf{x}) + U^b(\mathbf{x}), \tag{2.47}$$

where $U^b(\mathbf{x})$ is the bias potential. For a bias potential as a function of a CV $U^b(\eta(\mathbf{x}))$, the biasing force is calculated as follows [76]:

$$\mathbf{F}^b = \nabla_{\mathbf{x}} U^b(\eta(\mathbf{x})) = -\frac{\mathrm{d}U^b(\eta(\mathbf{x}))}{\mathrm{d}\eta(\mathbf{x})} \nabla_{\mathbf{x}} \eta(\mathbf{x}), \tag{2.48}$$

where $\nabla_{\mathbf{x}}\eta(\mathbf{x})$ is the Jacobian matrix. In this thesis, the *Colvars* module [79] coupled to NAMD simulation software [80] is used. This module requires that the Jacobian matrix be supplemented in row-major form. For a function of the form $f_i(x_j)$, the row-major Jacobian is: $\{\nabla_x f_1, \nabla_x f_2, \ldots\}$.

CVs are related to the reaction coordinate concept. RC is a coordinate that describes the pathway of a process, such as a chemical reaction, a conformational change, or a phase change. It can capture the reactant and product states separated by a transition state. If both the minima and the transition state are distinguished along the CV, it is a good candidate to serve as a RC [78]. An illustration of the concept of CV and its role in reducing the dimensionality of free energy is shown in Fig. 2.3.

In PT studies, a simple and commonly used CV is the linear combination of hydrogen-bonding distances (LinComb):

$$d = \sum_i r_i - \sum_j r_j, \tag{2.49}$$

where the first term is the sum over all the bond lengths in the reactant state, and the second term is the sum over all non-bonded distances in the reactant state. d works well for linear chains with two or three intervening waters, but its use is less robust for longer non-linear pathways [3].

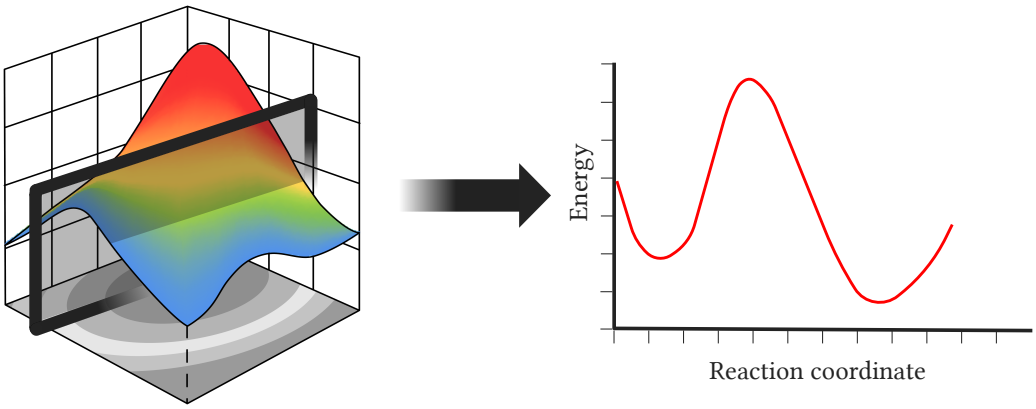


Figure 2.3: An illustration of how CV reduces the dimensionality of hypersurfaces by acting as a plane cutting through the high-dimensional free energy surface.

To account for charge migration and water reorientation in PT, another CV developed by Pomes and Roux [81] used the projected total dipole moment of the protonated water wire. This CV corresponds to the projection of the center of excess charge on the water wire axis. This CV is global, as it depends on the configuration of all water molecules in the water wire, making it sensitive to water orientation and susceptible to fluctuations [3]. Chakrabarti *et al.* [82] used a CV based on the number of protons coordinated to the oxygen atom with bonds switched off rather than an abrupt cutoff. This CV cannot distinguish between three hydrogen atoms bonded to an oxygen or coming close to it due to the collision of water molecules [3].

To overcome the previous shortcomings, König *et al.* [3] proposed the modified center of excess charge (mCEC), which has the following form:

$$\boldsymbol{\xi} = \sum_{i=1}^{N_H} \boldsymbol{r}^{H_i} - \sum_{j=1}^{N_X} w^{X_j} \boldsymbol{r}^{X_j} - \sum_{i=1}^{N_H} \sum_{j=1}^{N_X} f_{\mathrm{sw}}(d_{X_j,H_i}) \cdot (\boldsymbol{r}^{H_i} - \boldsymbol{r}^{X_j}). \tag{2.50}$$

The first term is a sum over all the positions of the hydrogen atoms involved ($\boldsymbol{r}^{H_i}$), the second term is the sum over all the heavy atoms X involved ($\boldsymbol{r}^{X_j}$), with each having a weight w^{X_j}, representing the minimum number of hydrogen atoms coordinated to the heavy atom X_j during PT. The third term is a sum of position differences between all the hydrogen and heavy atom pairs, representing the contribution from individual bonds. $f_{\mathrm{sw}}(d_{X_j,H_i})$ is a switching function defined as follows:

$$f_{\mathrm{sw}}(d) = \frac{1}{1 + \exp\left[(d - r_{\mathrm{sw}})/d_{\mathrm{sw}}\right]}, \tag{2.51}$$

where r_{sw} is a parameter representing the distance at which the bond is at its half-strength. d_{sw} parameter represents how quickly the bond strength decays to 0 (i.e., influences the slope). An illustration of the switching function can be seen in Fig. 2.4.

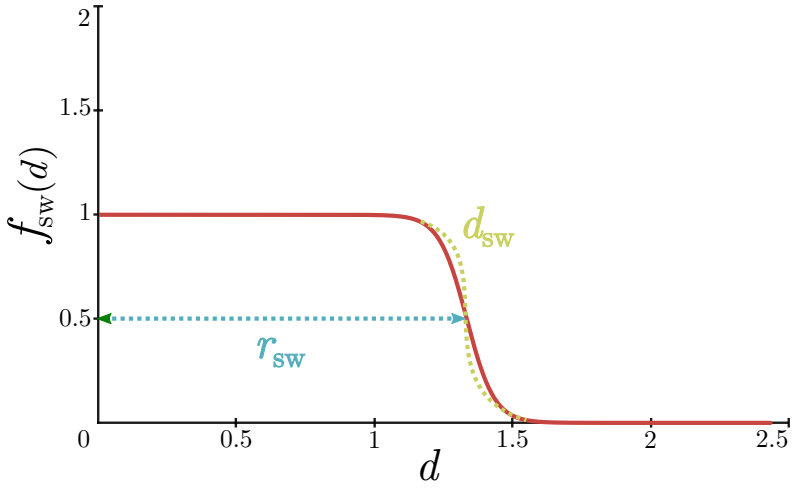


Figure 2.4: Plot of switching function for the bonds, along its parameters, used in mCEC CV.

This CV cannot account for simultaneous protonation/deprotonation of a single amino acid residue. In this scenario, one heavy atom accepts a proton while the other donates a proton. The two heavy atoms are said to be paired. That is why a fourth term is added to mCEC as follows:

$$\begin{aligned}\boldsymbol{\xi} = &\sum_{i=1}^{N_H} \boldsymbol{r}^{H_i} - \sum_{j=1}^{N_X} w^{X_j} \boldsymbol{r}^{X_j} - \sum_{i=1}^{N_H}\sum_{j=1}^{N_X} f_{\mathrm{sw}}(d_{X_j,H_i}) \cdot (\boldsymbol{r}^{H_i} - \boldsymbol{r}^{X_j}) \\ &+ \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2} [m(X_\alpha, \{H\}) \cdot (\boldsymbol{r}^{\beta} - \boldsymbol{r}^{\alpha}) + m(X_\beta, \{H\}) \cdot (\boldsymbol{r}^{\alpha} - \boldsymbol{r}^{\beta})],\end{aligned} \tag{2.52}$$

where the last term is the sum over all the pairs of paired atoms. $w_{\text{pair}}^{\alpha,\beta}$ is the minimum number of hydrogen atoms coordinated to the residue containing the α, β atom pair during the PT. $m(X, \{H\})$ is a function that encodes whether at least one proton is coordinated to the corresponding heavy atom or not. $m(X, \{H\})$ is defined as:

$$m(X, \{H\}) = \frac{\sum_i \left(f_{\mathrm{sw}}(d_{X,H_i})\right)^{n+1}}{\sum_i \left(f_{\mathrm{sw}}(d_{X,H_i})\right)^{n}}, \tag{2.53}$$

where n is a reasonably large number [3]. Both CVs (d and $\boldsymbol{\xi}$) are shown in Fig. 2.5.

One obstacle remains: vectors cannot be used in FES calculations, and a scalar ξ must be defined that follows the mCEC vector from the donor to the acceptor. The functional form of this scalar is not unique, with different suggestions in the literature [3]. In this thesis, the functional form described in the work of Kim *et al.* [83] is used:

$$\xi = \boldsymbol{\xi} \cdot \frac{(\boldsymbol{r}_a - \boldsymbol{r}_d)}{||\boldsymbol{r}_a - \boldsymbol{r}_d||}, \tag{2.54}$$

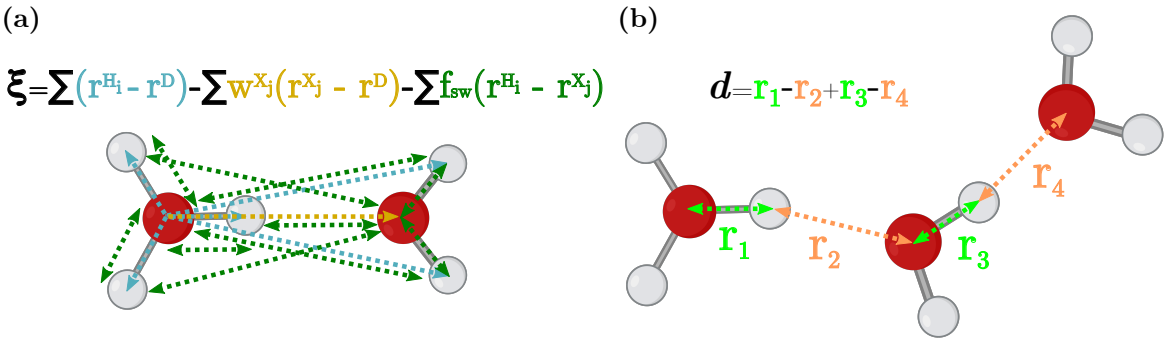


Figure 2.5: (a) Representation of the mCEC vector. The fourth term is neglected because, in this case, no simultaneous proton acceptance/donation occurs. (b) Representation of LinComb CV.

where the mCEC vector is projected on the normalized vector pointing from the donor to the acceptor. The mCEC itself is also slightly modified by moving the center of the coordinate system to the donor atom position $\mathbf{r}_d$:

$$\boldsymbol{\xi} = \sum_{i=1}^{N_H}(\boldsymbol{r}^{H_i} - \mathbf{r}_d) - \sum_{j=1}^{N_X} w^{X_j}(\boldsymbol{r}^{X_j} - \boldsymbol{r}_d) - \sum_{i=1}^{N_H}\sum_{j=1}^{N_X} f_{\mathrm{sw}}(d_{X_j,H_i}) \cdot (\boldsymbol{r}^{H_i} - \boldsymbol{r}^{X_j}) + \sum_{\text{pairs}} \frac{w^{\alpha,\beta}_{\text{pair}}}{2}[m(X_\alpha, \{H\}) \cdot (\boldsymbol{r}^\beta - \boldsymbol{r}^\alpha) + m(X_\beta, \{H\}) \cdot (\boldsymbol{r}^\alpha - \boldsymbol{r}^\beta)]. \tag{2.55}$$

Now that the concept of CV has been discussed in detail, it is important to note its relation to free energy. The free energy as a function of a collective variable $\mathbf{s}$ is called the free energy surface (FES) and denoted as $A(\mathbf{s})$. In many applications, the FES along a path connecting two states is the quantity of interest to determine reaction kinetics [75]. The FES, $A(\mathbf{s})$, is a function of the partially integrated configurational partition function:

$$A(\mathbf{s}) = -\beta^{-1} \ln \int \delta(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x}))\nu(\mathbf{x})\, \mathrm{d}\mathbf{x} = -\beta^{-1} \ln \rho(\mathbf{s}), \tag{2.56}$$

where $\rho(\mathbf{s})$ is called the marginal probability density, $\delta(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x}))$ is a multivariate Dirac delta function [76]. $\nu(\mathbf{x})$ is the configurational distribution and defined as follows:

$$\nu(\mathbf{x}) = \frac{e^{-\beta U(\mathbf{x})}}{\int e^{-\beta U(\mathbf{x})} \mathrm{d}\mathbf{x}}. \tag{2.57}$$

The FES is also referred to as the potential of mean force (PMF) [76], and throughout this thesis, both terms are used interchangeably. Moreover, as noted in Section 2.5, calculating FES is challenging due to undersampling in high-energy regions. Therefore, enhanced sampling techniques must be employed to achieve sufficient sampling of states.

2.6 Umbrella Sampling

Umbrella sampling is an enhanced sampling technique in which the CV is not constrained (fixed) but is harmonically restrained by adding an energy term that ensures sufficient and uniform sampling along a RC. Sampling is performed in multiple simulations, called windows, with overlapping CV distributions [84]. The unbiased distribution is written as follows [84]:

$$\rho^u(\mathbf{s}) = \frac{\int \delta(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x})) \exp\left(-\beta U(\mathbf{x})\right) \mathrm{d}\mathbf{x}}{\int \exp\left(-\beta U(\mathbf{x})\right) \mathrm{d}\mathbf{x}}, \tag{2.58}$$

where $\rho^u(\mathbf{s})$ is the probability distribution in an unbiased simulation as a function of the CV, $\delta(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x}))$ is a multivariate Dirac delta function. $U(\mathbf{x})$ is the potential energy of the system. The bias potential to keep the CV close to the reference value in a window is usually harmonic, having the following simple form:

$$U_i^b(\mathbf{s}) = \frac{K}{2}(\mathbf{s} - \mathbf{s}_i^0)^2, \tag{2.59}$$

where K is the force constant and $\mathbf{s}_i^0$ is the reference value in the i^{th} window. Since the simulation is biased, the distribution obtained from the simulation is biased. The relationship between the unbiased distribution and the biased one is as follows:

$$\rho_i^u(\mathbf{s}) = \rho_i^b(\mathbf{s}) \exp\left(\beta U_i^b(\boldsymbol{\eta}(\mathbf{x}))\right) \left\langle \exp\left(-\beta U_i^b(\boldsymbol{\eta}(\mathbf{x}))\right)\right\rangle, \tag{2.60}$$

where the i subscript denotes the i^{th} window and $\rho_i^b(\mathbf{s})$ is the biased distribution. Substituting this expression into Eq. (2.56), gives:

$$A_i(\mathbf{s}) = -\beta^{-1} \ln \rho_i^b(\mathbf{s}) - U_i^b(\mathbf{s}) - \beta^{-1} \ln \left\langle \exp\left(-\beta U_i^b(\mathbf{s})\right)\right\rangle. \tag{2.61}$$

The $\rho_i^b(\mathbf{s})$ in the first term can be approximated by a normalized histogram obtained from the MD simulation, given the assumption that the simulation is ergodic (see Section 2.5.1). The second term is the biasing potential. The last term is crucial when combining windows to form the full FES, $A(\mathbf{s})$ [84].

2.6.1 Weighted Histogram Analysis Method

The weighted histogram analysis method (WHAM) [85] is a way to estimate the last term mentioned in Eq. (2.61), where in this section it is denoted as F_i to simplify the notation. The global unbiased distribution $\rho^u(\mathbf{s})$ can be calculated as a weighted sum of the unbiased distribution in each window $\rho_i^u(\mathbf{s})$:

$$\rho^u(\mathbf{s}) = \sum_i^k \left(\frac{a_i}{\sum_j a_j}\right) \rho_i^u(\mathbf{s}), \tag{2.62}$$

where the denominator is for normalization and k is the number of windows. a_i is defined as follows:

$$a_i(\mathbf{s}) = N_i \exp\left(-\beta\, U_i^b(\mathbf{s}) + \beta\, F_i\right), \tag{2.63}$$

where N_i is the number of steps sampled in the i^{th} window. F_i is determined by the following equation:

$$\exp\left(-\beta\, F_i\right) = \int \rho^u(\mathbf{s}) \exp\left(-\beta\, U_i^b(\mathbf{s})\right)\, \mathrm{d}\mathbf{s}. \tag{2.64}$$

The unbiased distribution enters Eq. (2.64) and F_i enters Eq. (2.63) and consequently Eq. (2.62). Therefore, both must be solved iteratively until convergence [84].

Although the US method has been successfully used in many studies [36, 37, 86], its high computational cost can be a serious limitation when resources are limited. Consequently, other enhanced sampling methods, such as metadynamics, are preferred over US due to low computational cost.

2.7 Metadynamics

The idea behind metadynamics (MetaD) is to construct a time-dependent bias potential using repulsive Gaussian functions deposited at the position of the CV [87]. These Gaussian functions fill the energy minima, thus flattening the underlying FES [78]. To motivate this idea rigorously, consider the marginal probability distribution in Eq. (2.56). $\rho(\mathbf{s})$ can be thought of as an ensemble average of the delta function:

$$\rho(\mathbf{s}) = \left\langle \delta(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x})) \right\rangle. \tag{2.65}$$

By invoking the ergodic hypothesis, Eq. (2.65) can be written as a time average:

$$\rho(\mathbf{s}) = \lim_{t\to\infty} \frac{1}{t} \int_0^t \delta(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x}(t)))\, dt. \tag{2.66}$$

Next, the δ function is rewritten as the limiting case of a Gaussian function with its width approaching 0:

$$\rho(\mathbf{s}) = \lim_{t\to\infty} \lim_{\sigma\to 0} \frac{1}{\sigma\sqrt{2\pi t}} \int_0^t \exp \frac{(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x}(t)))^2}{2\sigma^2}\, dt, \tag{2.67}$$

where σ is the standard deviation. Next, switching into the discrete sum over a specific time step ΔT and assuming $\eta(\mathbf{x}(0)) = s$, yields:

$$\rho(\mathbf{s}) = \lim_{t\to\infty} \lim_{\sigma\to 0} \frac{1}{\sigma\sqrt{2\pi t}} \left[1 + \sum_{k=1}^{N-1} \exp \frac{(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x}(k\Delta t)))^2}{2\sigma^2}\right], \tag{2.68}$$

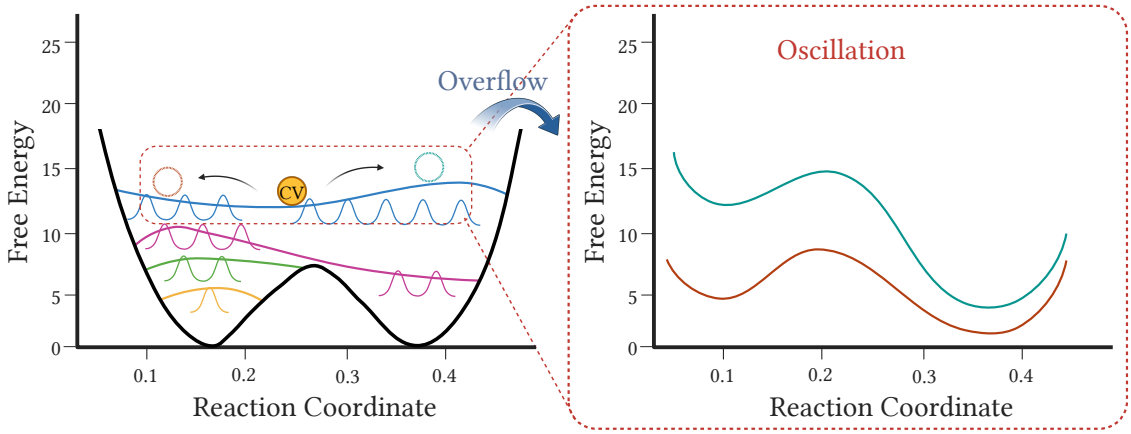


Figure 2.6: A schematic representation of deposited Gaussian functions filling the free energy minima. The oscillatory behavior of the resulting PMF and the risk of overflow are also shown.

where N is the number of time steps taken. Plugging the expression of ρ in Eq. (2.68) into the FES in Eq. (2.56), and using small amplitude Gaussian functions, $\ln(1+x) \approx x$, gives:

$$A(\mathbf{s}) = -k_B T \sum_{k=1}^{N-1} \exp \frac{(\mathbf{s} - \boldsymbol{\eta}(\mathbf{x}(k\Delta t)))^2}{2\sigma^2} + \text{const.} \tag{2.69}$$

This derivation can serve as a conceptual basis for constructing a bias potential as a sum of Gaussian functions [78]. An illustration of Gaussian functions filling minima in the free energy surface is shown in Fig. 2.6. Motivated by the above derivation, a bias potential of the following form is written:

$$U^b(\mathbf{s}, t) = \sum_{k=1}^{t/\tau} w_G \exp\left(\sum_{\alpha=1}^{N_{\text{CV}}} \frac{(s_\alpha - \eta_\alpha(\mathbf{x}(k\tau)))^2}{2\sigma_\alpha^2} \right), \tag{2.70}$$

where $\mathbf{s}$ is now written in component form (i.e., sum over number of CVs, N_{CV}). The sum is over all deposited Gaussian functions, with w_G the Gaussian height (or weight) and τ the deposition rate [87]. Next, the biased marginal probability distribution in the canonical ensemble is considered [88]:

$$\rho^b(\mathbf{s}) = \frac{\exp\left(-\beta(A(\mathbf{s}) + U^b(\mathbf{s}))\right)}{\int \exp\left(-\beta(A(\mathbf{s}) + U^b(\mathbf{s}))\right) d\mathbf{s}}. \tag{2.71}$$

Eq. (2.71) shows that as the bias potential approaches the negative of $A(\mathbf{s})$, the biased distribution ρ^b becomes a constant (i.e., a uniform distribution). This fact can serve as a criterion for stopping the biased simulation. However, in MetaD simulations, the PMF oscillates and does not converge. Therefore, it is not trivial to know when to stop the biased simulation [12].

From a practical standpoint, there are three parameters whose numerical values must be determined at the beginning of the simulation. The first one is the height of

the Gaussian functions w_G, where the numerical value is recommended to be less than k_BT [87]. This quantity at 310 K is about 0.616 kcal/mol. The second parameter is the Gaussian width 2σ. For this parameter, the CV trajectory is extracted from the unbiased run, and a normalized histogram is generated. Next, a Gaussian function must be fitted to the distribution. After that, the width of the deposited Gaussian functions is taken as a fraction ($\frac{1}{2}$ or $\frac{1}{3}$) of the fitted Gaussian distribution. The last parameter is the deposition rate of the Gaussian functions τ. The value of τ is chosen such that the error is minimized and the filling time is appropriate enough. The system should be allowed to relax between depositions; therefore, plotting the autocorrelation function in an unbiased simulation can help determine the numerical value of τ [87]. The autocorrelation function (AC) is given by the following formula:

$$\mathrm{AC}(k) = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}, \tag{2.72}$$

where n is the number of samples, k is the number of steps the data has shifted (lag), $\bar{x}$ is the sample mean, and x_t is the sample value at time t.

MetaD has the advantage of pushing the system over the lowest energy barrier, allowing the exploration of new reaction pathways. It explores the low-energy regions first, unlike umbrella sampling, which forces the system in a specific direction. Nevertheless, the method has shortcomings. For instance, the constructed bias potential oscillates around the FES due to continuous deposition of the Gaussian functions, overfilling the underlying FES and pushing the system to high-energy regions [12]. This is illustrated in Fig. 2.6. To overcome this problem, well-tempered metadynamics is developed.

2.7.1 Well-Tempered Metadynamics

In well-tempered metadynamics (WTMetaD), the bias potential enhances the sampling in a controlled manner by taking the bias potential as follows:

$$U^b(\mathbf{s}) = -\left(1 - \frac{1}{\gamma}\right) A(\mathbf{s}), \tag{2.73}$$

where γ is a positive coefficient defined as $\frac{T+\Delta T}{T}$, with T being the absolute temperature and ΔT is the bias temperature [87]. Considering the limit, $\gamma \to 1$, the unbiased simulation is recovered [88]. In this case, the unbiased marginal probability distribution is related to the biased one through the following form:

$$\rho^b(\mathbf{s}) = \frac{\rho^u(\mathbf{s})^{\frac{1}{\gamma}}}{\int \rho^u(\mathbf{s})^{\frac{1}{\gamma}}\, d\mathbf{s}}. \tag{2.74}$$

Eq. (2.74) shows that as γ increases, the distribution becomes wider and the fluctuation of CV increases. The numerical value of γ is an input parameter in the simulation and should be chosen such that the fluctuations are strong enough to overcome the energy barrier but small enough so that the calculations are not very demanding [88]. The bias potential in this case takes the following form:

$$U^b(\mathbf{s},t) = \sum_{k=1}^{t/\tau} w_G \exp\left(\sum_{\alpha=1}^{N_{\mathrm{CV}}} \frac{(s_\alpha - \eta_\alpha(\mathbf{x}(k\tau)))^2}{2\sigma_\alpha^2}\right) \exp\left[-\frac{1}{\gamma-1}\beta U^b_{k-1}(\mathbf{s}_k)\right]. \tag{2.75}$$

The second exponential is a scaling factor that can be shown to decrease by a factor of $\frac{1}{t/\tau}$ [88]. This means that, with each iteration, the bias potential grows less. It can be shown that the time evolution of the bias potential in Eq. (2.75) is given by the following differential equation [88]:

$$\frac{\mathrm{d}U^b(\mathbf{s},t)}{\mathrm{d}t} = \int G(\mathbf{s},\mathbf{s}') \exp\left[-\frac{1}{\gamma-1}\beta U^b_{k-1}(\mathbf{s}',t)\right] \frac{\exp\left(-\beta(A(\mathbf{s})+U^b(\mathbf{s},t))\right)}{\int \exp\left(-\beta(A(\mathbf{s})+U^b(\mathbf{s},t))\right)\,\mathrm{d}\mathbf{s}}\,\mathrm{d}\mathbf{s}', \tag{2.76}$$

where the deposited Gaussian functions are represented as $G(\mathbf{s},\mathbf{s}')$. The differential equation in Eq. (2.76) has an asymptotic solution as follows:

$$U^b(\mathbf{s},t) = -\left(1-\frac{1}{\gamma}\right)A(\mathbf{s}) + \frac{1}{\beta}\log\frac{\int \exp\left(-\beta A(\mathbf{s})\right)\mathrm{d}\mathbf{s}}{\int \exp\left(-\beta(A(\mathbf{s})+U^b(\mathbf{s},t))\right)\,\mathrm{d}\mathbf{s}}. \tag{2.77}$$

The last term is independent of $\mathbf{s}$ and therefore shifts the FES by a constant [88]. WTMetaD has the advantage of bias potential convergence. Therefore, the constructed PMF does not oscillate around the underlying FES.

The optimal choice of parameters for metadynamics in hybrid QM/MM simulation is unknown. Therefore, different parameters in both conventional and well-tempered variants of metadynamics are tested. Furthermore, the choice of a RC is not unique, and different reaction coordinates are investigated. The mCEC CV will be implemented and tested, and the results will be compared with those from the LinComb CV.

3. Biochemical Background

3.1 Biological Building Blocks

Proteins, lipids, nucleic acids, and carbohydrates are the building blocks of life. Proteins are made of 20 different amino acids (AA), which are carboxylic acids with an amino group in the α position. Amino acids are chiral molecules, usually found in the L-form in living organisms. At pH 7, AA are zwitterions possessing both positive and negative charges. Amino acids are linked by peptide bonds to form polypeptides. Proteins are polypeptides that perform biological functions. The AA sequence in protein is called the primary structure. Secondary structures are folding patterns stabilized by hydrogen bonds, such as the α-helix and β-strand. Tertiary structure represents the three-dimensional organization of the secondary structures stabilized by hydrogen bonds, salt bridges, hydrophobic interactions, and disulfide bonds. Quaternary structure describes the coupling of polypeptides together to form a single functional protein [89].

3.2 Respiratory Complex I

Respiratory complex I (CI) is of great interest to the bioenergetics community, as it is the major player in creating the proton motive force across the inner membrane of the mitochondria [90]. CI is made of a peripheral arm and a membrane domain (see Fig. 3.1). The peripheral arm consists of two modules: the N module, where NADH is oxidized, and the Q module, where quinone is reduced. The membrane part has a proximal and a distal PT module [91]. CI functions by coupling the oxidation of nicotinamide adenine dinucleotide hydride (NADH) and the subsequent reduction of ubiquinone (Q) to proton pumping across the inner mitochondrial membrane. Electrons are donated from NADH to flavin mononucleotide and are passed along a chain made of seven iron-sulfur clusters to reach ubiquinone. By reduction of ubiquinone, ubiquinol is formed, and subsequently, the released energy drives the proton pumping across the inner membrane [90]. CI is made of 14 core subunits, where 7 of them make up the hydrophilic domain encoded in the nuclear genome, and the other 7 make up the

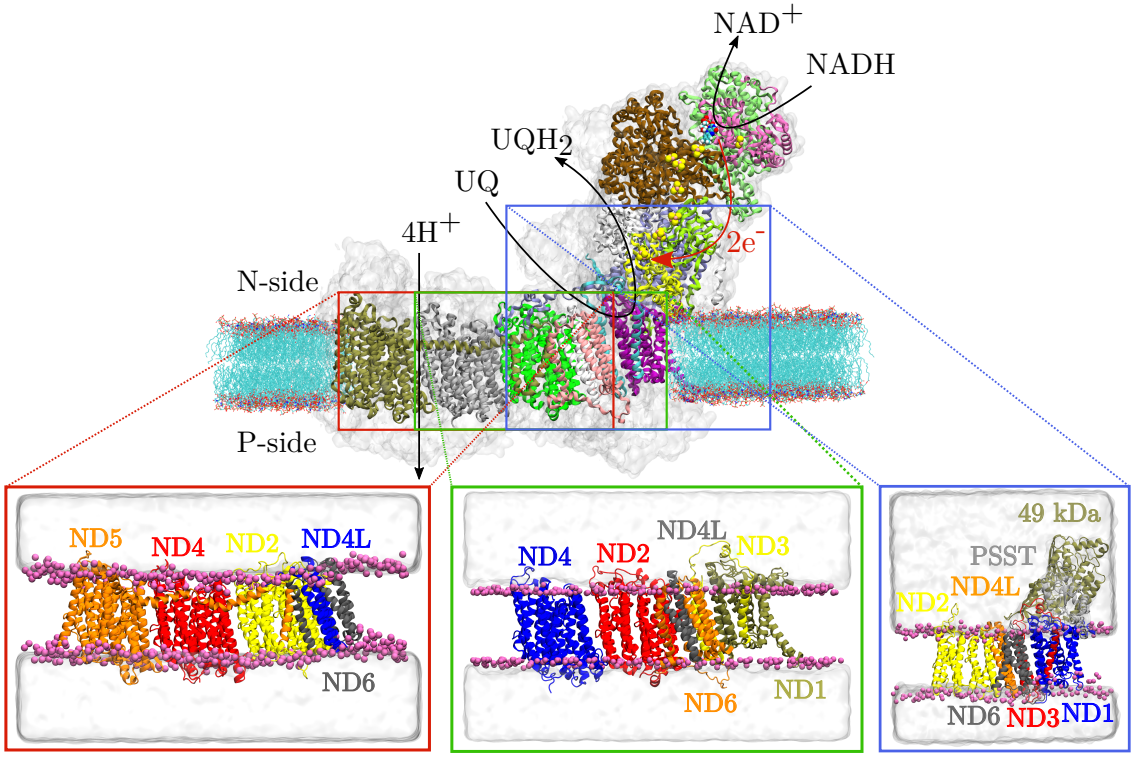


Figure 3.1: Top inset: A schematic representation of respiratory complex I. Bottom inset: a close-up representation of the simulated systems in this thesis work. From left to right: for PT pathways in ND5 subunit, for PT pathways in ND2 subunit, for PT pathways in E-channel.

hydrophobic domain embedded in the inner mitochondrial membrane and encoded by the mitochondrial genome. Mitochondrial complexes have a variety of supernumerary subunits that are not necessary for the catalysis but might have other roles [92].

There are many unanswered questions regarding how the redox energy is coupled to proton pumping across the membrane [92]. It is accepted that antiporter-like subunits (ND2, ND4, ND5) transport protons across the membrane [86] with some molecular details of this transport function still missing. Therefore, conducting bioenergetic studies of different PT pathways in CI is necessary to develop a clear mechanistic understanding of its function.

In this thesis, different subunits of CI are used as model systems for free energy calculations of PT reactions. The work focuses on PT reactions occurring in the ND2 and ND5 subunits and in a highly conserved region called the E-channel. The E-channel consists mainly of glutamic acid and is hydrated in many high-resolution structures. It connects the quinone tunnel to the membrane domain of CI, and it is hypothesized to play a key role in CI function by coupling the redox reactions to proton pumping [36]. The QM/MM setup involving the ND2 subunit has tyrosine and lysine residues, and the ND5 setup contains basic residues, including a histidine. E-channel setups, on the other hand, involve acidic residues.

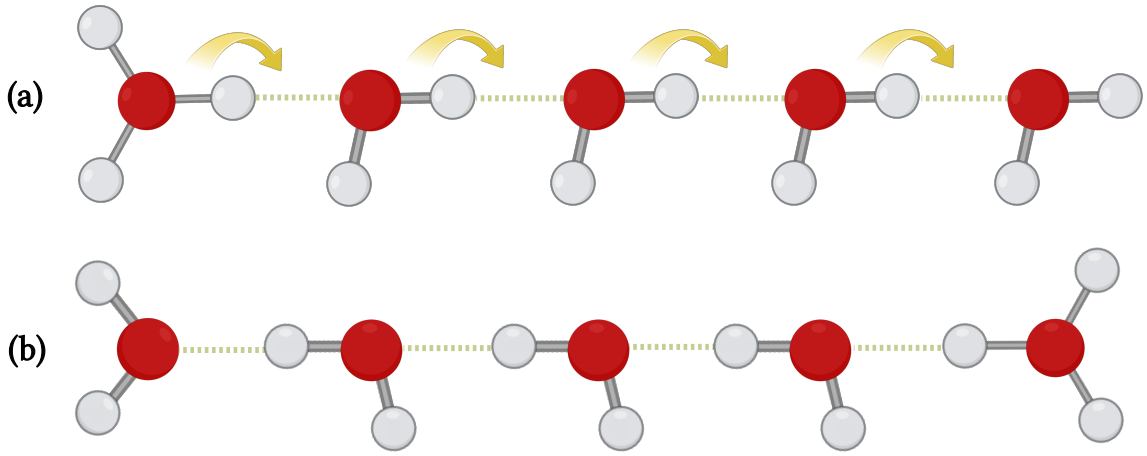


Figure 3.2: A schematic illustration of PT along a water wire. (a) initial state, (b) final state.

3.3 Proton Transfer

To understand PT, it must be recognized that excess protons in bulk water exist as the Eigen cation ($H_9O_4^+$), in which the hydronium ion (H_3O^+) is hydrogen-bonded to three water molecules, known as the first hydration shell. Another form is the Zundel cation, $H_5O_2^+$, where a proton is shared between two water molecules [93]. The first hydration shell of the Eigen cation is characterized by strong hydrogen bonds. However, in the second shell, where proton transfer is initiated, a normal hydrogen bond is broken, thereby forming the Zundel cation. Further fluctuations and reorientation of water molecules lead to complete proton migration to the neighboring water molecule, leading to the formation of the Eigen cation again [93]. Therefore, there is no single proton hopping, but many protons are involved to shuttle the excess proton [1]. An illustration of PT along a water wire is shown in Fig. 3.2.

In proteins, in addition to charged AA side chains, ions are present. Ions in the proximity of a hydrogen bond can facilitate proton transfer by transferring excess protonic charge from one group to another. This behavior is due to electrostatic effects, in which positive charges repel one another, causing the proton to reside in another group with a lower pK. The reverse pulling effect by opposite charges is also valid. Furthermore, small conformational changes in proteins can modify the angular characteristics of a hydrogen bond (i.e., the dipole moment vector), thereby altering the protonation states of residues. This results in PT favoring the proton when the proton affinity of one group is higher than that of the other, simply by reorienting the geometry of the two groups involved [2]. Overall, the environment has a substantial effect on PT in proteins, making the study of PT in this context complex.

4. Results

In this thesis, various QM/MM metadynamics parameters are tested, and the optimal choice is applied to different PT pathways with varying amino acid compositions to validate the parameters. Furthermore, two different reaction coordinates are investigated: one is the linear combination of hydrogen-bonding distances (LinComb), and the other is the modified center of excess charge (mCEC), which is implemented from scratch and tested on various PT pathways with different amino acid compositions. A flowchart of the thesis work is shown in Fig. 4.1.

4.1 Simulation Protocol

Simulation setups in this thesis followed the same protocol: model building and classical equilibration are accomplished in different studies [36, 37, 86, 94], and comprise the initial steps before any QM/MM simulation. The output data is used to perform QM/MM energy minimization for 300 steps and an unbiased hybrid QM/MM dynamics simulation for 1 ps. The final frame from the unbiased QM/MM simulation is used to initiate a biased QM/MM simulation and obtain the PMF along a specific PT pathway. The exceptions are the PT pathways in E-channel, where QM/MM energy minimization and the unbiased QM/MM simulation are already carried out in the work of Simsive *et al.* [36], and only the final frame is taken from that project to launch a biased QM/MM simulation. In all simulations, the Verlet integration algorithm [60] with a 1 fs timestep is used. Non-bonded interactions are calculated using the Verlet [95] cutoff scheme with a 12 Å cutoff distance, 10 Å switching, and 14 Å pairlist distance. In addition, a constant temperature of 310 K is maintained by the Langevin thermostat [63]. In the NPT simulations, the Langevin piston barostat [67] is used with an oscillation period of 100 fs and a decay constant of 50 fs to maintain the pressure at 1.01325 atm. All QM/MM simulations are performed in the NVT ensemble, except for the E-channel simulations, which are in the NPT ensemble. QM region is treated with DFT, with the hybrid B3LYP XC functional [39, 40, 41, 42] and def2-SVP basis set [96]. Furthermore, to account for the long-range dispersion effects, DFT-D3 dispersion correction [44] is employed. The SCF energy tolerance is set to 10×10^{-8} au. The electrostatic interaction between the

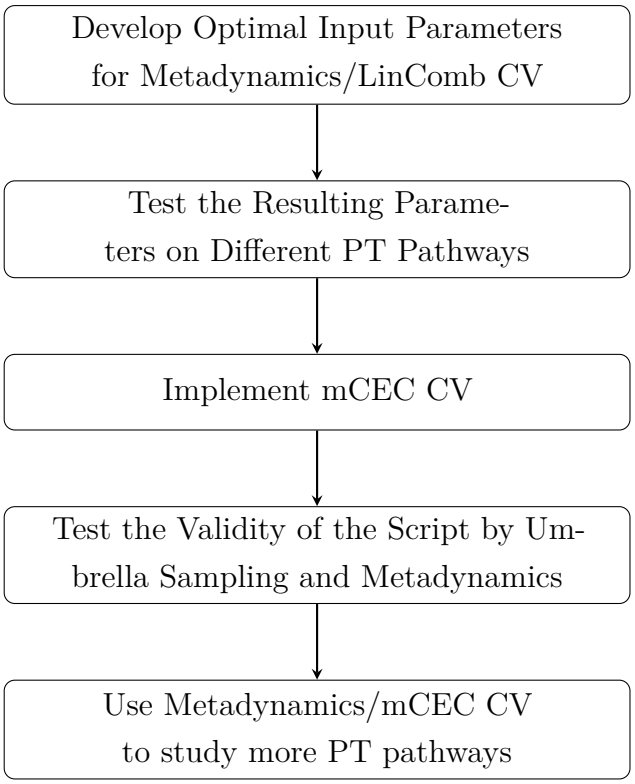


Figure 4.1: Workflow of the results in this thesis.

QM and MM regions is treated using the additive electrostatic embedding [72]. For the MM region, the CHARMM36 force field [52] is used. The QM/MM simulation is performed by coupling the NAMD software [80], which is responsible for the MM part of the simulation, with the ORCA software [73], which is responsible for the QM simulation. The *Colvars* module [79] is used to define the CV in all simulations.

4.2 Testing Metadynamics Parameters

The first goal of this thesis is to identify the optimal input parameters for MetaD to study PT within the QM/MM framework. This necessitates comparing the results against an existing PMF profile generated in prior work [37, 86]. This facilitates comparison of results across different parameter sets, ensuring that results are qualitatively matching and are quantitatively similar.

A PT pathway in the ND2 subunit of CI (from Lys211 (K211) to Tyr225 (Y225), see Fig. 4.2) is chosen. This pathway is investigated by Djurabekova *et al.* [86], using the QM/MM umbrella sampling enhanced sampling method. The PMF profile (see Fig. 4.2) is obtained using the US/WHAM [97] and the errors are calculated using the bootstrapping method [98]. Analyzing the PMF profile reveals three points. First, three minima are seen in the PMF profile, which is an important qualitative feature. Next,

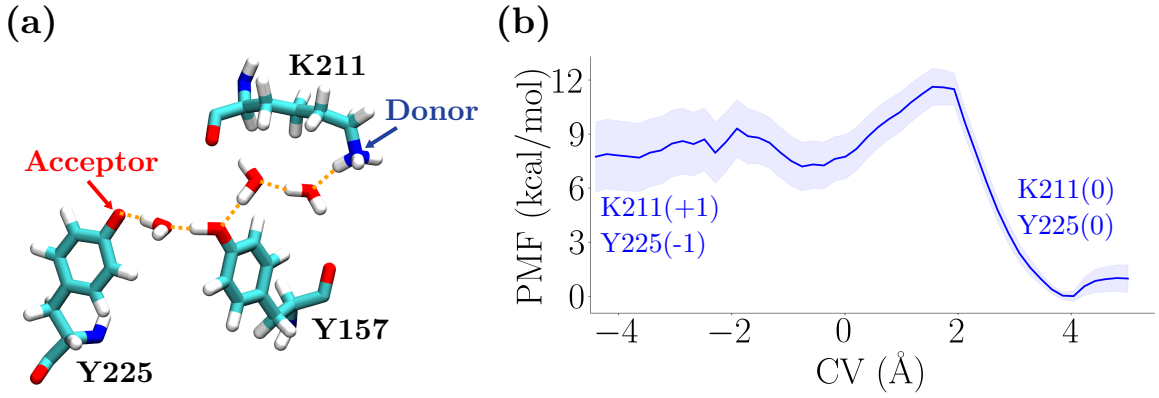


Figure 4.2: (a) PT pathway between K211 and Y225. The dotted orange line represents the selected PT pathway. (b) The corresponding PMF profile obtained by US/WHAM [86]. The shaded area is the standard error calculated by the bootstrapping method.

the barrier heights for the first and second minima are about 1.0 and 3.5 kcal/mol, respectively. Finally, the energy difference between the reactant and product states is about 8.0 kcal/mol.

The parameters have three hill weights (Gaussian height) of 0.1, 0.3, and 0.5 kcal/mol, adhering to the recommendation that the height should be less than 1 k_BT as mentioned in Section 2.7. The *Colvars* module refers to the Gaussian functions as hills, and this terminology will be used here as well.

To determine the hill width, a histogram of the CV trajectory is made, and a Gaussian function is fitted to the histogram. The results are shown in Fig. 4.3. The width of the fitted Gaussian function is about 0.73 Å, and the deposited hills are chosen to be 0.3 Å wide. However, smaller widths of 0.2 and 0.1 Å are also tested. From a practical standpoint, a grid width is defined that determines the number of discrete states of a CV, and the width of the deposited hills is a multiple of this width. This number has to be chosen such that it does not exceed the standard deviation of the CV trajectory in an unbiased run. The standard deviation of the selected CV is about 0.4 Å, and therefore a grid width of 0.2 Å is chosen. The finer the grid, the more computationally expensive the simulation, and the coarser the grid, the fewer details captured. Only in the case of 0.1 Å hill width, the grid width is chosen to be 0.1 Å according to the recommendations from the *Colvars* module [99] that the multiplier of the grid width (which determines the hill width) should be between 1.0 and 3.0.

Next, the deposition rate is determined by plotting the autocorrelation function (see Section 2.7) of the CV trajectory from the unbiased run and identifying the time at which it crosses 0. The plot is shown in Fig. 4.3. At a lag time of 219 fs, the function is 0, so a longer time must be chosen. However, the autocorrelation approaches 0 at a lag time of approximately 100 fs. This shows that the optimal deposition rate is

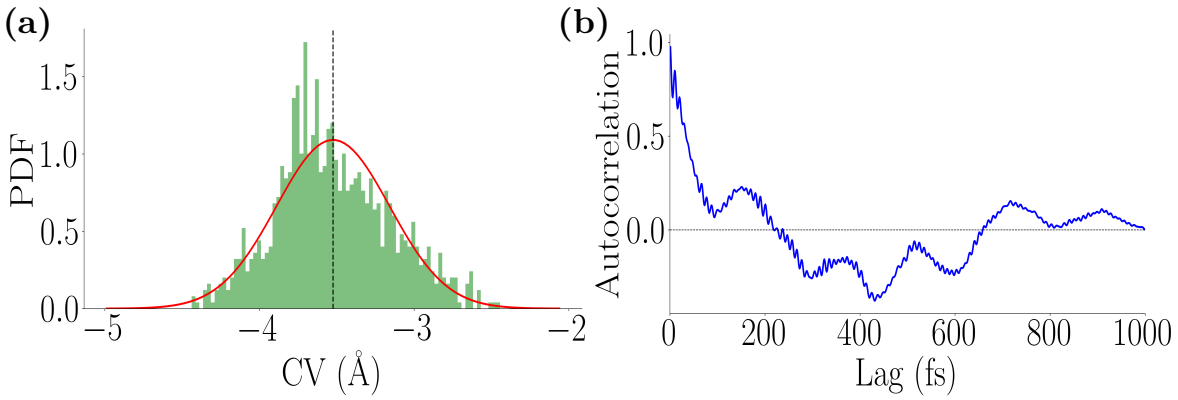


Figure 4.3: (a) Normalized histogram of the LinComb CV for PT from K211 to Y226. The fitted Gaussian function is plotted in red with a width of 0.73 Å and a mean of -3.52 Å represented by the black dashed line. (b) Autocorrelation function of the LinComb CV values along the unbiased trajectory for the PT from K211 to Y226. The function crosses 0 at lag time 219 fs.

system-dependent and smaller values can still be valid. In any case, high deposition rates are not feasible for QM/MM simulations due to computational cost, underscoring that the parameter choice recommendations in Section 2.7 are not suitable here. Two deposition rates, 10 and 50 fs, are tested, representing the optimal and the slowest possible choices, respectively. The tested parameters are listed in Table 4.1. The constructed system is shown in the bottom, middle panel of Fig. 3.1. The results for each test run are shown in Appendix A.

To properly estimate the errors of PMF calculations, the block-averaging technique [100] must be applied to the uncorrelated data points in the collected PMF dataset after the CV has become diffusive and the filling phase is complete (biased simulation has converged). Unfortunately, the MetaD method in QM/MM simulations is limited to short time intervals relative to classical MD simulations. In this case, sampling the reaction coordinate many times to obtain sufficient data points (as recommended in [87]) is often infeasible. Therefore, for conventional metadynamics, the PMF profile is constructed by collecting all the output PMF profiles within a specific range that starts after the completion of the filling phase and ends after the CV has diffused back once, and calculating the average and the standard deviation using the *NumPy* module [101] in Python. For the well-tempered variant of metadynamics (see Section 2.7), determining PMF convergence within the QM/MM framework is also challenging [102]. Therefore, the converged PMF is extracted as the final result.

4.2.1 Conventional Metadynamics

A comparison of all the results is shown in Fig. 4.4. First, comparing conventional 1, 3, and 5, as these differ only in the hill weight (Fig. 4.4 a, see also Table 4.1). Conventional

Table 4.1: Test simulations with their corresponding parameters set. (R) stands for replica.

Test Simulation	Hill Weight (kcal/mol)	Hill Width (Å)	Deposition rate (fs)	Simulation Time (ps)	ΔT (K)
Conventional 1	0.1	0.3	50	46.54	-
Conventional 2	0.1	0.3	10	24.62	-
Conventional 3	0.3	0.3	50	29.03	-
Conventional 4	0.3	0.3	10	8.73	-
Conventional 5	0.5	0.3	50	28.52	-
Conventional 6	0.5	0.3	10	8.74	-
Conventional 7	0.3	0.2	50	23.26	-
Conventional 8	0.3	0.1	50	16.70	-
Well-Tempered 1	0.1	0.3	50	26.95	1860
Well-Tempered 2	0.1	0.3	10	18.85	1860
Well-Tempered 3	0.3	0.3	10	14.98	1860
Well-Tempered 3 (R)	0.3	0.3	10	13.85	1860

5 with the largest hill has the largest standard deviation and differs the most from the US PMF (Fig. 4.2). Conventional 1 and 3 are quite similar, both reporting a barrier of about 3.0 kcal/mol and an energy difference between the reactant and product minima of about 7.5-8.5 kcal/mol. Importantly, conventional 3 produced the result in a shorter time, 17.5 ps faster (see Table 4.1). Next, conventional 3, 7, and 8 are compared as these only differ by the hill width, having the same 0.3 kcal/mol hill height (Fig. 4.4 b, see also Table 4.1). The results show that as the width decreases, the PMF profile becomes less smooth and the barriers diminish (the PMF profile becomes flatter). Smaller hill widths yield stronger biasing forces, causing the minimum to be undersampled and the barriers to be underestimated. In this case, although conventional 7 and 8 are faster, conventional 3 yields the best accuracy. Now comparing conventional 2, 4, and 6 as the only difference is the hill height, while having a 10 fs deposition rate (Fig. 4.4 c, see also Table 4.1). At this deposition rate, conventional 6 exhibits a very large standard deviation. Conventional 2 and 4 yield similar results for standard deviation and barrier size. Conventional 4 is consequently superior in this case, as it yields results faster (see Table 4.1). Finally, comparing conventional 3 and 4, as these are the best so far, and only differ in deposition rate (Fig. 4.4 d, see also Table 4.1). The results are similar, and conventional 3 has a smaller standard deviation. However, conventional 4 consumes fewer computational resources, making this parameter set favorable. So far, the conventional 4 parameters are the most accurate and fastest.

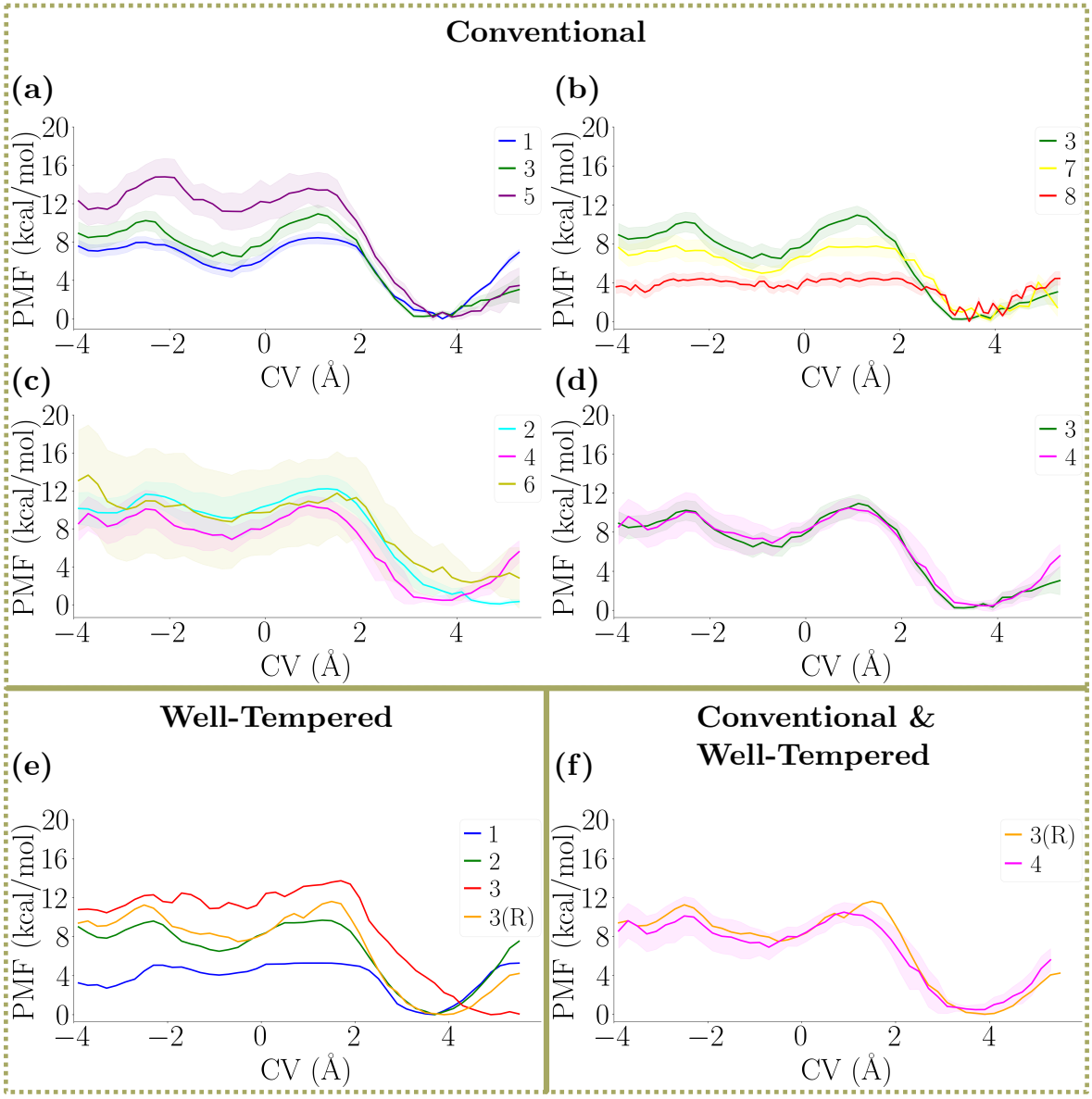


Figure 4.4: Comparison of all the test runs results. The PMF profile is obtained by averaging the PMF, and the shaded area represents the standard deviation. (a) conventional 1 (blue), 3 (green), and 5 (purple), (b) Comparison of the PMF generated from 3 test runs: conventional 3 (green), 7 (yellow), and 8 (red), (c) conventional 2 (cyan), 4 (purple), and 6 (olive), (d) conventional 3 (green) and 4 (purple), (e) well-tempered 1 (blue), 2 (green), 3 (red), and a replica of 3 (3(R), orange), (f) comparison of well-tempered 3 replica (orange) and conventional 4 (magenta). The setup parameters are listed in Table 4.1.

4.2.2 Well-Tempered Metadynamics

The next step is to test parameters using WTMetaD, an attractive choice for conducting PMF investigations due to its convergence properties (see Section 2.7). Three sets of parameters are also tested with WTMetaD. The results are shown in Fig. 4.4 e,f. Based on the results, the PMF profile closest to the US one (see Fig. 4.2) is obtained from the well-tempered 2 and 3(R) parameter sets, with 3(R) being faster (see Table 4.1). In well-tempered 3, where the parameter set is the same as conventional 4 (besides bias temperature, of course), the minimum has shifted, and the underlying PMF profile is quite different. The reason is that a water molecule between Y225 and Y157 (see Fig. 4.2) has rotated, and the PT pathway is reformed using the other O−H bond of the water that is not part of the CV. This led to a substantially different PMF profile, which otherwise should have been similar to well-tempered 2, as the only difference is the hill height, which should not change the results to this degree. This illustrates the shortcomings of the LinComb CV, which can be problematic in highly dynamic PT pathways. As a result, another replica of the well-tempered 3 parameters is relaunched from the same coordinates, with the only difference being the seed number (Fig. 4.4 e, orange line). In this case, the results are comparable both quantitatively and qualitatively to those of well-tempered 2 and conventional 4, with the added benefit of being less computationally expensive than well-tempered 2 and exhibiting better convergence properties than conventional 4.

Since the optimal parameters for a WTMetaD simulation to study PT have been identified, another system is chosen to test their validity. This time, two PT pathways in the ND5 subunit are analyzed (see Fig. 3.1, bottom panel, left system), where the majority of the amino acids involved are basic residues. These PT pathways are part of the Endres *et al.* study [94], and the convergence of PMF profiles is shown in Appendix B. The same parameters as well-tempered 3 (Table 4.1) are chosen, with the difference being the higher ΔT of 4808 K. The PT pathways and the corresponding PMF profiles are shown in Fig. 4.5. For the K305 to K342 pathway, the trajectory is 30 ps long, and the barrier is approximately 12 kcal/mol, with PT being favorable. The product state is about 4 kcal/mol lower in energy. For PT between K229 and H254, the simulation is 20 ps long, and the energy barrier is 12 kcal/mol, but the product state is about 4 kcal/mol higher in energy.

So far, the optimal parameters for biased hybrid QM/MM simulations with both the conventional and well-tempered metadynamics have been identified. Furthermore, the WTMetaD parameters have been tested in other PT pathways. These parameters enable accurate and efficient investigation of PT energetics. However, the results depend on the choice of a reaction coordinate. In the following section, a new RC is investigated.

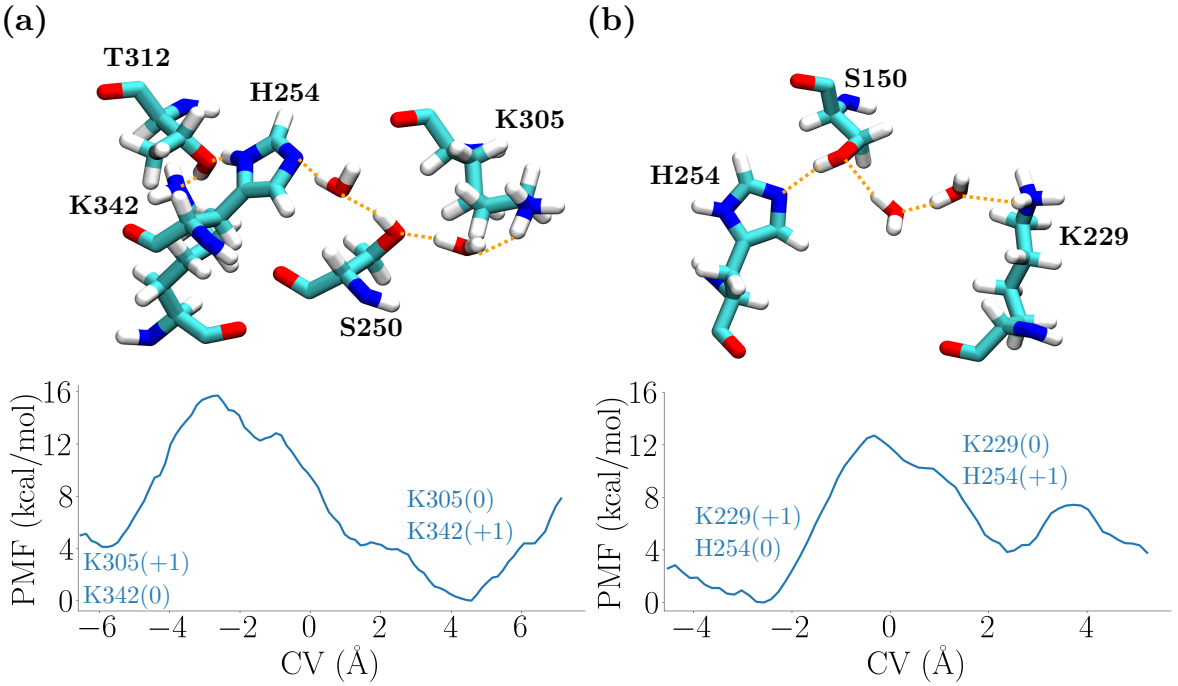


Figure 4.5: (a) PT pathway between K305 and K342 in the ND5 subunit. (b) PT pathway between K229 and H254 in the ND5 subunit. Orange dotted lines represent the selected PT pathway.

4.3 Modified Center of Excess Charge

The LinComb CV suffers from two shortcomings: (1) the PT pathway has to be strictly decided before biasing the RC, which can pose a challenge in highly hydrated regions where there could be multiple possible pathways for the PT, (2) during biasing, the PT pathway could be unstable, and certain water molecules might undergo rotations where the PT pathway is reformed by the other O−H covalent bond of the water that is not selected initially. To circumvent these difficulties, mCEC CV (see Section 2.5.3) is used to study long-range PT pathways, where these issues are more recurring. However, as this RC is not implemented in the current release of the *Colvars* module (2025-04-30) [79], it has been implemented in the framework of this thesis. The first step is to derive the analytical expression for the derivative of mCEC CV with respect to (x, y, z), which is given in Appendix C. The next step is to implement mCEC CV in a TCL script and test it to ensure it functions as expected. The TCL script can be found in https://github.com/TorabiM98/Center-of-Excess-Charge-CV.

To test the mCEC CV, the same PT pathway in the ND2 subunit of CI is chosen (see Section 4.2), and two scenarios are considered. In the first, the same water molecules as in Section 4.2 are selected, and in the second, three extra water molecules in the proximity are also added to the CV (see Fig. 4.6, top panel). For each, two bias simulations are performed, one using US and the other using WTMetaD.

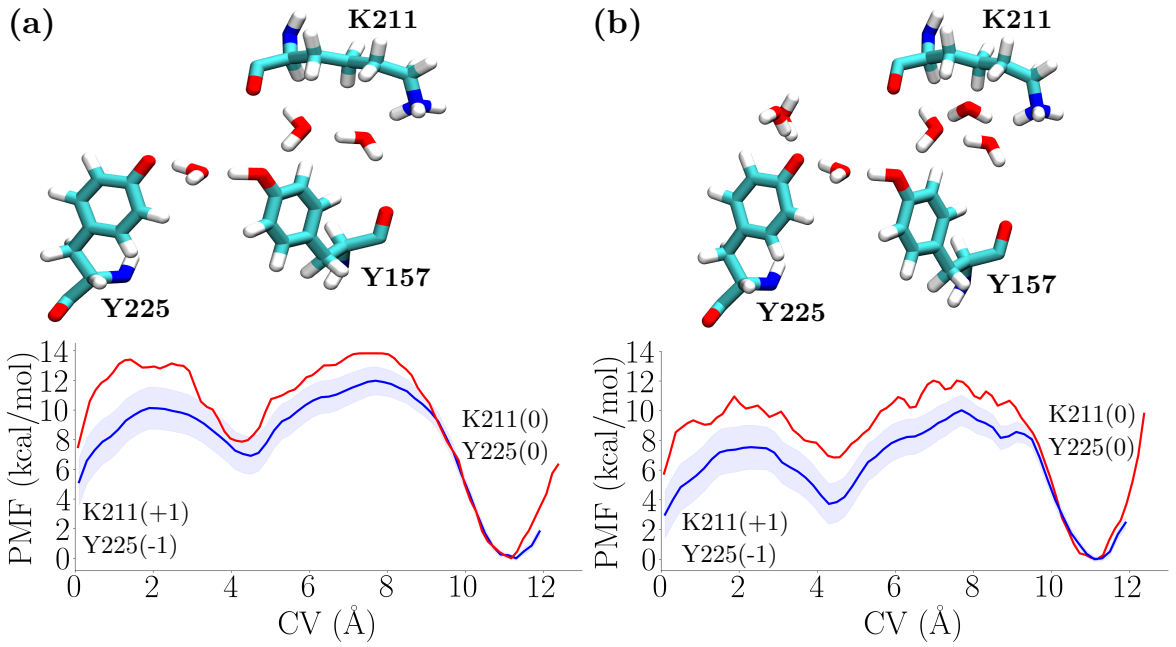


Figure 4.6: PMF profile of the PT between K211 and Y225 using mCEC as a reaction coordinate with (a) only water molecules making the pathway are included in the reaction coordinate, (b) surrounding water molecules from the QM region are added to the reaction coordinate. Blue curves represent profiles obtained using the US/WHAM technique, with the shaded area representing the standard error estimated by bootstrapping. Red curves are obtained by WTMetaD.

For the WTMetaD case, the same parameters as in well-tempered 3 are used (see Table 4.1) with a trajectory length of 15 ps. In US simulations, 60 windows are used. A harmonic potential with a force constant of 50 kcal/mol/Å^2 is employed in each window to restrain the CV near the reference CV value in that window. Each window is simulated for 4 ps, and the last 2 ps are used to calculate the PMF using the WHAM method. The errors are calculated using a bootstrapping procedure with a correlation time of 50 fs and 1000 Monte Carlo trials. Larger correlation times can lower the number of data points, and the results often does not converge. The results are shown in Fig. 4.6. For US window distributions and metadynamics PMF convergence see Appendix D. Comparing the PMF in Fig. 4.6, bottom left panel, to the PMF in Fig. 4.2, it shows that both US/mCEC (blue curve) and WTMetaD/mCEC (red curve) match the US/LinComb qualitatively. All the results exhibit three minima in the PMF profile. US/mCEC has a higher barrier of about 5 kcal/mol than the PMF from US/LinComb (see Fig. 4.2) and a lower PMF difference between the reactant and product states of about 6 kcal/mol. However, using WTMetaD/mCEC yields a barrier comparable to that of US/mCEC and energetics comparable to those of US/LinComb shown in Fig. 4.2. This highlights that, when using mCEC CV as the reaction coordinate, both enhanced sampling methods yield similar barriers, whereas only WTMetaD yields the more favorable energetics. Including additional water molecules in the QM region (see

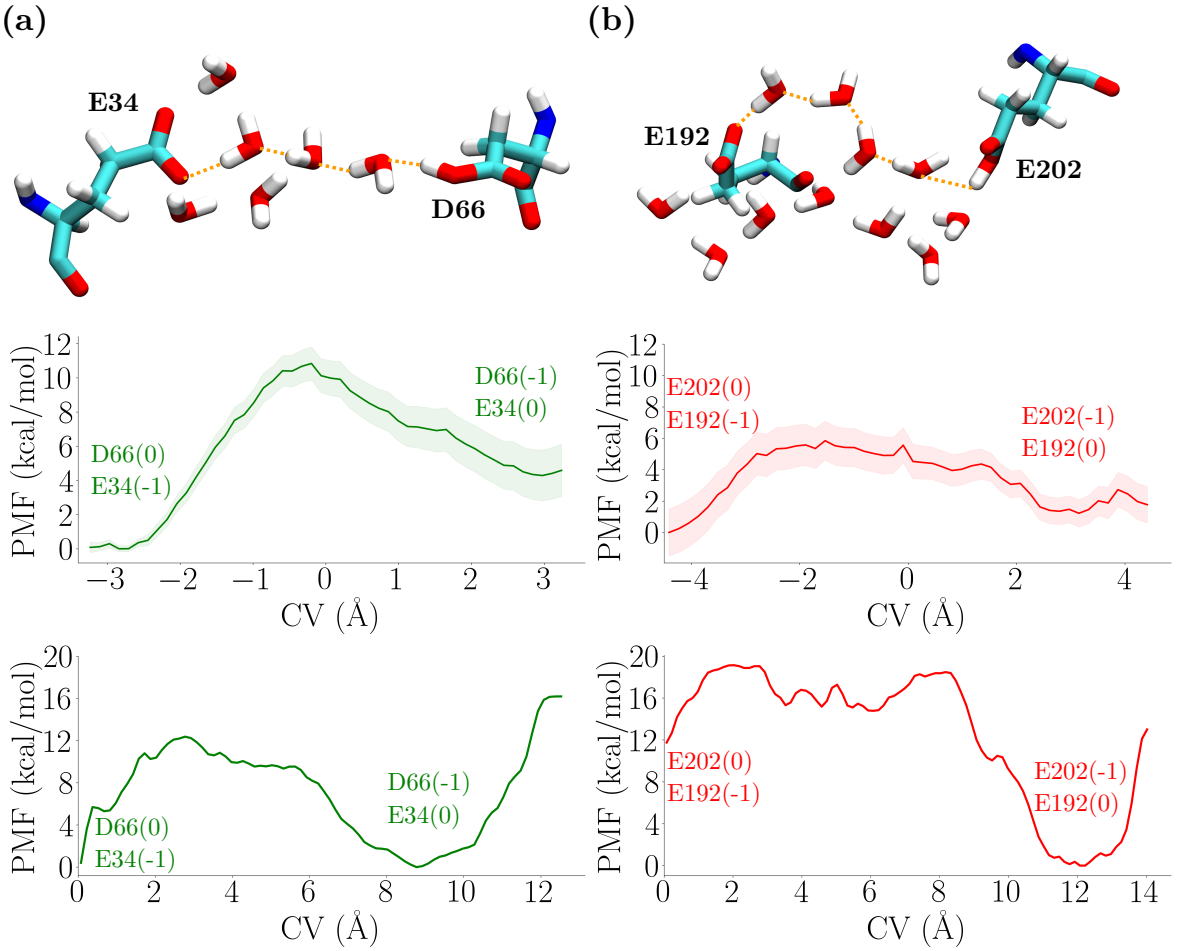


Figure 4.7: Top panel: (a) PT from ASP66 to GLU34, (b) PT from E202 to E192. The water molecules linked by the orange dotted lines are used as part of the LinComb CV, while the surrounding water molecules are only involved in the mCEC CV. Middle panel: the PMF profiles corresponding to the top panel, obtained using the umbrella sampling and LinComb CV. The results are adapted from Simsive *et al.* [36]. Bottom panel: the PMF profiles corresponding to the top panel, obtained using the well-tempered metadynamics and mCEC CV.

Fig. 4.6, right side) has further changed both the energy barriers and the energetics. The barriers are slightly higher, and the product state is comparatively less favorable. However, the same pattern is observed: WTMetaD yields more favorable energetics than US, which can be attributed to the absence of an exact, predetermined path in the biased QM/MM simulation, allowing the exploration of low-energy regions first. The results show that without an exact path to bias the CV, the system identifies a more appropriate path in the free energy landscape. This can drive the PT energetics into a more favorable direction.

Furthermore, to ensure the validity of the mCEC script and the WTMetaD parameters set, two extra PT pathways are explored. These pathways are located in the E channel of CI (see Fig. 3.1, bottom panel, right system), where the region is quite

hydrated, and the participating residues are acidic. These are previously investigated by Simsive *et al.* [36] using a LinComb CV and US enhanced sampling method. The PT pathways and the corresponding PMF profiles are shown in Fig. 4.7, top, and the middle panels, respectively. These pathways are re-investigated using mCEC CV with the WTMetaD method. The biased QM/MM simulation had the same parameters as well-tempered 3 (see Section 4.2) with a bias temperature equal to 4800 K. The results are shown in Fig. 4.7, bottom panel. The WTMetaD/mCEC simulation for the D66 and E34 PT pathway is 18.25 ps long and exhibits a barrier height similar to the US/LinComb, with the major difference that both the reactant and product states have the same PMF value, in contrast to previous results showing that the product state is unfavorable. For E202 to E192 PT, the simulation is 17.15 ps long, and the barrier obtained by mCEC is slightly higher, about 6.0 kcal/mol. However, the PMF difference between the reactant and product states is about 12.0 kcal/mol, indicating that the product state is favored. This result improved the energetics compared to the US/LinComb CV (see Fig. 4.7), where the product state is unfavorable by about 2 kcal/mol, due to both changing the RC to mCEC, where extra water molecules could be included in the RC and changing the enhanced sampling method to WTMetaD where the system goes over the smallest PMF barrier during the biased QM/MM simulation. In addition, a local minimum in the center is more pronounced with a small barrier of about 3 kcal/mol.

In summary, adding extra water molecules can improve the energetics of PT reactions in a particular PMF profile. The newly implemented reaction coordinate (mCEC) has been shown to be a powerful reaction coordinate in both QM/MM umbrella sampling and metadynamics enhanced sampling methods, and is more universal and reliable for longer pathways.

Moreover, running all the simulations mentioned above has shown that MetaD is at least one order of magnitude faster than US in CPU time for the same setup. This means that employing MetaD for PT investigations constitutes a more appropriate choice, especially if computational resources are limited.

5. Conclusions

Proton transfer (PT) is a complex chemical process that is challenging to track experimentally due to its fast nature and lack of direct measurements. In addition, the complexity of side-chain dynamics makes PT reactions in proteins difficult to explore. At the same time, computational methods provide important atomistic insights into PT reactions and enable us to track their dynamics, thereby enhancing the value of the experimental data. Specifically, hybrid quantum mechanics/molecular mechanics (QM/MM) free energy calculations provide valuable insights by tracking the energetics of individual proton pathways. However, the parameters for these methods are not well-defined, where the choice of free energy method, the values of the imposed bias, and the selection of a reaction coordinate remain ambiguous when exploring PT reactions.

A powerful enhanced sampling method is metadynamics, which, unlike other methods, such as umbrella sampling, can substantially reduce the computational time required to investigate a particular PT pathway. In addition, given the nature of applied bias (Gaussian functions deposited at the immediate position of the collective variable), the system can explore additional low-energy regions that remain unsampled in the umbrella sampling simulations. However, the optimal input parameters for PT reactions within the QM/MM framework remain challenging to identify. In this thesis, different values of the Gaussian height, width, and deposition rate have been tested and compared with the umbrella sampling method to assess energy barriers and thermodynamic differences. Based on this research, the optimal parameters for both conventional and well-tempered metadynamics have been determined. The results are in good agreement with previous studies that used other enhanced sampling methods.

Furthermore, to study the dependence of the free energy profiles on the chosen reaction coordinate (RC), the modified center of excess charge (mCEC) collective variable (CV) has been implemented in the *Colvars* module. Testing it along various PT pathways, it has been shown that using the mCEC CV as the RC can lower energy barriers and yield more favorable PT energetics, leading to more reliable outcomes.

Overall, using the well-tempered metadynamics enhanced sampling method with mCEC CV can streamline the study of PT reactions by providing an efficient and accurate means of determining the PMF profile relevant to a particular PT pathway.

Bibliography

[1] G. A. Voth, "Computer simulation of proton solvation and transport in aqueous and biomolecular systems", *Accounts of Chemical Research*, vol. 39, no. 2, pp. 143–150, Feb. 2006, ISSN: 1520-4898. DOI: 10.1021/ar0402098. [Online]. Available: http://dx.doi.org/10.1021/ar0402098.

[2] S. Scheiner, P. Redfern, and E. A. Hillenbrand, "Factors influencing proton positions in biomolecules", *International Journal of Quantum Chemistry*, vol. 29, no. 4, pp. 817–827, Apr. 1986, ISSN: 1097-461X. DOI: 10.1002/qua.560290420. [Online]. Available: http://dx.doi.org/10.1002/qua.560290420.

[3] P. H. König, N. Ghosh, M. Hoffmann, M. Elstner, E. Tajkhorshid, T. Frauenheim, and Q. Cui, "Toward theoretical analyis of long-range proton transfer kinetics in biomolecular pumps", *The Journal of Physical Chemistry A*, vol. 110, no. 2, pp. 548–563, Jul. 2005, ISSN: 1520-5215. DOI: 10.1021/jp052328q. [Online]. Available: http://dx.doi.org/10.1021/jp052328q.

[4] C. de Grotthuss, "Memoir on the decomposition of water and of the bodies that it holds in solution by means of galvanic electricity", *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, vol. 1757, no. 8, pp. 871–875, Aug. 2006, ISSN: 0005-2728. DOI: 10.1016/j.bbabio.2006.07.004. [Online]. Available: http://dx.doi.org/10.1016/j.bbabio.2006.07.004.

[5] A. Warshel and M. Levitt, "Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme", *Journal of Molecular Biology*, vol. 103, no. 2, pp. 227–249, May 1976, ISSN: 0022-2836. DOI: 10.1016/0022-2836(76)90311-9. [Online]. Available: http://dx.doi.org/10.1016/0022-2836(76)90311-9.

[6] A. Shurki and A. Warshel, "Structure/function correlations of proteins using mm, qm/mm, and related approaches: Methods, concepts, pitfalls, and current progress", in *Protein Simulations*. Elsevier, 2003, pp. 249–313, ISBN: 9780120342662. DOI: 10.1016/s0065-3233(03)66007-9. [Online]. Available: http://dx.doi.org/10.1016/S0065-3233(03)66007-9.

[7] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding", *Chemical Physics Letters*, vol. 314, no. 1–2, pp. 141–151, Nov. 1999, ISSN: 0009-2614. DOI: 10.1016/s0009-2614(99)01123-9. [Online]. Available: http://dx.doi.org/10.1016/S0009-2614(99)01123-9.

[8] E. Darve, D. Rodríguez-Gómez, and A. Pohorille, "Adaptive biasing force method for scalar and vector free energy calculations", *The Journal of Chemical Physics*, vol. 128, no. 14, Apr. 2008, ISSN: 1089-7690. DOI: 10.1063/1.2829861. [Online]. Available: http://dx.doi.org/10.1063/1.2829861.

[9] G. M. Torrie and J. P. Valleau, "Monte carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid", *Chemical Physics Letters*, vol. 28, no. 4, pp. 578–581, Oct. 1974, ISSN: 0009-2614. DOI: 10.1016/0009-2614(74)80109-0. [Online]. Available: http://dx.doi.org/10.1016/0009-2614(74)80109-0.

[10] G. Torrie and J. Valleau, "Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling", *Journal of Computational Physics*, vol. 23, no. 2, pp. 187–199, Feb. 1977, ISSN: 0021-9991. DOI: 10.1016/0021-9991(77)90121-8. [Online]. Available: http://dx.doi.org/10.1016/0021-9991(77)90121-8.

[11] A. Laio and M. Parrinello, "Escaping free-energy minima", *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12 562–12 566, Sep. 2002, ISSN: 1091-6490. DOI: 10.1073/pnas.202427399. [Online]. Available: http://dx.doi.org/10.1073/pnas.202427399.

[12] A. Barducci, M. Bonomi, and M. Parrinello, "Metadynamics", *WIREs Computational Molecular Science*, vol. 1, no. 5, pp. 826–843, Feb. 2011, ISSN: 1759-0884. DOI: 10.1002/wcms.31. [Online]. Available: http://dx.doi.org/10.1002/wcms.31.

[13] B. Roux, *Computational modeling and simulations of biomolecular systems*. World Scientific, 2022.

[14] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln", *Annalen der Physik*, vol. 389, no. 20, pp. 457–484, Jan. 1927, ISSN: 1521-3889. DOI: 10.1002/andp.19273892002. [Online]. Available: http://dx.doi.org/10.1002/andp.19273892002.

[15] D. R. Hartree, "The wave mechanics of an atom with a non-coulomb central field. part i. theory and methods", in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge university press, vol. 24, 1928, pp. 89–110.

[16] C. D. Sherrill, "An introduction to hartree-fock molecular orbital theory", *School of Chemistry and Biochemistry Georgia Institute of Technology*, vol. 1, 2000.

[17] J. C. Slater, "The theory of complex spectra", *Physical Review*, vol. 34, no. 10, pp. 1293–1322, Nov. 1929, ISSN: 0031-899X. DOI: 10.1103/physrev.34.1293. [Online]. Available: http://dx.doi.org/10.1103/PhysRev.34.1293.

[18] T. Tsuneda, *Density Functional Theory in Quantum Chemistry*. Springer Japan, 2014, ISBN: 9784431548256. DOI: 10.1007/978-4-431-54825-6. [Online]. Available: http://dx.doi.org/10.1007/978-4-431-54825-6.

[19] Y. Shikano, H. C. Watanabe, K. M. Nakanishi, and Y.-y. Ohnishi, "Post-hartree–fock method in quantum chemistry for quantum computer", *The European Physical Journal Special Topics*, vol. 230, no. 4, pp. 1037–1051, Apr. 2021, ISSN: 1951-6401. DOI: 10.1140/epjs/s11734-021-00087-z. [Online]. Available: http://dx.doi.org/10.1140/epjs/s11734-021-00087-z.

[20] G. Berthier, M. Defranceschi, and C. Le Bris, "Shortcomings in computational chemistry", *International Journal of Quantum Chemistry*, vol. 93, no. 3, pp. 156–165, Jan. 2003, ISSN: 1097-461X. DOI: 10.1002/qua.10550. [Online]. Available: http://dx.doi.org/10.1002/qua.10550.

[21] R. Car, "Introduction to density-functional theory and ab-initio molecular dynamics", *Quantitative Structure-Activity Relationships*, vol. 21, no. 2, pp. 97–104, Jul. 2002, ISSN: 1521-3838. DOI: 10.1002/1521-3838(200207)21:2<97::aid-qsar97>3.0.co;2-6. [Online]. Available: http://dx.doi.org/10.1002/1521-3838(200207)21:2%3C97::AID-QSAR97%3E3.0.CO;2-6.

[22] P. A. M. Dirac, "Note on exchange phenomena in the thomas atom", *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, no. 3, pp. 376–385, Jul. 1930, ISSN: 1469-8064. DOI: 10.1017/s0305004100016108. [Online]. Available: http://dx.doi.org/10.1017/S0305004100016108.

[23] L. H. Thomas, "The calculation of atomic fields", *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 23, no. 5, pp. 542–548, Jan. 1927, ISSN: 1469-8064. DOI: 10.1017/s0305004100011683. [Online]. Available: http://dx.doi.org/10.1017/S0305004100011683.

[24] E. Fermi, "Un metodo statistico per la determinazione di alcune priorieta dell'atome", *Rend. Accad. Naz. Lincei*, vol. 6, no. 602-607, p. 32, 1927.

[25] F. Jensen, *Introduction to computational chemistry*. John wiley & sons, 2017.

[26] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas", *Physical Review*, vol. 136, no. 3B, B864–B871, Nov. 1964, ISSN: 0031-899X. DOI: 10.1103/physrev.136.b864. [Online]. Available: http://dx.doi.org/10.1103/PhysRev.136.B864.

[27] P. Geerlings, F. De Proft, and W. Langenaeker, "Conceptual density functional theory", *Chemical Reviews*, vol. 103, no. 5, pp. 1793–1874, Apr. 2003, ISSN: 1520-6890. DOI: 10.1021/cr990029p. [Online]. Available: http://dx.doi.org/10.1021/cr990029p.

[28] W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", *Physical Review*, vol. 140, no. 4A, A1133–A1138, Nov. 1965, ISSN: 0031-899X. DOI: 10.1103/physrev.140.a1133. [Online]. Available: http://dx.doi.org/10.1103/PhysRev.140.A1133.

[29] N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals", *Molecular Physics*, vol. 115, no. 19, pp. 2315–2372, Jun. 2017, ISSN: 1362-3028. DOI: 10.1080/00268976.2017.1333644. [Online]. Available: http://dx.doi.org/10.1080/00268976.2017.1333644.

[30] S. Kurth and J. P. Perdew, "Role of the exchange-correlation energy: Nature's glue", *International Journal of Quantum Chemistry*, vol. 77, no. 5, pp. 814–818, 2000, ISSN: 1097-461X. DOI: 10.1002/(sici)1097-461x(2000)77:5<814::aid-qua3>3.0.co;2-f. [Online]. Available: http://dx.doi.org/10.1002/(SICI)1097-461X(2000)77:5%3C814::AID-QUA3%3E3.0.CO;2-F.

[31] D. Bagayoko, "Understanding density functional theory (dft) and completing it in practice", *AIP Advances*, vol. 4, no. 12, Dec. 2014, ISSN: 2158-3226. DOI: 10.1063/1.4903408. [Online]. Available: http://dx.doi.org/10.1063/1.4903408.

[32] K. P. Zois and D. Tzeli, "A critical look at density functional theory in chemistry: Untangling its strengths and weaknesses", *Atoms*, vol. 12, no. 12, p. 65, Dec. 2024, ISSN: 2218-2004. DOI: 10.3390/atoms12120065. [Online]. Available: http://dx.doi.org/10.3390/atoms12120065.

[33] J. Simons, "Why is quantum chemistry so complicated?", *Journal of the American Chemical Society*, vol. 145, no. 8, pp. 4343–4354, Feb. 2023, ISSN: 1520-5126. DOI: 10.1021/jacs.2c13042. [Online]. Available: http://dx.doi.org/10.1021/jacs.2c13042.

[34] J. P. Perdew and A. Zunger, "Self-interaction correction to density-functional approximations for many-electron systems", *Physical Review B*, vol. 23, no. 10, pp. 5048–5079, May 1981, ISSN: 0163-1829. DOI: 10.1103/physrevb.23.5048. [Online]. Available: http://dx.doi.org/10.1103/PhysRevB.23.5048.

[35] V. R. I. Kaila, "Resolving chemical dynamics in biological energy conversion: Long-range proton-coupled electron transfer in respiratory complex i", *Accounts of Chemical Research*, vol. 54, no. 24, pp. 4462–4473, Dec. 2021, ISSN: 1520-4898. DOI: 10.1021/acs.accounts.1c00524. [Online]. Available: http://dx.doi.org/10.1021/acs.accounts.1c00524.

[36] L. Simsive, O. Zdorevskyi, and V. Sharma, "Proton transfer through a charged conduit in respiratory complex I: Long-range effects and conformational gating", *Journal of Chemical Information and Modeling*, vol. 65, no. 19, pp. 10 600–10 612, Sep. 2025, ISSN: 1549-960X. DOI: 10.1021/acs.jcim.5c01365. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.5c01365.

[37] O. Zdorevskyi, A. Djurabekova, J. Lasham, and V. Sharma, "Horizontal proton transfer across the antiporter-like subunits in mitochondrial respiratory complex i", *Chemical Science*, vol. 14, no. 23, pp. 6309–6318, 2023, ISSN: 2041-6539. DOI: 10.1039/d3sc01427d. [Online]. Available: http://dx.doi.org/10.1039/D3SC01427D.

[38] F. Jensen, "Atomic orbital basis sets", *WIREs Computational Molecular Science*, vol. 3, no. 3, pp. 273–295, Oct. 2012, ISSN: 1759-0884. DOI: 10.1002/wcms.1123. [Online]. Available: http://dx.doi.org/10.1002/wcms.1123.

[39] A. D. Becke, "Becke's three parameter hybrid method using the lyp correlation functional", *J. Chem. Phys*, vol. 98, no. 492, pp. 5648–5652, 1993.

[40] C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of the electron density", *Physical Review B*, vol. 37, no. 2, pp. 785–789, Jan. 1988, ISSN: 0163-1829. DOI: 10.1103/physrevb.37.785. [Online]. Available: http://dx.doi.org/10.1103/PhysRevB.37.785.

[41] S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: A critical analysis", *Canadian Journal of Physics*, vol. 58, no. 8, pp. 1200–1211, Aug. 1980, ISSN: 1208-6045. DOI: 10.1139/p80-159. [Online]. Available: http://dx.doi.org/10.1139/p80-159.

[42] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields", *The Journal of physical chemistry*, vol. 98, no. 45, pp. 11 623–11 627, 1994.

[43] J. Klimeš and A. Michaelides, "Perspective: Advances and challenges in treating van der waals dispersion forces in density functional theory", *The Journal of Chemical Physics*, vol. 137, no. 12, Sep. 2012, ISSN: 1089-7690. DOI: 10.1063/1.4754130. [Online]. Available: http://dx.doi.org/10.1063/1.4754130.

[44] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurateab initioparametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu", *The Journal of Chemical Physics*, vol. 132, no. 15, Apr. 2010, ISSN: 1089-7690. DOI: 10.1063/1.3382344. [Online]. Available: http://dx.doi.org/10.1063/1.3382344.

[45] A. D. Becke and E. R. Johnson, "A density-functional model of the dispersion interaction", *The Journal of Chemical Physics*, vol. 123, no. 15, Oct. 2005, ISSN: 1089-7690. DOI: 10.1063/1.2065267. [Online]. Available: http://dx.doi.org/10.1063/1.2065267.

[46] K. Vanommeslaeghe, O. Guvench, and A. D. MacKerell Jr, "Molecular mechanics", *Current pharmaceutical design*, vol. 20, no. 20, pp. 3281–3292, 2014.

[47] C. Sagui and T. A. Darden, "Molecular dynamics simulations of biomolecules: Long-range electrostatic effects", *Annual Review of Biophysics and Biomolecular Structure*, vol. 28, no. 1, pp. 155–179, Jun. 1999, ISSN: 1545-4266. DOI: 10.1146/annurev.biophys.28.1.155. [Online]. Available: http://dx.doi.org/10.1146/annurev.biophys.28.1.155.

[48] P. P. Ewald, "Die berechnung optischer und elektrostatischer gitterpotentiale", *Annalen der Physik*, vol. 369, no. 3, pp. 253–287, Jan. 1921, ISSN: 1521-3889. DOI: 10.1002/andp.19213690304. [Online]. Available: http://dx.doi.org/10.1002/andp.19213690304.

[49] T. Darden, D. York, and L. Pedersen, "Particle mesh ewald: An n·log(n) method for ewald sums in large systems", *The Journal of Chemical Physics*, vol. 98, no. 12, pp. 10 089–10 092, Jun. 1993, ISSN: 1089-7690. DOI: 10.1063/1.464397. [Online]. Available: http://dx.doi.org/10.1063/1.464397.

[50] H. G. Petersen, "Accuracy and efficiency of the particle mesh ewald method", *The Journal of Chemical Physics*, vol. 103, no. 9, pp. 3668–3679, Sep. 1995, ISSN: 1089-7690. DOI: 10.1063/1.470043. [Online]. Available: http://dx.doi.org/10.1063/1.470043.

[51] J. W. Ponder and D. A. Case, "Force fields for protein simulations", in *Protein Simulations*. Elsevier, 2003, pp. 27–85, ISBN: 9780120342662. DOI: 10.1016/s0065-3233(03)66002-x. [Online]. Available: http://dx.doi.org/10.1016/S0065-3233(03)66002-X.

[52] R. B. Best, X. Zhu, J. Shim, P. E. M. Lopes, J. Mittal, M. Feig, and A. D. MacKerell, "Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ_1 and χ_2 dihedral angles", *Journal of Chemical Theory and Computation*, vol. 8, no. 9, pp. 3257–3273, Aug. 2012, ISSN: 1549-9626. DOI: 10.1021/ct300400x. [Online]. Available: http://dx.doi.org/10.1021/ct300400x.

[53] J. B. Klauda, R. M. Venable, J. A. Freites, J. W. O'Connor, D. J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. D. MacKerell, and R. W. Pastor, "Update of the charmm all-atom additive force field for lipids: Validation on six lipid types", *The Journal of Physical Chemistry B*, vol. 114, no. 23, pp. 7830–7843, May 2010, ISSN: 1520-5207. DOI: 10.1021/jp101759q. [Online]. Available: http://dx.doi.org/10.1021/jp101759q.

[54] D. Beglov and B. Roux, "Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations", *The Journal of Chemical Physics*, vol. 100, no. 12, pp. 9050–9063, Jun. 1994, ISSN: 1089-7690. DOI: 10.1063/1.466711. [Online]. Available: http://dx.doi.org/10.1063/1.466711.

[55] S. Hug, "Classical molecular dynamics in a nutshell", in *Biomolecular Simulations*. Humana Press, Aug. 2012, pp. 127–152, ISBN: 9781627030175. DOI: 10.1007/978-1-62703-017-5_6. [Online]. Available: http://dx.doi.org/10.1007/978-1-62703-017-5_6.

[56] M. Karplus and G. A. Petsko, "Molecular dynamics simulations in biology", *Nature*, vol. 347, no. 6294, pp. 631–639, 1990.

[57] S. Gorbunov, A. Volkov, and R. Voronkov, "Periodic boundary conditions effects on atomic dynamics analysis", *Computer Physics Communications*, vol. 279, p. 108 454, Oct. 2022, ISSN: 0010-4655. DOI: 10.1016/j.cpc.2022.108454. [Online]. Available: http://dx.doi.org/10.1016/j.cpc.2022.108454.

[58] M. Bhandarkar, R. Brunner, C. Chipot, A. Dalke, S. Dixit, P. Grayson, J. Gullingsrud, A. Gursoy, D. Hardy, W. Humphrey, *et al.*, "Namd user's guide", *Urbana*, vol. 51, p. 61 801, 2003.

[59] S. Patodia, "Molecular dynamics simulation of proteins: A brief overview", *Journal of Physical Chemistry & amp; Biophysics*, vol. 4, no. 6, 2014, ISSN: 2161-0398. DOI: 10.4172/2161-0398.1000166. [Online]. Available: http://dx.doi.org/10.4172/2161-0398.1000166.

[60] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*. Oxford university press, 2017.

[61] Z. Huang, L. Zheng, and W. Yang, "An accurate and efficient npt ensemble molecular dynamics simulation method", *The Journal of Chemical Physics*, vol. 164, no. 1, Jan. 2026, ISSN: 1089-7690. DOI: 10.1063/5.0307402. [Online]. Available: http://dx.doi.org/10.1063/5.0307402.

[62] N. Shuichi, "Constant temperature molecular dynamics methods", *Progress of Theoretical Physics Supplement*, vol. 103, pp. 1–46, 1991, ISSN: 0375-9687. DOI: 10.1143/ptps.103.1. [Online]. Available: http://dx.doi.org/10.1143/PTPS.103.1.

[63] O. Farago, "Langevin thermostat for robust configurational and kinetic sampling", *Physica A: Statistical Mechanics and its Applications*, vol. 534, p. 122 210, Nov. 2019, ISSN: 0378-4371. DOI: 10.1016/j.physa.2019.122210. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2019.122210.

[64] D. Toton, C. D. Lorenz, N. Rompotis, N. Martsinovich, and L. Kantorovich, "Temperature control in molecular dynamic simulations of non-equilibrium processes", *Journal of Physics: Condensed Matter*, vol. 22, no. 7, p. 074 205, Feb. 2010, ISSN: 1361-648X. DOI: 10.1088/0953-8984/22/7/074205. [Online]. Available: http://dx.doi.org/10.1088/0953-8984/22/7/074205.

[65] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath", *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984, ISSN: 1089-7690. DOI: 10.1063/1.448118. [Online]. Available: http://dx.doi.org/10.1063/1.448118.

[66] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method", *Journal of Applied Physics*, vol. 52, no. 12, pp. 7182–7190, Dec. 1981, ISSN: 1089-7550. DOI: 10.1063/1.328693. [Online]. Available: http://dx.doi.org/10.1063/1.328693.

[67] S. E. Feller, Y. Zhang, R. W. Pastor, and B. R. Brooks, "Constant pressure molecular dynamics simulation: The langevin piston method", *The Journal of Chemical Physics*, vol. 103, no. 11, pp. 4613–4621, Sep. 1995, ISSN: 1089-7690.

DOI: 10.1063/1.470648. [Online]. Available: http://dx.doi.org/10.1063/1.470648.

[68] Q. Ke, X. Gong, S. Liao, C. Duan, and L. Li, "Effects of thermostats/barostats on physical properties of liquids by molecular dynamics simulations", *Journal of Molecular Liquids*, vol. 365, p. 120 116, Nov. 2022, ISSN: 0167-7322. DOI: 10.1016/j.molliq.2022.120116. [Online]. Available: http://dx.doi.org/10.1016/j.molliq.2022.120116.

[69] C. M. Clemente, L. Capece, and M. A. Martı, "Best practices on qm/mm simulations of biological systems", *Journal of Chemical Information and Modeling*, vol. 63, no. 9, pp. 2609–2627, Apr. 2023, ISSN: 1549-960X. DOI: 10.1021/acs.jcim.2c01522. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.2c01522.

[70] H. J. Kulik, J. Zhang, J. P. Klinman, and T. J. Martínez, "How large should the qm region be in qm/mm calculations? the case of catecholo-methyltransferase", *The Journal of Physical Chemistry B*, vol. 120, no. 44, pp. 11 381–11 394, Oct. 2016, ISSN: 1520-5207. DOI: 10.1021/acs.jpcb.6b07814. [Online]. Available: http://dx.doi.org/10.1021/acs.jpcb.6b07814.

[71] G. Groenhof, "Introduction to qm/mm simulations", in *Biomolecular Simulations*. Humana Press, Aug. 2012, pp. 43–66, ISBN: 9781627030175. DOI: 10.1007/978-1-62703-017-5_3. [Online]. Available: http://dx.doi.org/10.1007/978-1-62703-017-5_3.

[72] M. C. Melo, R. C. Bernardi, T. Rudack, M. Scheurer, C. Riplinger, J. C. Phillips, J. D. Maia, G. B. Rocha, J. V. Ribeiro, J. E. Stone, *et al.*, "Namd goes quantum: An integrative suite for hybrid simulations", *Nature Methods*, vol. 15, no. 5, pp. 351–354, Mar. 2018, ISSN: 1548-7105. DOI: 10.1038/nmeth.4638. [Online]. Available: http://dx.doi.org/10.1038/nmeth.4638.

[73] F. Neese, "The orca program system", *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 2, no. 1, pp. 73–78, 2012.

[74] J. K. Szántó, J. C. B. Dietschreit, M. Shein, A. K. Schütz, and C. Ochsenfeld, "Systematic qm/mm study for predicting 31p nmr chemical shifts of adenosine nucleotides in solution and stages of atp hydrolysis in a protein environment", *Journal of Chemical Theory and Computation*, vol. 20, no. 6, pp. 2433–2444, Mar. 2024, ISSN: 1549-9626. DOI: 10.1021/acs.jctc.3c01280. [Online]. Available: http://dx.doi.org/10.1021/acs.jctc.3c01280.

[75] D. A. Kofke, "Free energy methods in molecular simulation", *Fluid Phase Equilibria*, vol. 228–229, pp. 41–48, Feb. 2005, ISSN: 0378-3812. DOI: 10.1016/j.fluid.2004.09.017. [Online]. Available: http://dx.doi.org/10.1016/j.fluid.2004.09.017.

[76] J. Hénin, T. Leliévre, M. R. Shirts, O. Valsson, and L. Delemotte, "Enhanced sampling methods for molecular dynamics simulations [article v1.0]", *Living Journal of Computational Molecular Science*, vol. 4, no. 1, p. 1583, Dec. 2022, ISSN: 2575-6524. DOI: 10.33011/livecoms.4.1.1583. [Online]. Available: http://dx.doi.org/10.33011/livecoms.4.1.1583.

[77] D. L. Beveridge and F. M. Dicapua, "Free energy via molecular simulation: Applications to chemical and biomolecular systems", *Annual review of biophysics and biophysical chemistry*, vol. 18, no. 1, pp. 431–492, 1989.

[78] M. E. Tuckerman, *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2023.

[79] G. Fiorin, M. L. Klein, and J. Hénin, "Using collective variables to drive molecular dynamics simulations", *Molecular Physics*, vol. 111, no. 22–23, pp. 3345–3362, Dec. 2013, ISSN: 1362-3028. DOI: 10.1080/00268976.2013.813594. [Online]. Available: http://dx.doi.org/10.1080/00268976.2013.813594.

[80] J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot, and E. Tajkhorshid, "Scalable molecular dynamics on cpu and gpu architectures with namd", *The Journal of Chemical Physics*, vol. 153, no. 4, Jul. 2020, ISSN: 1089-7690. DOI: 10.1063/5.0014475. [Online]. Available: http://dx.doi.org/10.1063/5.0014475.

[81] R. Pomès and B. Roux, "Free energy profiles for H^+ conduction along hydrogen-bonded chains of water molecules", *Biophysical Journal*, vol. 75, no. 1, pp. 33–40, Jul. 1998, ISSN: 0006-3495. DOI: 10.1016/s0006-3495(98)77492-2. [Online]. Available: http://dx.doi.org/10.1016/S0006-3495(98)77492-2.

[82] N. Chakrabarti, E. Tajkhorshid, B. Roux, and R. Pomès, "Molecular basis of proton blockage in aquaporins", *Structure*, vol. 12, no. 1, pp. 65–74, Mar. 2004, ISSN: 0969-2126. DOI: 10.1016/j.str.2003.11.017. [Online]. Available: http://dx.doi.org/10.1016/j.str.2003.11.017.

[83] H. Kim, P. Saura, M. C. Pöverlein, A. P. Gamiz-Hernandez, and V. R. I. Kaila, "Quinone catalysis modulates proton transfer reactions in the membrane domain of respiratory complex I", *Journal of the American Chemical Society*, vol. 145, no. 31, pp. 17 075–17 086, Jul. 2023, ISSN: 1520-5126. DOI: 10.1021/jacs.3c03086. [Online]. Available: http://dx.doi.org/10.1021/jacs.3c03086.

[84] J. Kästner, "Umbrella sampling", *WIREs Computational Molecular Science*, vol. 1, no. 6, pp. 932–942, May 2011, ISSN: 1759-0884. DOI: 10.1002/wcms.66. [Online]. Available: http://dx.doi.org/10.1002/wcms.66.

[85] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method", *Journal of Computational Chemistry*, vol. 13, no. 8, pp. 1011–1021, Oct. 1992, ISSN: 1096-987X. DOI: 10.1002/jcc.540130812. [Online]. Available: http://dx.doi.org/10.1002/jcc.540130812.

[86] A. Djurabekova, J. Lasham, O. Zdorevskyi, V. Zickermann, and V. Sharma, "Long-range electron proton coupling in respiratory complex I — insights from molecular simulations of the quinone chamber and antiporter-like subunits", *Biochemical Journal*, vol. 481, no. 7, pp. 499–514, Apr. 2024, ISSN: 1470-8728. DOI: 10.1042/bcj20240009. [Online]. Available: http://dx.doi.org/10.1042/BCJ20240009.

[87] G. Bussi and D. Branduardi, *Free-energy calculations with metadynamics: Theory and practice*, Apr. 2015. DOI: 10.1002/9781118889886.ch1. [Online]. Available: http://dx.doi.org/10.1002/9781118889886.ch1.

[88] O. Valsson, P. Tiwary, and M. Parrinello, "Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint", *Annual Review of Physical Chemistry*, vol. 67, no. 1, pp. 159–184, May 2016, ISSN: 1545-1593. DOI: 10.1146/annurev-physchem-040215-112229. [Online]. Available: http://dx.doi.org/10.1146/annurev-physchem-040215-112229.

[89] E. Buxbaum *et al.*, *Fundamentals of protein structure and function*. Springer, 2007, vol. 31.

[90] W. Kühlbrandt, L. A. Carreira, and Ö. Yildiz, "Cryo-em of mitochondrial complex I and atp synthase", *Annual Review of Biophysics*, vol. 54, no. 1, pp. 209–226, May 2025, ISSN: 1936-1238. DOI: 10.1146/annurev-biophys-060724-110838. [Online]. Available: http://dx.doi.org/10.1146/annurev-biophys-060724-110838.

[91] K. Parey, J. Lasham, D. J. Mills, A. Djurabekova, O. Haapanen, E. G. Yoga, H. Xie, W. Kühlbrandt, V. Sharma, J. Vonck, and V. Zickermann, "High-resolution structure and dynamics of mitochondrial complex I-insights into the proton pumping mechanism", *Science Advances*, vol. 7, no. 46, Nov. 2021, ISSN: 2375-2548. DOI: 10.1126/sciadv.abj3221. [Online]. Available: http://dx.doi.org/10.1126/sciadv.abj3221.

[92] J. Hirst, "Mitochondrial complex I", *Annual Review of Biochemistry*, vol. 82, no. 1, pp. 551–575, Jun. 2013, ISSN: 1545-4509. DOI: 10.1146/annurev-biochem-070511-103700. [Online]. Available: http://dx.doi.org/10.1146/annurev-biochem-070511-103700.

[93] N. Agmon, "The grotthuss mechanism", *Chemical Physics Letters*, vol. 244, no. 5-6, pp. 456–462, 1995.

[94] E. Endres, M. Torabi, M. Jousmäki, K. V. Huynh, C. Pecorilla, O. Zdorevskyi, V. Zickermann, and V. Sharma, "Molecular and energetic basis of histidine switch dynamics in respiratory complex I", *bioRxiv*, 2026. DOI: 10.64898/2026.01.12.698557. eprint: https://www.biorxiv.org/content/early/2026/01/12/2026.01.12.698557.full.pdf. [Online]. Available: https://www.biorxiv.org/content/early/2026/01/12/2026.01.12.698557.

[95] L. Verlet, "Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules", *Physical Review*, vol. 159, no. 1, pp. 98–103, Jul. 1967, ISSN: 0031-899X. DOI: 10.1103/physrev.159.98. [Online]. Available: http://dx.doi.org/10.1103/PhysRev.159.98.

[96] F. Weigend and R. Ahlrichs, "Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy", *Physical Chemistry Chemical Physics*, vol. 7, no. 18, p. 3297, 2005, ISSN: 1463-9084. DOI: 10.1039/b508541a. [Online]. Available: http://dx.doi.org/10.1039/B508541A.

[97] A. Grossfield, *WHAM: The weighted histogram analysis method*, version 2.1.0, http://membrane.urmc.rochester.edu/wordpress/?page_id=126, 2012.

[98] B. Efron, "Bootstrap methods: Another look at the jackknife", in *Breakthroughs in Statistics*. Springer New York, 1992, pp. 569–593, ISBN: 9781461243809. DOI: 10.1007/978-1-4612-4380-9_41. [Online]. Available: http://dx.doi.org/10.1007/978-1-4612-4380-9_41.

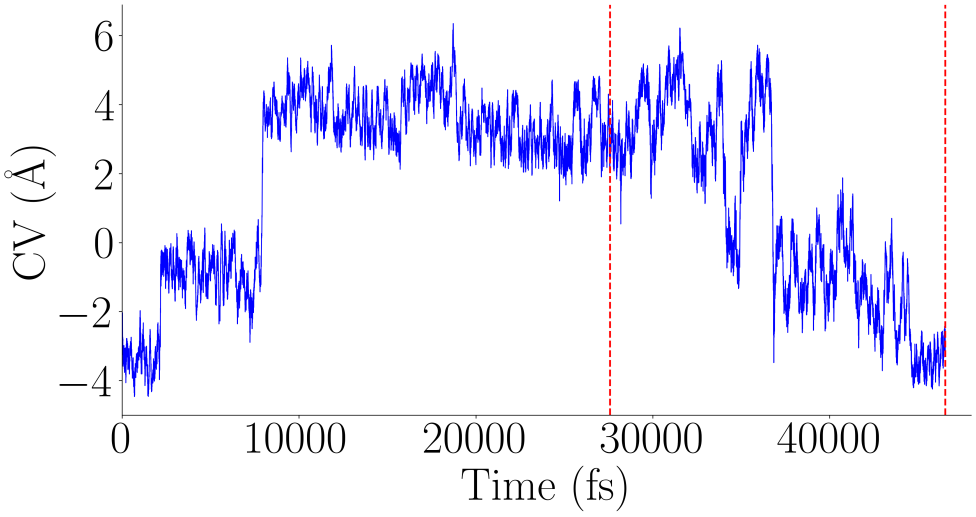
[99] A. Bernardin, H. Chen, J. R. Comer, G. Fiorin, H. Fu, J. Hénin, A. Kohlmeyer, F. Marinelli, H. Santuz, J. V. Vermaas, *et al.*, *Collective variables module reference manual for namd*, 2023.

[100] G. Bussi and G. A. Tribello, "Analyzing and biasing simulations with plumed", in *Biomolecular Simulations*. Springer New York, 2019, pp. 529–578, ISBN: 9781493996087. DOI: 10.1007/978-1-4939-9608-7_21. [Online]. Available: http://dx.doi.org/10.1007/978-1-4939-9608-7_21.

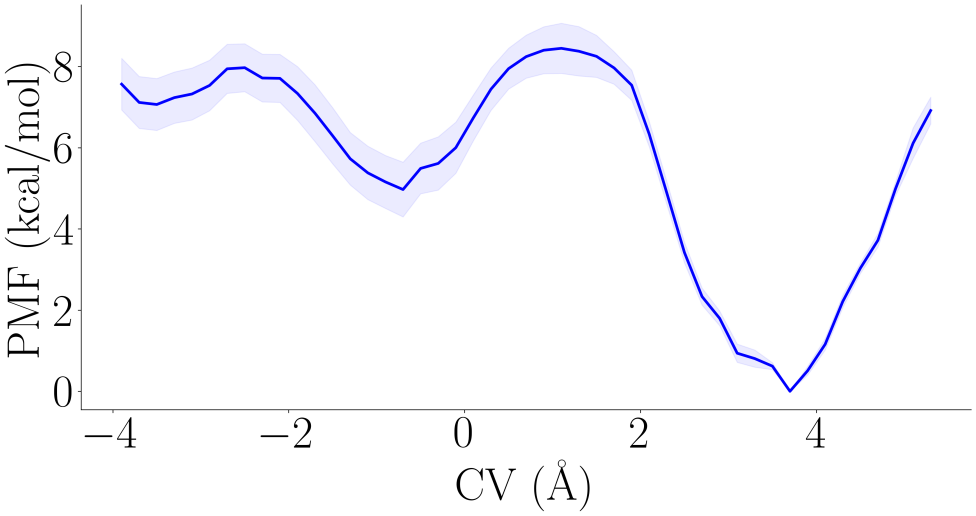
[101] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy", *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[102] R. Sun, O. Sode, J. F. Dama, and G. A. Voth, "Simulating protein mediated hydrolysis of ATP and other nucleoside triphosphates by combining QM/MM molecular dynamics with advances in metadynamics", *Journal of Chemical Theory and Computation*, vol. 13, no. 5, pp. 2332–2341, Apr. 2017, ISSN: 1549-9626. DOI: 10.1021/acs.jctc.7b00077. [Online]. Available: http://dx.doi.org/10.1021/acs.jctc.7b00077.

Appendix A. Parameter Testing Results

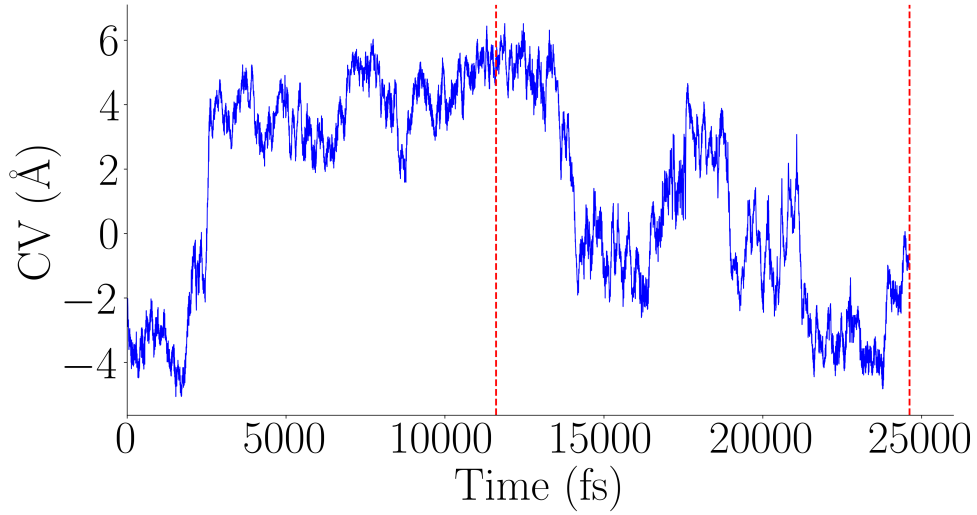


(a)

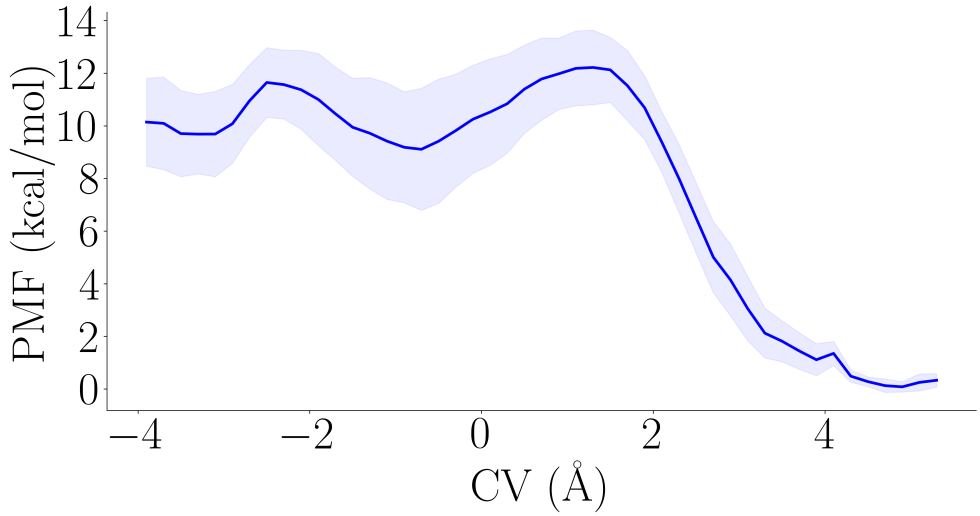


(b)

Figure A.1: (a) The linear combination CV trajectory during the biased simulation using the conventional-1 parameters. The red vertical dashed lines (at 27590 and 46540 fs) represent the limits for the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-1 parameters. A total of 1897 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).

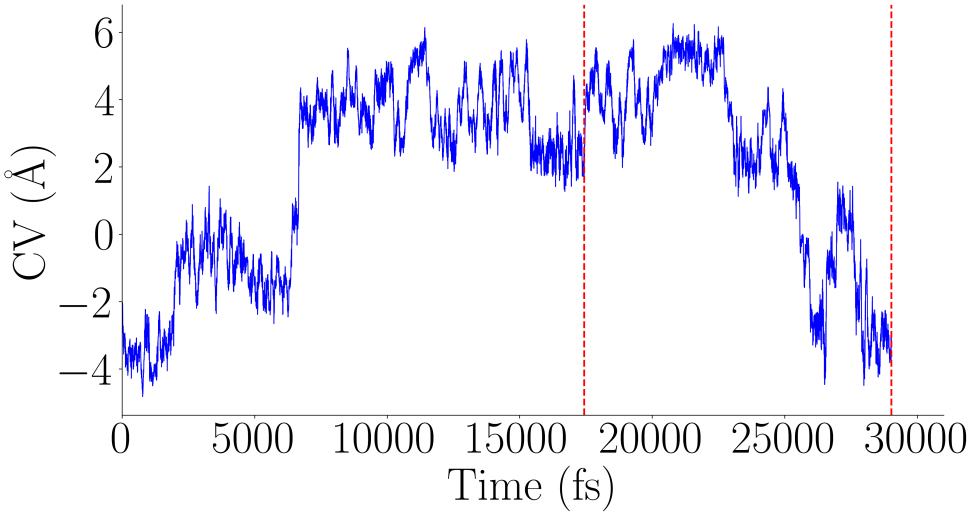


(a)

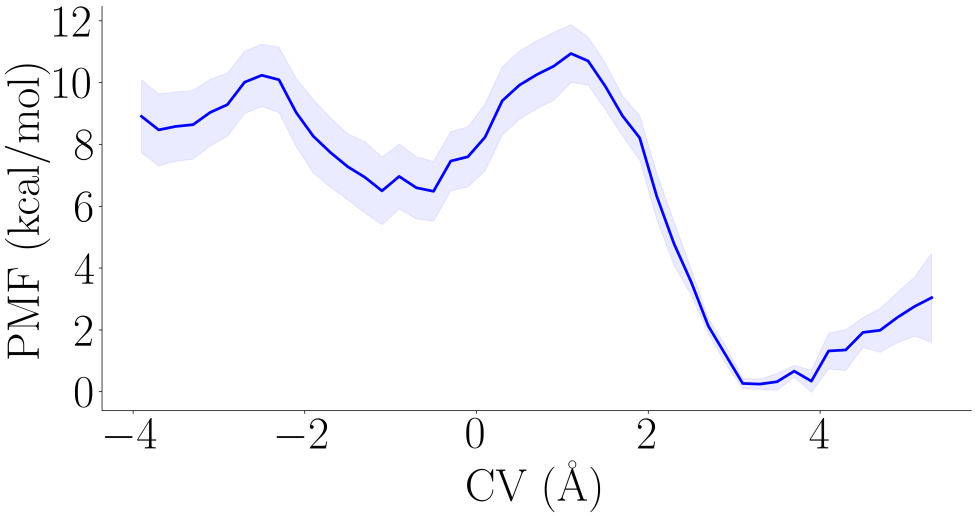


(b)

Figure A.2: (a) The linear combination CV trajectory during the biased simulation using the conventional-2 parameters. The red dashed vertical lines (at 11610 and 24620 fs) mark the limits of the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-2 parameters. A total of 1302 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).

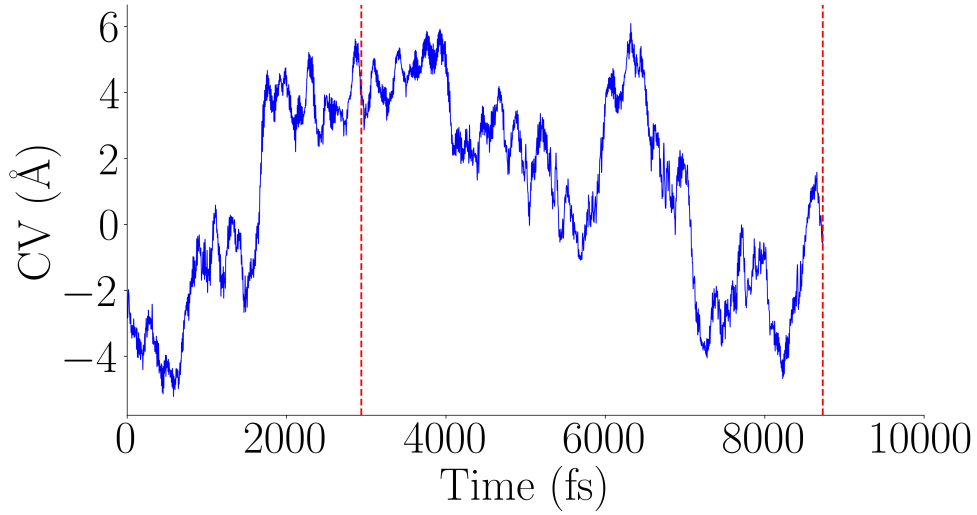


(a)

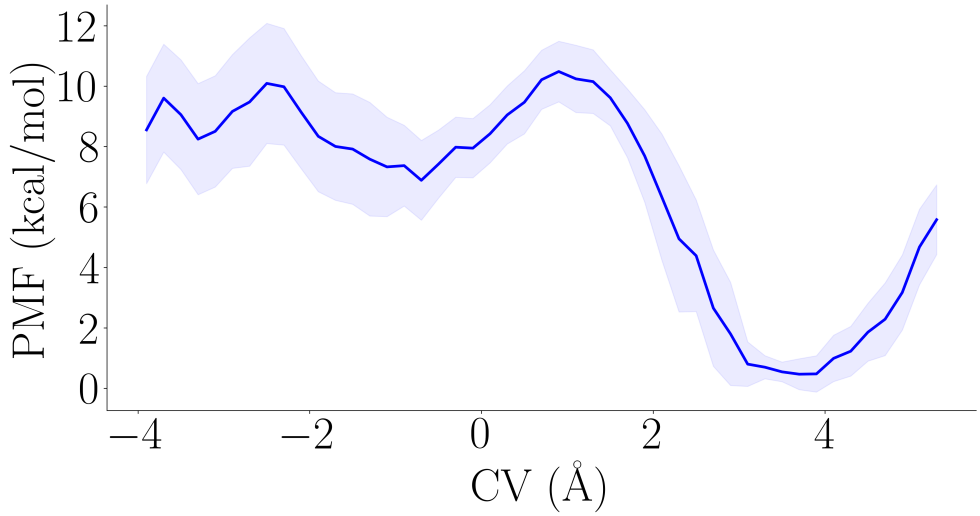


(b)

Figure A.3: (a) The linear combination CV trajectory during the biased simulation using the conventional-3 parameters. The red dashed vertical lines (at 17440 and 29030 fs) mark the limits of the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-3 parameters. A total of 1160 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).

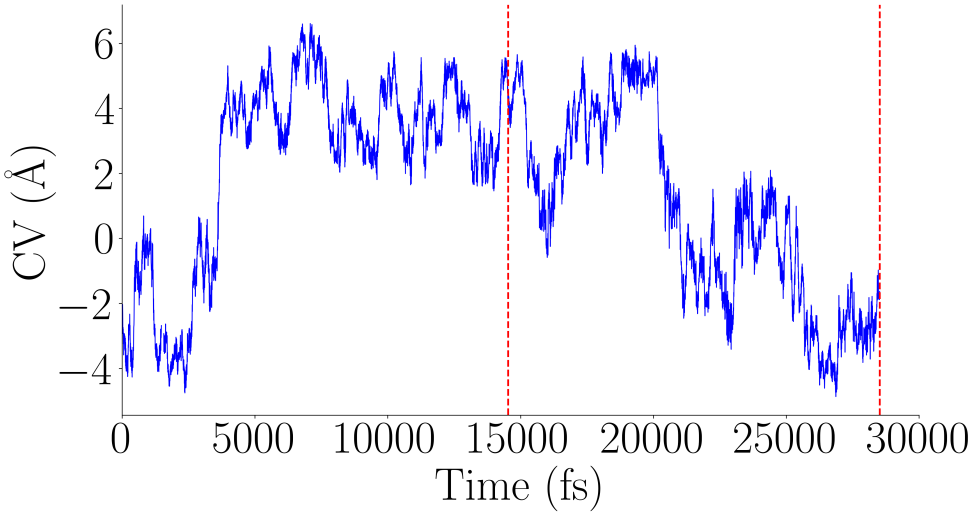


(a)

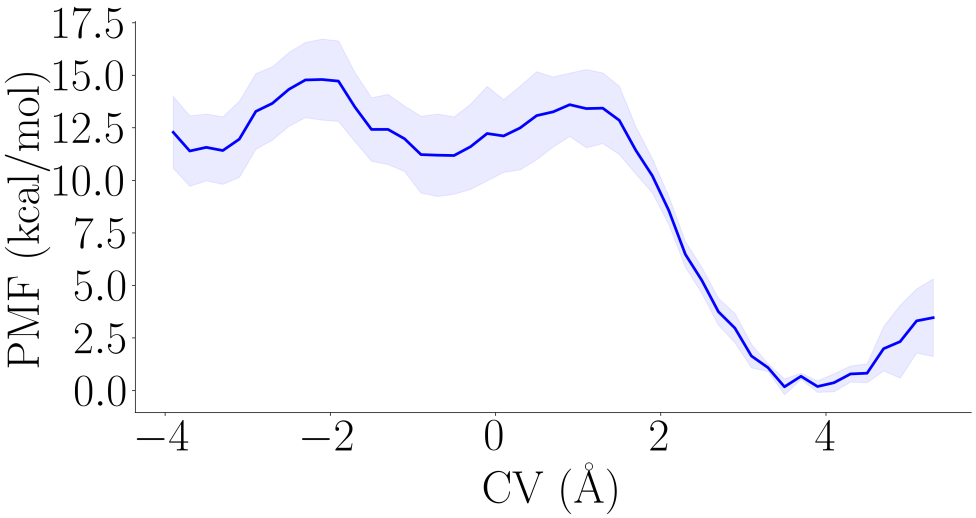


(b)

Figure A.4: (a) The linear combination CV trajectory during the biased simulation using the conventional-4 parameters. The red dashed vertical lines (at 2940 and 8730 fs) mark the limits of the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-4 parameters. A total of 580 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).

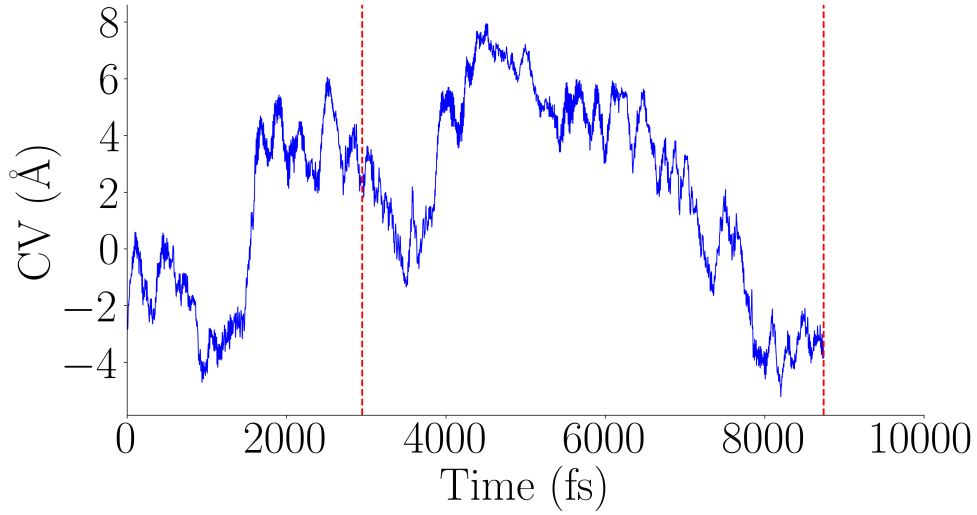


(a)

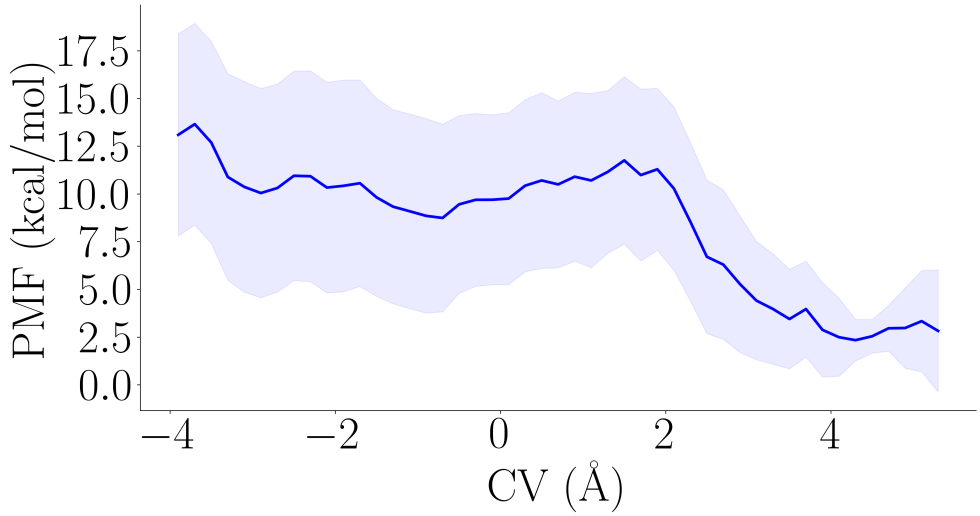


(b)

Figure A.5: (a) The linear combination CV trajectory during the biased simulation using the conventional-5 parameters. The red dashed vertical lines (at 14530 and 28520 fs) mark the limits of the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-5 parameters. A total of 1258 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).

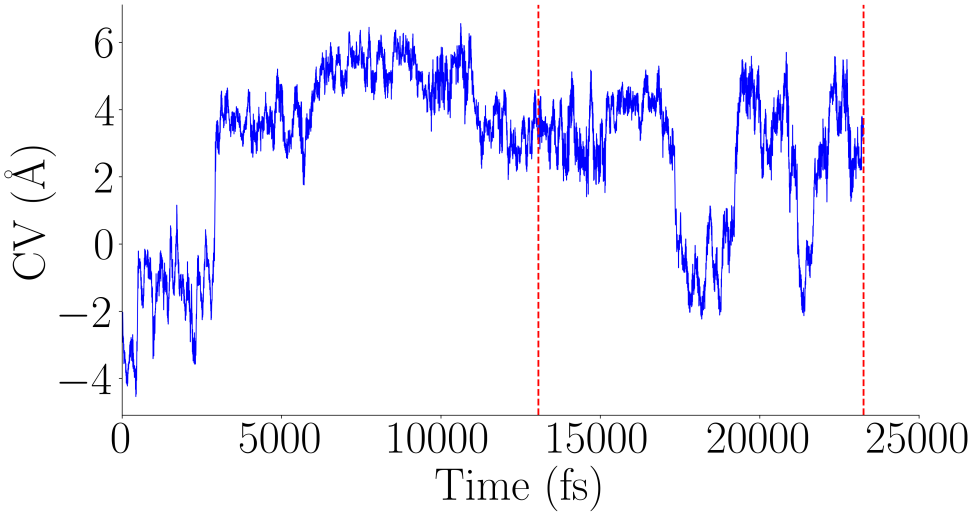


(a)

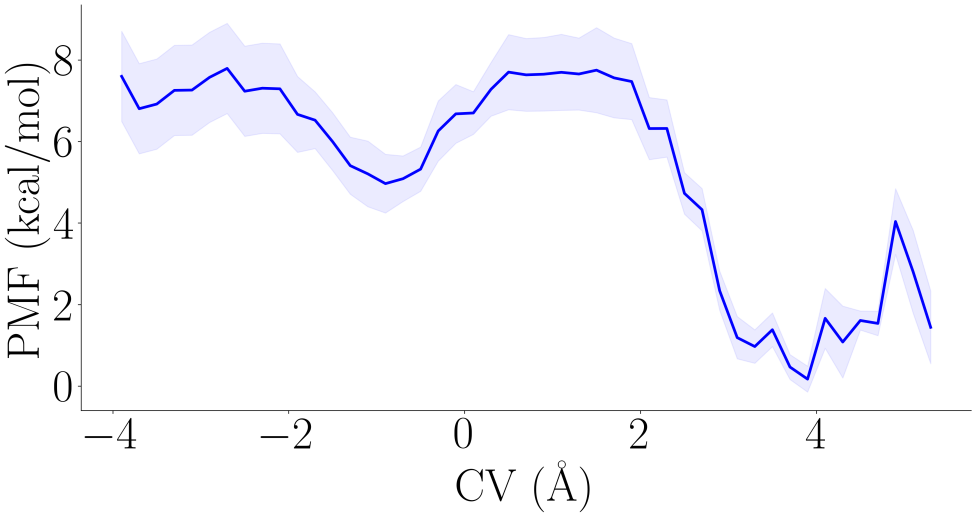


(b)

Figure A.6: (a) The linear combination CV trajectory during the biased simulation using the conventional-6 parameters. The red dashed vertical lines (at 2950 and 8740 fs) mark the limits of the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-6 parameters. A total of 580 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).

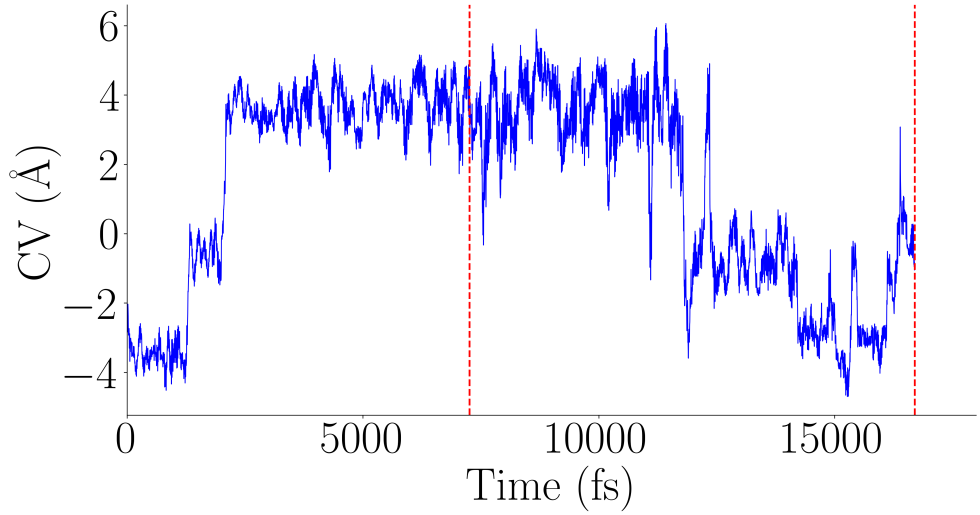


(a)

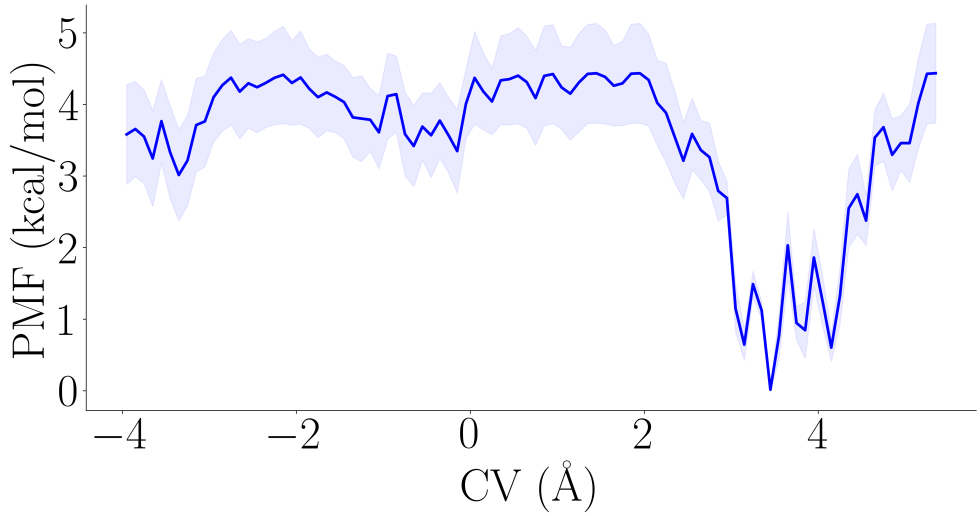


(b)

Figure A.7: (a) The linear combination CV trajectory during the biased simulation using the conventional-7 parameters. The red dashed vertical lines (at 13060 and 23260 fs) mark the limits of the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-7 parameters. A total of 1022 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).



(a)



(b)

Figure A.8: (a) The linear combination CV trajectory during the biased simulation using the conventional-8 parameters. The red dashed vertical lines (at 7260 and 16700 fs) mark the limits of the PMF calculation. (b) The resulting PMF profile for PT between K211 and Y225 using the conventional-8 parameters. A total of 945 PMF profiles were used to calculate the average (blue curve) and standard deviation (shaded light blue area).

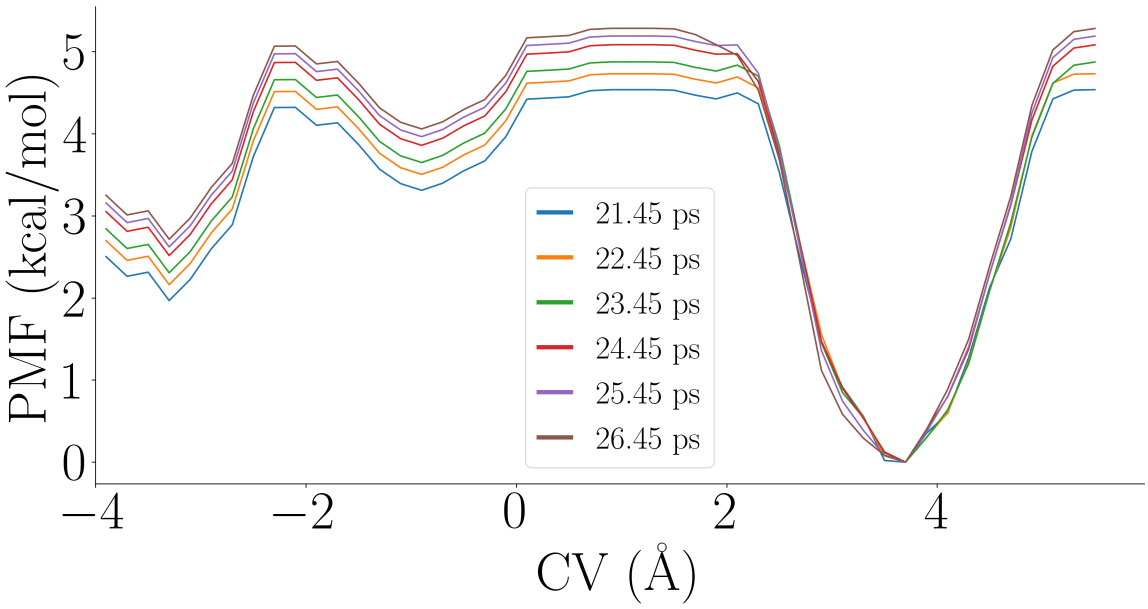


Figure A.9: The plot of convergence for the PMF profiles using the well-tempered 1 parameters.

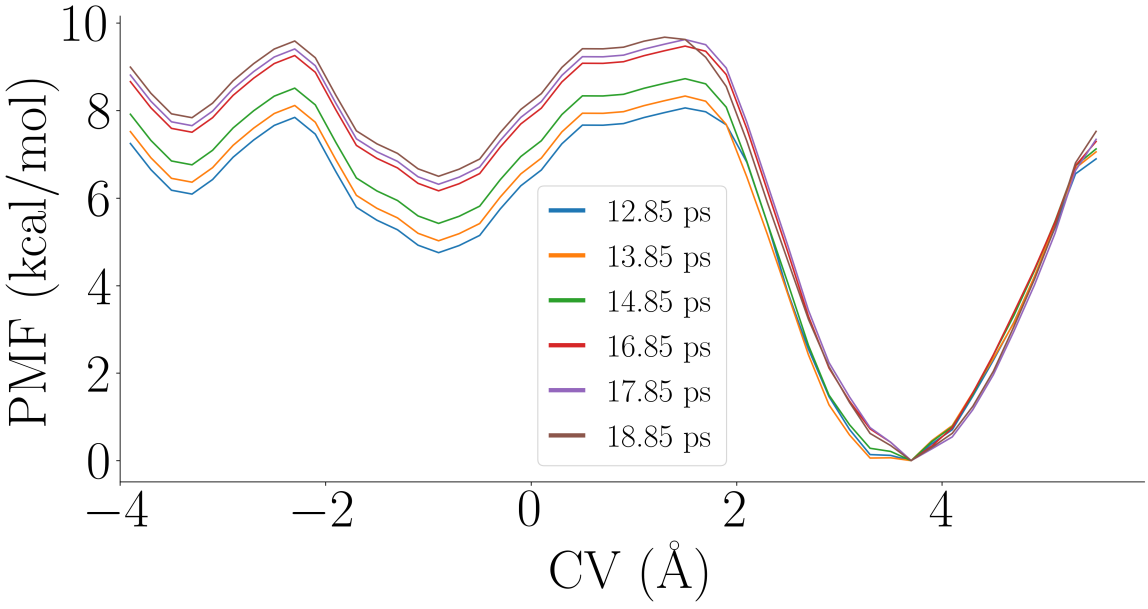


Figure A.10: The plot of convergence for the PMF profiles using the well-tempered 2 parameters.

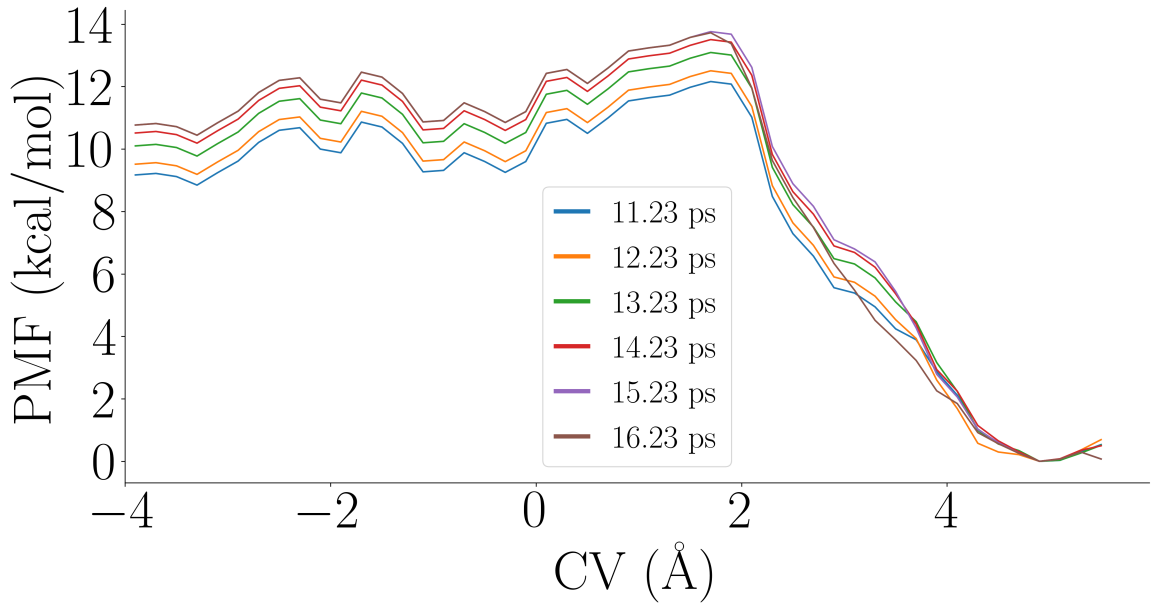


Figure A.11: The plot of convergence for the PMF profiles using the well-tempered 3 parameters.

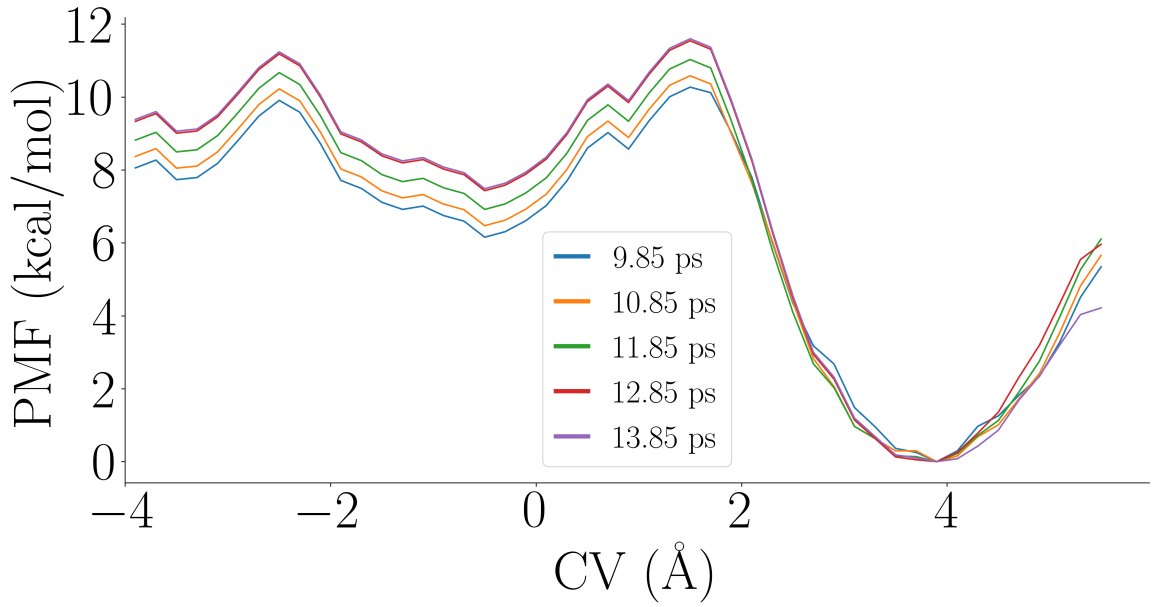


Figure A.12: The plot of convergence for the PMF profiles for the replica of the well-tempered 3 parameters.

Appendix B. ND5 PMF Convergence Plots

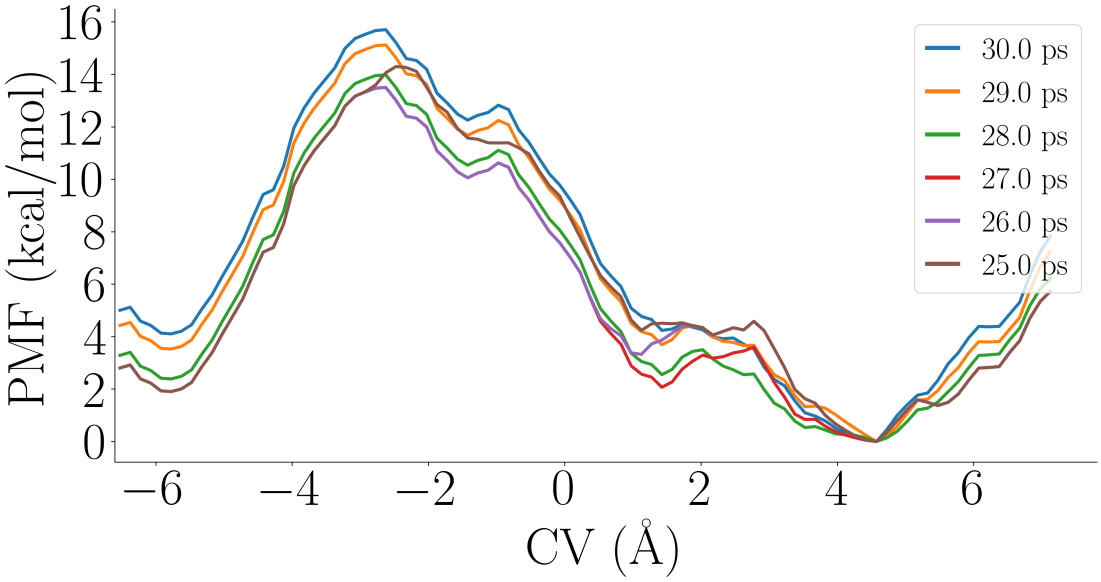


Figure B.1: The convergence plot for the PMF profiles of K305 to K342.

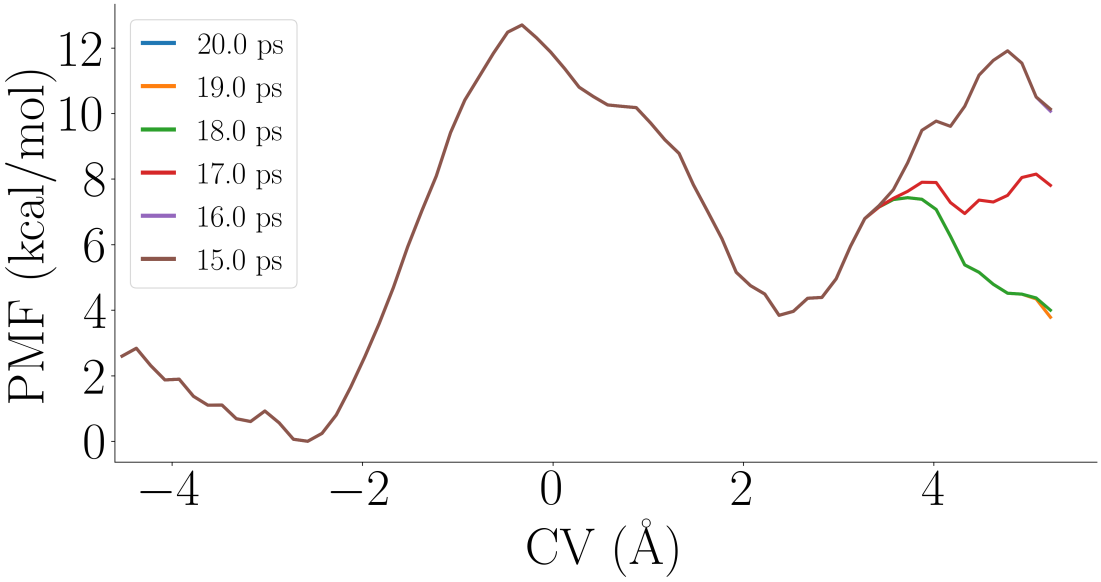


Figure B.2: The convergence plot for the PMF profiles of K305 to K342.

Appendix C. Derivative of mCEC (Jacobian Matrix)

$$\begin{aligned}\xi &= \boldsymbol{\xi} \cdot \frac{(\boldsymbol{r}_a - \boldsymbol{r}_d)}{||\boldsymbol{r}_a - \boldsymbol{r}_d||} \\ &= (\xi_x, \xi_y, \xi_z) \cdot \frac{(x_a - x_d, y_a - y_d, z_a - z_d)}{[(x_a - x_d)^2 + (y_a - y_d)^2 + (z_a - z_d)^2]^{1/2}} \\ &= \frac{\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)}{[(x_a - x_d)^2 + (y_a - y_d)^2 + (z_a - z_d)^2]^{1/2}}\end{aligned} \tag{C.1}$$

$$\begin{aligned}\xi_x =& \sum_i (x_i - x_d) - \sum_j w_j(x_j - x_d) - \sum_i \sum_{j,j\neq i} f(d_{ij}) \cdot (x_i - x_j) \\ &+ \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2} [m(X_\alpha, \{H\}) - m(X_\beta, \{H\})] \cdot (x_\beta - x_\alpha)\end{aligned} \tag{C.2}$$

$$\begin{aligned}\xi_y =& \sum_i (y_i - y_d) - \sum_j w_j(y_j - y_d) - \sum_i \sum_{j,j\neq i} f(d_{ij}) \cdot (y_i - y_j) \\ &+ \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2} [m(X_\alpha, \{H\}) - m(X_\beta, \{H\})] \cdot (y_\beta - y_\alpha)\end{aligned} \tag{C.3}$$

$$\begin{aligned}\xi_z =& \sum_i (z_i - z_d) - \sum_j w_j(z_j - z_d) - \sum_i \sum_{j,j\neq i} f(d_{ij}) \cdot (z_i - z_j) \\ &+ \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2} [m(X_\alpha, \{H\}) - m(X_\beta, \{H\})] \cdot (z_\beta - z_\alpha)\end{aligned} \tag{C.4}$$

Derivative with respect to x:

$$\begin{aligned}\frac{\partial}{\partial x_k}\xi =& \frac{\partial}{\partial x_k}\frac{1}{d_{ad}} \cdot [\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)] \\ &+ \frac{1}{d_{ad}} \cdot \frac{\partial}{\partial x_k}[\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)]\end{aligned} \tag{C.5}$$

Where d_{ad} is the Euclidean norm between the acceptor and donor atom positions. k is the index that goes through all the atoms involved in the collective variable, i.e. $k = 0, \cdots, (N_{\text{atoms}} - 1)$.
Now, there are two derivatives to solve. The first one:

$$\frac{\partial}{\partial x_k}\frac{1}{d_{ad}} = \frac{-1}{2}\frac{1}{[(x_a-x_d)^2+(y_a-y_d)^2+(z_a-z_d)^2]^{3/2}}\cdot$$
$$2[(x_a-x_d)\frac{\partial}{\partial x_k}(x_a-x_d)+(y_a-y_d)\frac{\partial}{\partial x_k}(y_a-y_d)+(z_a-z_d)\frac{\partial}{\partial x_k}(z_a-z_d)]$$
$$=\frac{-1}{[(x_a-x_d)^2+(y_a-y_d)^2+(z_a-z_d)^2]^{3/2}}\cdot[(x_a-x_d)\frac{\partial}{\partial x_k}(x_a-x_d)] \quad \text{(C.6)}$$

Where:

$$\frac{\partial}{\partial x_k}(x_a-x_d)=\begin{cases}1 & k=a\\ -1 & k=d\\ 0 & \text{otherwise}\end{cases}$$

The second one:

$$\frac{\partial}{\partial x_k}[\xi_x(x_a-x_d)+\xi_y(y_a-y_d)+\xi_z(z_a-z_d)] =$$
$$\frac{\partial}{\partial x_k}\xi_x(x_a-x_d)+\frac{\partial}{\partial x_k}\xi_y(y_a-y_d)+\frac{\partial}{\partial x_k}\xi_z(z_a-z_d)$$
$$=\frac{\partial}{\partial x_k}\xi_x\cdot(x_a-x_d)+\xi_x\cdot\frac{\partial}{\partial x_k}(x_a-x_d)+$$
$$\frac{\partial}{\partial x_k}\xi_y\cdot(y_a-y_d)+\xi_y\cdot\frac{\partial}{\partial x_k}(y_a-y_d)+$$
$$\frac{\partial}{\partial x_k}\xi_z\cdot(z_a-z_d)+\xi_z\cdot\frac{\partial}{\partial x_k}(z_a-z_d)$$
$$=\frac{\partial}{\partial x_k}\xi_x\cdot(x_a-x_d)+\xi_x\cdot\frac{\partial}{\partial x_k}(x_a-x_d)+$$
$$\frac{\partial}{\partial x_k}\xi_y\cdot(y_a-y_d)+\frac{\partial}{\partial x_k}\xi_z\cdot(z_a-z_d) \quad \text{(C.7)}$$

Now there are 3 derivatives to figure out:

$$\frac{\partial}{\partial x_k}\xi_x=\sum_i\frac{\partial}{\partial x_k}(x_i-x_d)-\sum_j w_j\frac{\partial}{\partial x_k}(x_j-x_d)$$
$$-\sum_i\sum_{j,j\neq i}(\frac{\partial}{\partial x_k}f(d_{ij})\cdot(x_i-x_j)+f(d_{ij})\cdot\frac{\partial}{\partial x_k}(x_i-x_j))$$
$$+\sum_{\text{pairs}}\frac{w_{\text{pair}}^{\alpha,\beta}}{2}\Bigg([\frac{\partial}{\partial x_k}m(X_\alpha,\{H\})-\frac{\partial}{\partial x_k}m(X_\beta,\{H\})]\cdot(x_\beta-x_\alpha)$$
$$+[m(X_\alpha,\{H\})-m(X_\beta,\{H\})]\cdot\frac{\partial}{\partial x_k}(x_\beta-x_\alpha)\Bigg) \quad \text{(C.8)}$$

Where:

$$\frac{\partial}{\partial x_k}(x_i - x_d) = \begin{cases} 1 & k = i \\ -1 & k = d \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial x_k}(x_j - x_d) = \begin{cases} 1 & k = j, k \neq d \\ -1 & k = d, k \neq j \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial x_k}(x_\beta - x_\alpha) = \begin{cases} 1 & k = \beta \\ -1 & k = \alpha \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial x_k} f(d_{ij}) = \frac{\partial}{\partial x_k}[1 + e^{\frac{d_{ij}-B}{A}}]^{-1} = -1[1 + e^{\frac{d_{ij}-B}{A}}]^{-2} e^{\frac{d_{ij}-B}{A}} \frac{1}{A} \frac{\partial}{\partial x_k} d_{ij} \tag{C.9}$$

where:

$$\begin{aligned} \frac{\partial}{\partial x_k} d_{ij} &= \frac{\partial}{\partial x_k}[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{1/2} \\ &= \frac{1}{2}[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{-1/2} 2[(x_i - x_j) \cdot \frac{\partial}{\partial x_k}(x_i - x_j)] \\ &= \frac{1}{d_{ij}}(x_i - x_j) \cdot \frac{\partial}{\partial x_k}(x_i - x_j) \end{aligned} \tag{C.10}$$

where:

$$\frac{\partial}{\partial x_k}(x_i - x_j) = \begin{cases} 1 & k = i \\ -1 & k = j \\ 0 & \text{otherwise} \end{cases}$$

Now, checking the other derivative:

$$\begin{aligned} \frac{\partial}{\partial x_k} m(X, \{H\}) =& \frac{\partial}{\partial x_k} \frac{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n+1}}{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n}} \\ &= \frac{\partial}{\partial x_k} \sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n+1} \cdot \frac{1}{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n}} \\ &+ \sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n+1} \frac{\partial}{\partial x_k} \frac{1}{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n}} \end{aligned} \tag{C.11}$$

Where:

$$\frac{\partial}{\partial x_k}\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n+1} = (n+1)\sum_i \left(\left(f_{\text{sw}}(d_{X,H_i})\right)^n \cdot \frac{\partial}{\partial x_k} f_{\text{sw}}(d_{X,H_i})\right) \tag{C.12}$$

and

$$\frac{\partial}{\partial x_k}\frac{1}{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n} = \frac{-1\cdot(n)\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n-1}\frac{\partial}{\partial x_k} f_{\text{sw}}(d_{X,H_i})}{\left(\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n\right)^2} \tag{C.13}$$

so for $\frac{\partial}{\partial x_k} m(X,\{H\})$:

$$\begin{aligned}\frac{\partial}{\partial x_k} m(X,\{H\}) =& (n+1)\cdot \frac{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n \frac{\partial}{\partial x_k} f_{\text{sw}}(d_{X,H_i})}{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n} \\ &- (n)\cdot \frac{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n+1} \cdot \sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n-1}\frac{\partial}{\partial x_k} f_{\text{sw}}(d_{X,H_i})}{\left(\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n\right)^2}\end{aligned} \tag{C.14}$$

Now, going back to the previous terms:

$$\begin{aligned}\frac{\partial}{\partial x_k}\xi_y &= 0 - 0 - \sum_i \sum_{j,j\neq i} \left(\frac{\partial}{\partial x_k} f(d_{ij})\cdot(y_i - y_j) + f(d_{ij})\cdot\frac{\partial}{\partial x_k}(y_i - y_j)\right) \\ &\quad + \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2}\left(\left[\frac{\partial}{\partial x_k} m(X_\alpha,\{H\}) - \frac{\partial}{\partial x_k} m(X_\beta,\{H\})\right]\cdot(y_\beta - y_\alpha)\right. \\ &\quad \left. + \left[m(X_\alpha,\{H\}) - m(X_\beta,\{H\})\right]\cdot\frac{\partial}{\partial x_k}(y_\beta - y_\alpha)\right) \\ &= -\sum_i \sum_{j,j\neq i} \frac{\partial}{\partial x_k} f(d_{ij})\cdot(y_i - y_j) \\ &\quad + \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2}\left(\left[\frac{\partial}{\partial x_k} m(X_\alpha,\{H\}) - \frac{\partial}{\partial x_k} m(X_\beta,\{H\})\right]\cdot(y_\beta - y_\alpha)\right)\end{aligned} \tag{C.15}$$

and

$$\begin{aligned}\frac{\partial}{\partial x_k}\xi_z &= -\sum_i \sum_{j,j\neq i} \frac{\partial}{\partial x_k} f(d_{ij})\cdot(z_i - z_j) \\ &\quad + \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2}\left(\left[\frac{\partial}{\partial x_k} m(X_\alpha,\{H\}) - \frac{\partial}{\partial x_k} m(X_\beta,\{H\})\right]\cdot(z_\beta - z_\alpha)\right)\end{aligned} \tag{C.16}$$

Next is the derivative with respect to y:

$$\begin{aligned}\frac{\partial}{\partial y_k}\xi = \frac{\partial}{\partial y_k}\frac{1}{d_{ad}} \cdot [\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)] \\ + \frac{1}{d_{ad}} \cdot \frac{\partial}{\partial y_k}[\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)]\end{aligned} \tag{C.17}$$

There are two derivatives to solve:
The first one:

$$\begin{aligned}\frac{\partial}{\partial y_k}\frac{1}{d_{ad}} = \frac{-1}{2}\frac{1}{[(x_a - x_d)^2 + (y_a - y_d)^2 + (z_a - z_d)^2]^{3/2}} \cdot \\ 2[(x_a - x_d)\frac{\partial}{\partial y_k}(x_a - x_d) + (y_a - y_d)\frac{\partial}{\partial y_k}(y_a - y_d) + (z_a - z_d)\frac{\partial}{\partial y_k}(z_a - z_d)] \\ = \frac{-1}{[(x_a - x_d)^2 + (y_a - y_d)^2 + (z_a - z_d)^2]^{3/2}} \cdot [(y_a - y_d)\frac{\partial}{\partial y_k}(y_a - y_d)]\end{aligned} \tag{C.18}$$

Where:

$$\frac{\partial}{\partial y_k}(y_a - y_d) = \begin{cases} 1 & k = a \\ -1 & k = d \\ 0 & \text{otherwise} \end{cases}$$

The second one:

$$\begin{aligned}\frac{\partial}{\partial y_k}[\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)] = \\ \frac{\partial}{\partial y_k}\xi_x(x_a - x_d) + \frac{\partial}{\partial y_k}\xi_y(y_a - y_d) + \frac{\partial}{\partial y_k}\xi_z(z_a - z_d) \\ = \frac{\partial}{\partial y_k}\xi_x \cdot (x_a - x_d) + \xi_x \cdot \frac{\partial}{\partial y_k}(x_a - x_d) \\ + \frac{\partial}{\partial y_k}\xi_y \cdot (y_a - y_d) + \xi_y \cdot \frac{\partial}{\partial y_k}(y_a - y_d) \\ + \frac{\partial}{\partial y_k}\xi_z \cdot (z_a - z_d) + \xi_z \cdot \frac{\partial}{\partial y_k}(z_a - z_d) \\ = \frac{\partial}{\partial y_k}\xi_x \cdot (x_a - x_d) + \frac{\partial}{\partial y_k}\xi_y \cdot (y_a - y_d) \\ + \xi_y \cdot \frac{\partial}{\partial y_k}(y_a - y_d) + \frac{\partial}{\partial y_k}\xi_z \cdot (z_a - z_d)\end{aligned} \tag{C.19}$$

We have 3 derivatives to figure out:

$$\begin{aligned}\frac{\partial}{\partial y_k}\xi_y = &\sum_i \frac{\partial}{\partial y_k}(y_i - y_d) - \sum_j w_j \frac{\partial}{\partial y_k}(y_j - y_d) \\ &- \sum_i \sum_{j,j\neq i} \left(\frac{\partial}{\partial y_k} f(d_{ij}) \cdot (y_i - y_j) + f(d_{ij}) \cdot \frac{\partial}{\partial y_k}(y_i - y_j)\right) \\ &+ \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2} \Bigg(\left[\frac{\partial}{\partial y_k} m(X_\alpha, \{H\}) - \frac{\partial}{\partial y_k} m(X_\beta, \{H\})\right] \cdot (y_\beta - y_\alpha) \\ &+ \left[m(X_\alpha, \{H\}) - m(X_\beta, \{H\})\right] \cdot \frac{\partial}{\partial y_k}(y_\beta - y_\alpha) \Bigg)\end{aligned} \tag{C.20}$$

Where:

$$\frac{\partial}{\partial y_k}(y_i - y_d) = \begin{cases} 1 & k = i \\ -1 & k = d \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial y_k}(y_j - y_d) = \begin{cases} 1 & k = j, k \neq d \\ -1 & k = d, k \neq j \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial y_k}(y_\beta - y_\alpha) = \begin{cases} 1 & k = \beta \\ -1 & k = \alpha \\ 0 & \text{otherwise} \end{cases}$$

and for $\frac{\partial}{\partial y_k} m(X, \{H\})$:

$$\begin{aligned}\frac{\partial}{\partial y_k} m(X, \{H\}) = &(n+1) \cdot \frac{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n \frac{\partial}{\partial y_k} f_{\text{sw}}(d_{X,H_i})}{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n} \\ &- (n) \cdot \frac{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n+1} \cdot \sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n-1} \frac{\partial}{\partial y_k} f_{\text{sw}}(d_{X,H_i})}{\left(\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n\right)^2}\end{aligned} \tag{C.21}$$

$$\frac{\partial}{\partial y_k} f(d_{ij}) = \frac{\partial}{\partial y_k}[1 + e^{\frac{d_{ij}-B}{A}}]^{-1} = -1[1 + e^{\frac{d_{ij}-B}{A}}]^{-2} e^{\frac{d_{ij}-B}{A}} \frac{1}{A} \frac{\partial}{\partial y_k} d_{ij} \tag{C.22}$$

where:

$$\begin{aligned}\frac{\partial}{\partial y_k} d_{ij} &= \frac{\partial}{\partial y_k}[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{1/2} \\ &= \frac{1}{2}[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{-1/2} 2[(y_i - y_j) \cdot \frac{\partial}{\partial y_k}(y_i - y_j)] \\ &= \frac{1}{d_{ij}}(y_i - y_j) \cdot \frac{\partial}{\partial y_k}(y_i - y_j) \end{aligned} \tag{C.23}$$

where:

$$\frac{\partial}{\partial y_k}(y_i - y_j) = \begin{cases} 1 & k = i \\ -1 & k = j \\ 0 & \text{otherwise} \end{cases}$$

Now, going back to the previous terms:

$$\begin{aligned}\frac{\partial}{\partial y_k}\xi_x = &- \sum_i \sum_{j,j\neq i} \frac{\partial}{\partial y_k} f(d_{ij}) \cdot (x_i - x_j) \\ &+ \sum_{\text{pairs}} \left(\frac{w_{\text{pair}}^{\alpha,\beta}}{2} [\frac{\partial}{\partial y_k} m(X_\alpha, \{H\}) - \frac{\partial}{\partial y_k} m(X_\beta, \{H\})] \cdot (x_\beta - x_\alpha) \right) \end{aligned} \tag{C.24}$$

and

$$\begin{aligned}\frac{\partial}{\partial y_k}\xi_z = &- \sum_i \sum_{j,j\neq i} \frac{\partial}{\partial y_k} f(d_{ij}) \cdot (z_i - z_j) \\ &+ \sum_{\text{pairs}} \left(\frac{w_{\text{pair}}^{\alpha,\beta}}{2} [\frac{\partial}{\partial y_k} m(X_\alpha, \{H\}) - \frac{\partial}{\partial y_k} m(X_\beta, \{H\})] \cdot (z_\beta - z_\alpha) \right) \end{aligned} \tag{C.25}$$

Now derivative with respect to the z component:

$$\begin{aligned}\frac{\partial}{\partial z_k}\xi = &\frac{\partial}{\partial z_k}\frac{1}{d_{ad}} \cdot [\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)] \\ &+ \frac{1}{d_{ad}} \cdot \frac{\partial}{\partial z_k}[\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)] \end{aligned} \tag{C.26}$$

Now, there are two derivatives to solve. The first one:

$$\begin{aligned}\frac{\partial}{\partial z_k}\frac{1}{d_{ad}} = &\frac{-1}{2} \frac{1}{[(x_a - x_d)^2 + (y_a - y_d)^2 + (z_a - z_d)^2]^{3/2}} \cdot \\ &2[(x_a - x_d)\frac{\partial}{\partial z_k}(x_a - x_d) + (y_a - y_d)\frac{\partial}{\partial z_k}(y_a - y_d) + (z_a - z_d)\frac{\partial}{\partial z_k}(z_a - z_d)] \\ = &\frac{-1}{[(x_a - x_d)^2 + (y_a - y_d)^2 + (z_a - z_d)^2]^{3/2}} \cdot [(z_a - z_d)\frac{\partial}{\partial z_k}(z_a - z_d)] \end{aligned} \tag{C.27}$$

Where:

$$\frac{\partial}{\partial z_k}(z_a - z_d) = \begin{cases} 1 & k = a \\ -1 & k = d \\ 0 & \text{otherwise} \end{cases}$$

The second one:

$$\begin{aligned} \frac{\partial}{\partial z_k}[\xi_x(x_a - x_d) + \xi_y(y_a - y_d) + \xi_z(z_a - z_d)] = \\ & \frac{\partial}{\partial z_k}\xi_x(x_a - x_d) + \frac{\partial}{\partial z_k}\xi_y(y_a - y_d) + \frac{\partial}{\partial z_k}\xi_z(z_a - z_d) \\ & = \frac{\partial}{\partial z_k}\xi_x \cdot (x_a - x_d) + \xi_x \cdot \frac{\partial}{\partial z_k}(x_a - x_d) + \\ & \frac{\partial}{\partial z_k}\xi_y \cdot (y_a - y_d) + \xi_y \cdot \frac{\partial}{\partial z_k}(y_a - y_d) + \\ & \frac{\partial}{\partial z_k}\xi_z \cdot (z_a - z_d) + \xi_z \cdot \frac{\partial}{\partial z_k}(z_a - z_d) \\ & = \frac{\partial}{\partial z_k}\xi_x \cdot (x_a - x_d) + \frac{\partial}{\partial z_k}\xi_y \cdot (y_a - y_d) + \\ & \frac{\partial}{\partial z_k}\xi_z \cdot (z_a - z_d) + \xi_z \cdot \frac{\partial}{\partial z_k}(z_a - z_d) \end{aligned} \tag{C.28}$$

Now there are 3 derivatives to figure out:

$$\begin{aligned} \frac{\partial}{\partial z_k}\xi_z = & \sum_i \frac{\partial}{\partial z_k}(z_i - z_d) - \sum_j w_j \frac{\partial}{\partial z_k}(z_j - z_d) \\ & - \sum_i \sum_{j,j\neq i} \left(\frac{\partial}{\partial z_k} f(d_{ij}) \cdot (z_i - z_j) + f(d_{ij}) \cdot \frac{\partial}{\partial z_k}(z_i - z_j)\right) \\ & + \sum_{\text{pairs}} \frac{w_{\text{pair}}^{\alpha,\beta}}{2} \Bigg(\left[\frac{\partial}{\partial z_k} m(X_\alpha, \{H\}) - \frac{\partial}{\partial z_k} m(X_\beta, \{H\})\right] \cdot (z_\beta - z_\alpha) \\ & + [m(X_\alpha, \{H\}) - m(X_\beta, \{H\})] \cdot \frac{\partial}{\partial z_k}(z_\beta - z_\alpha) \Bigg) \end{aligned} \tag{C.29}$$

Where:

$$\frac{\partial}{\partial z_k}(z_i - z_d) = \begin{cases} 1 & k = i \\ -1 & k = d \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial z_k}(z_j - z_d) = \begin{cases} 1 & k = j, k \neq d \\ -1 & k = d, k \neq j \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial}{\partial z_k}(z_\beta - z_\alpha) = \begin{cases} 1 & k = \beta \\ -1 & k = \alpha \\ 0 & \text{otherwise} \end{cases}$$

and for $\frac{\partial}{\partial z_k} m(X, \{H\})$ we will have:

$$\begin{aligned} \frac{\partial}{\partial z_k} m(X, \{H\}) =& (n+1) \cdot \frac{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n \frac{\partial}{\partial z_k} f_{\text{sw}}(d_{X,H_i})}{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n} \\ & - (n) \cdot \frac{\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n+1} \cdot \sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^{n-1} \frac{\partial}{\partial z_k} f_{\text{sw}}(d_{X,H_i})}{\left(\sum_i \left(f_{\text{sw}}(d_{X,H_i})\right)^n\right)^2} \end{aligned} \tag{C.30}$$

$$\frac{\partial}{\partial z_k} f(d_{ij}) = \frac{\partial}{\partial z_k}[1 + e^{\frac{d_{ij}-B}{A}}]^{-1} = -1[1 + e^{\frac{d_{ij}-B}{A}}]^{-2} e^{\frac{d_{ij}-B}{A}} \frac{1}{A} \frac{\partial}{\partial z_k} d_{ij} \tag{C.31}$$

where:

$$\begin{aligned} \frac{\partial}{\partial z_k} d_{ij} &= \frac{\partial}{\partial z_k}[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{1/2} \\ &= \frac{1}{2}[(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{-1/2} 2[(z_i - z_j) \cdot \frac{\partial}{\partial z_k}(z_i - z_j)] \\ &= \frac{1}{d_{ij}}(z_i - z_j) \cdot \frac{\partial}{\partial z_k}(z_i - z_j) \end{aligned} \tag{C.32}$$

where:

$$\frac{\partial}{\partial z_k}(z_i - z_j) = \begin{cases} 1 & k = i \\ -1 & k = j \\ 0 & \text{otherwise} \end{cases}$$

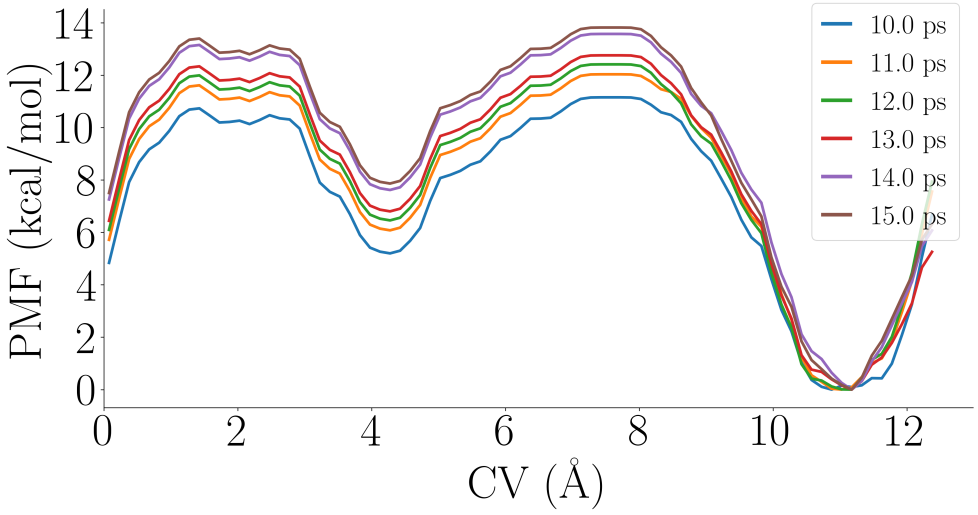
Now, going back to the previous terms:

$$\frac{\partial}{\partial z_k}\xi_x = -\sum_i \sum_{j,j\neq i} \frac{\partial}{\partial z_k} f(d_{ij}) \cdot (x_i - x_j) + \sum_{\text{pairs}} \left(\frac{w_{\text{pair}}^{\alpha,\beta}}{2} [\frac{\partial}{\partial z_k} m(X_\alpha, \{H\}) - \frac{\partial}{\partial z_k} m(X_\beta, \{H\})] \cdot (x_\beta - x_\alpha) \right) \tag{C.33}$$

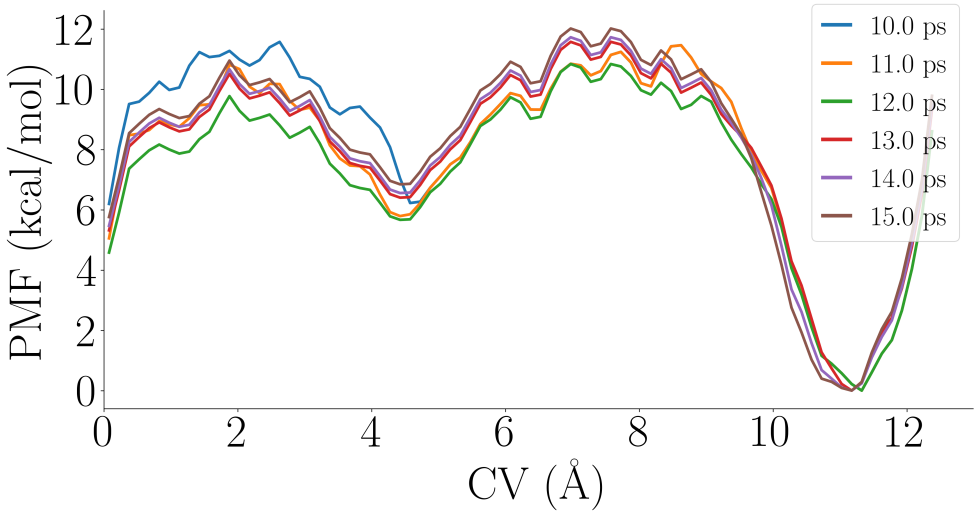
and

$$\frac{\partial}{\partial z_k}\xi_y = -\sum_i \sum_{j,j\neq i} \frac{\partial}{\partial z_k} f(d_{ij}) \cdot (y_i - y_j) + \sum_{\text{pairs}} \left(\frac{w_{\text{pair}}^{\alpha,\beta}}{2} [\frac{\partial}{\partial z_k} m(X_\alpha, \{H\}) - \frac{\partial}{\partial z_k} m(X_\beta, \{H\})] \cdot (y_\beta - y_\alpha) \right) \tag{C.34}$$

Appendix D. ND2/mCEC Results

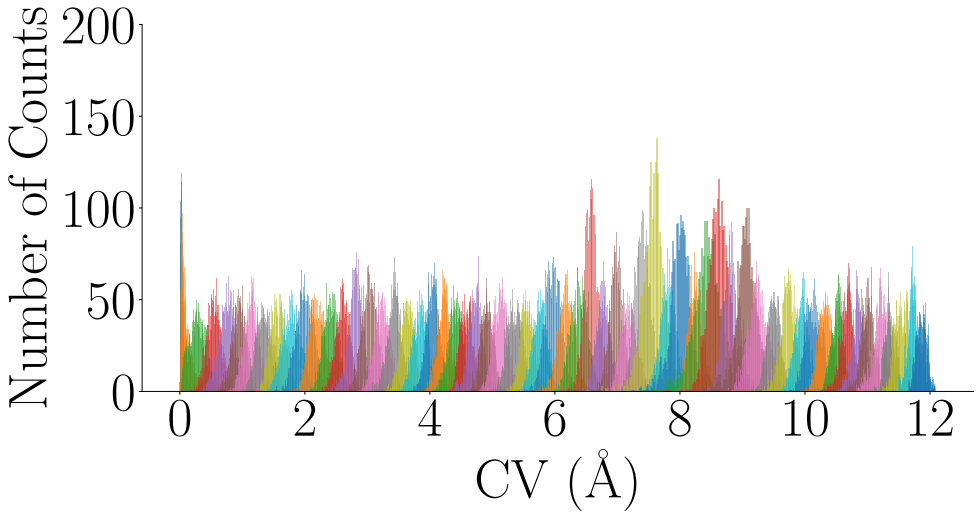


(a)

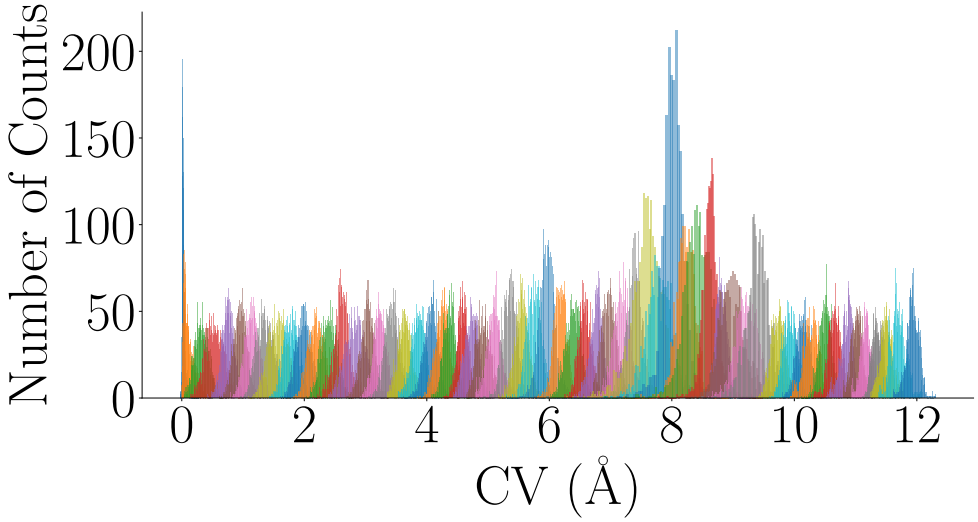


(b)

Figure D.1: The convergence plot for the PMF profiles of K211 to Y226. (a) No additional water molecules included. (b) Additional water molecules included.



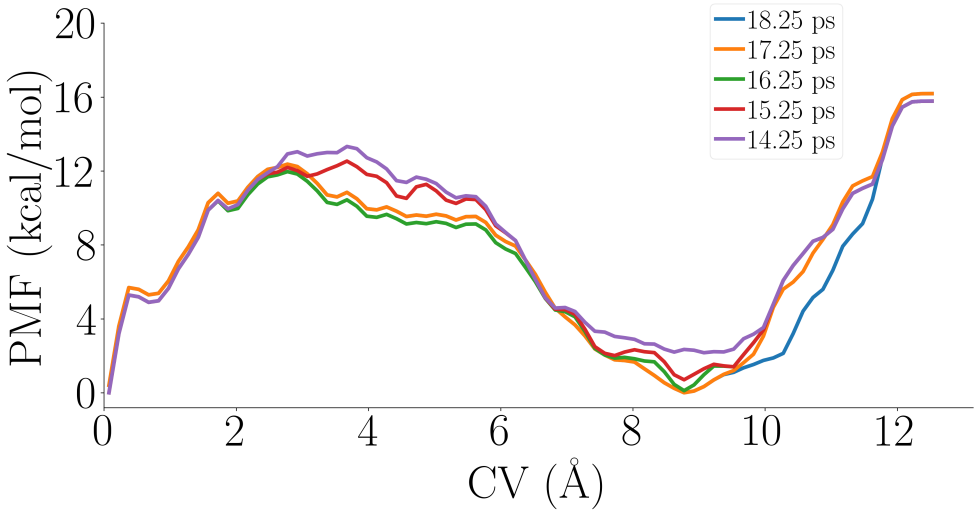
(a)



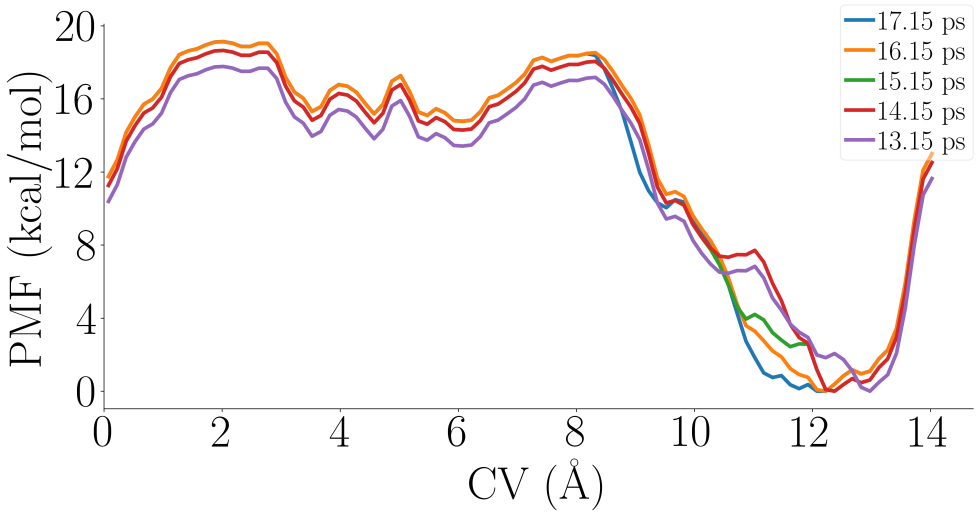
(b)

Figure D.2: The counts of the value of mCEC CV in all the simulated windows. (a) No additional water molecules included. (b) Additional water molecules included.

Appendix E. E-channel PMF Convergence Plots



(a)



(b)

Figure E.1: (a) PMF convergence of D66 to E34 proton transfer pathway. (b) PMF convergence of E202 to E192.